Selected Papers of #AoIR2022:
The 23rd Annual Conference of the
Association of Internet Researchers
Dublin, Ireland / 2-5 Nov 2022

EXAMINING THE EFFECTIVENESS OF ARTIFICIAL INTELLIGENCE-BASED CYBERBULLYING MODERATION ON ONLINE PLATFORMS: TRANSPARENCY IMPLICATIONS

Kanishk Verma
ADAPT Centre, DCU Anti Bullying Centre, Dublin City University

Tijana Milosevic
DCU Anti Bullying Centre, Dublin City University

Brian Davis
ADAPT Centre, Dublin City University

James O'Higgins Norman
DCU Anti Bullying Centre, Dublin City University

Introduction

The COVID-19 pandemic in 2020 led to an increased usage of Online Social Network (OSN) and Multiplayer Online Gaming (MOG) platforms by children, accompanied by a rise in cyberbullying [1]–[7]. Consequently, there is a need for OSN and MOG platforms to enhance the effectiveness of content moderation, often relying on automation using Artificial Intelligence (AI) [8]–[10].

Measuring the efficacy of automated moderation mechanisms for detecting online bullying-related content is challenging due to their proprietary nature. Our study analyses the research literature, transparency reports, and platform blogs to understand the efficacy of AI tools used by OSN and MOG platforms to regulate cyberbullying. To that effect, our research questions include,

1. How does AI identify and mitigate cyberbullying, and what user interactions does it monitor?
2. What do we know about the effectiveness of AI tools for detecting cyberbullying and enforcing policies against it?
3. How does the information available about these processes shed light on the transparency of online platforms?

Suggested Citation (APA): Verma, K., Milosevic, T., Davis, B., and O'Higgins Norman, J. (2022, November). Examining the Effectiveness of Artificial Intelligence-Based Cyberbullying Moderation on Online Platforms: Transparency Implications. Paper presented at AoIR 2022: The 23rd Annual Conference of the Association of Internet Researchers. Dublin, Ireland: AoIR. Retrieved from http://spir.aoir.org.

Search Strategy

To examine how OSN/MOGs moderate cyberbullying content, we leveraged a two-method search strategy,

I. Examine platforms for information about any AI-enabled solutions for cyberbullying *detection, prevention,* and *intervention.*
II. Querying search engines with specific keywords to retrieve news articles or blog posts that discuss aplatform's AI-based solutions.

We inspected 18 platforms (See Appendix A) and their web resources throughout 2020 and 2021. We excluded articles and blog posts that did not provide information about new system/mechanism developments or designs by the platform, or shed light on the transparency of such systems The proprietary nature of AI-based platform solutions can be obstacle to their evaluation hence, as a an alternative look for computational studies that leverage conceptually similar AI algorithms for detecting online bullying. To that effect, we conducted a scoping study review of studies across academic repositories (See Appendix B), to identify studies and datasets from two perspectives:

I. Data perspective - exploring types of cyberbullying-related data, data annotation strategy
II. Effectiveness perspective - helps assess the capabilities of the state-of-the-art AI-enabled solutions to detect cyberbullying content.

We excluded articles that do not discuss an online bullying dataset, an AI evaluation technique or benchmarking mechanism or is not a novel or reproducible system to detect cyberbullying.

Key Findings - Industry Initiatives

Our search yielded 150 relevant web-resources, indicating that AI solutions encompass various techniques to recognise and extract user features from multimodal human interaction data. We categorised these solutions as proactive intervention, where platforms intervene before user reporting, and reactive intervention, where platforms intervene after user reporting. Apart from *Facebook, Instagram's* proactive solutions *and Whisper*, *Minecraft*'s reactive solutions, we did not find information about whether and how the other 14 platforms leverage AI to counter cyberbullying on their platforms.

Descriptive information on AI techniques for content moderation is publicly available by Facebook and Google. Facebook AI developed DeepText [11], Linformer [12], Whole Post Integrity Embeddings (WPIE) - Reinforced Integrity Optimizer (RIO) [13] to identify harmful content across different contexts. However, the application and effectiveness of these advancements in addressing cyberbullying remain *largely unknown*.
Google and Jigsaw[1] developed, *Perspective*, to address toxic and abusive online comments. To facilitate algorithmic transparency and fairness, the information on

[1] https://jigsaw.google.com

Perspective's training process, architecture, and training data are publicly available [14]–[16]. However, our analysis and previous study [17] revealed that Perspective can be deceived by subtle phrase modifications, resulting in reduced toxicity scores provided by Perspective.

Key Findings - Academic literature

A total of 157 computational articles were reviewed, resulting in selection of 119 articles, including 15 dataset studies, focused on cyberbullying. Twitter emerged as the most examined platform, followed by Instagram, AskFM, and MySpace. High quality datasets are vital for stimulating complex computational processes and are characterised by reliable annotations and inter-annotator agreement. Only four datasets by [18]–[21] met these criteria, covering diverse aspects of cyberbullying and user-interactivity information.

Cyberbullying, encompassing various behaviour and involving not only perpetrators and victims but also bystanders, has been extensively studied. However, computational literature lacks diversity, with a majority of studies focusing on text-based communication, and only 20% of the studies scoped rely on other modes of communication than text. Moreover, almost 73% of the studies scoped focus on classifying cyberbullying as binary, and 87% focus only on the English language. (See Appendix C).

Conclusion

Co-relating our search results with the complex cyberbullying nomenclature, we found that despite the novel efforts by both academia and industry, the publicly available resources for independent researchers to scrutinise industry efforts, as well as to design a novel, effective, efficient and most importantly *explainable* cyberbullying detection model are **extremely scarce.** The availability of only two fine-grained first-rate datasets by [18], [21] is an indication that cyberbullying research progress in academia is slow. On the other hand, due to the proprietary nature of Facebook AI's DeepText, Linformer, RIO and WPIE, **not much** can be known about its effectiveness in tackling cyberbullying. Although the progress on Google and Jigsaw's *Perspective* is commendable in tackling toxicity on online platforms, it still can be deceived by subtle modifications to the text. Our search also revealed that current platform policies fail to recognise "*bystanders*" in bullying incidents.

Despite many technological advancements and having witnessed an increase in cyberbullying incidents on OSN/MOG platforms, the AI-enabled solutions for cyberbullying remain imperfect. Through this study, we were able to ascertain the scarcity of qualitative and quantitative research to devise better and more effective tools to counter cyberbullying. Going forward, AI tools that leverage not only text but also multimodal data to detect all varied forms and roles in cyberbullying should be developed. To close this gap and devise effective solutions, a multi-disciplinary

environment of computational and social science researchers with teens and youth at focus must be leveraged.

Bibliography

[1] P. Babvey, F. Capela, C. Cappa, C. Lipizzi, N. Petrowski, and J. Ramirez-Marquez, 'Using social media data for assessing children's exposure to violence during the COVID-19 pandemic', *Child Abuse Negl.*, vol. 116, p. 104747, Jun. 2021, doi: 10.1016/j.chiabu.2020.104747.

[2] 'Facebook Reports Fourth Quarter and Full Year 2020 Results'. https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx (accessed Feb. 25, 2022).

[3] D. T. Milosevic, D. Laffan, and J. O. Norman, 'A Study on Digital Practices, Safety and Wellbeing', p. 16.

[4] 'Coronavirus Disease (COVID-19) and Its Implications for Protecting Children Online'. https://www.unicef.org/documents/covid-19-and-implications-protecting-children-online (accessed Feb. 25, 2022).

[5] G. U. Utemissova, S. Danna, and V. N. Nikolaevna, 'Cyberbullying during the COVID-19 pandemic', *Glob. J. Guid. Couns. Sch. Curr. Perspect.*, vol. 11, no. 2, Art. no. 2, Jul. 2021, doi: 10.18844/gjgc.v11i2.5471.

[6] V. Charisi *et al.*, 'Artificial Intelligence and the Rights of the Child : Towards an Integrated Agenda for Research and Policy', *JRC Publications Repository*, Jun. 07, 2022. https://publications.jrc.ec.europa.eu/repository/handle/JRC127564 (accessed Oct. 01, 2022).

[7] European Commission. Joint Research Centre., *How children (10-18) experienced online risks during the Covid-19 lockdown :spring 2020 : key findings from surveying families in 11 European countries.* LU: Publications Office, 2021. Accessed: Feb. 25, 2022. [Online]. Available: https://data.europa.eu/doi/10.2760/066196

[8] T. Milosevic, *Protecting Children Online?: Cyberbullying Policies of Social Media Companies*. 2018. doi: 10.7551/mitpress/11008.001.0001.

[9] T. Gillespie, *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press, 2018. doi: 10.12987/9780300235029.

[10] T. Gillespie *et al.*, 'Expanding the debate about content moderation: scholarly research agendas for the coming policy debates', *Internet Policy Rev.*, vol. 9, no. 4, Oct. 2020, Accessed: Feb. 25, 2022. [Online]. Available: https://policyreview.info/articles/analysis/expanding-debate-about-content-moderation-scholarly-research-agendas-coming-policy

[11] 'Introducing DeepText: Facebook's text understanding engine - Engineering at Meta'. https://engineering.fb.com/2016/06/01/core-data/introducing-deeptext-facebook-s-text-understanding-engine/ (accessed Feb. 28, 2022).

[12] S. Wang, B. Z. Li, M. Khabsa, H. Fang, and H. Ma, 'Linformer: Self-Attention with Linear Complexity', *ArXiv200604768 Cs Stat*, Jun. 2020, Accessed: Feb. 28, 2022. [Online]. Available: http://arxiv.org/abs/2006.04768

[13] 'How AI is getting better at detecting hate speech'. https://ai.facebook.com/blog/how-ai-is-getting-better-at-detecting-hate-speech/ (accessed Feb. 28, 2022).

[14] 'Perspective API - Case Studies'. https://perspectiveapi.com/case-studies/ (accessed Dec.

08, 2022).
[15] ‘About the API - Attributes and Languages’. https://support.perspectiveapi.com/s/about-the-api-attributes-and-languages (accessed Feb. 28, 2022).
[16] ‘Jigsaw Unintended Bias in Toxicity Classification’. https://kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification (accessed Feb. 28, 2022).
[17] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, ‘Deceiving Google’s Perspective API Built for Detecting Toxic Comments’, *ArXiv170208138 Cs*, Feb. 2017, Accessed: Feb. 28, 2022. [Online]. Available: http://arxiv.org/abs/1702.08138
[18] C. V. Hee *et al.*, ‘Automatic detection of cyberbullying in social media text’, *PLOS ONE*, vol. 13, no. 10, p. e0203794, Oct. 2018, doi: 10.1371/journal.pone.0203794.
[19] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, and S. Mishra, ‘Detection of Cyberbullying Incidents on the Instagram Social Network’, *ArXiv150303909 Cs*, Mar. 2015, Accessed: Feb. 28, 2022. [Online]. Available: http://arxiv.org/abs/1503.03909
[20] ‘Analysis and detection of labeled cyberbullying instances in Vine, a video-based social network | SpringerLink’. https://link.springer.com/article/10.1007/s13278-016-0398-x (accessed Feb. 28, 2022).
[21] R. Sprugnoli, S. Menini, S. Tonelli, F. Oncini, and E. Piras, ‘Creating a WhatsApp Dataset to Study Pre-teen Cyberbullying’, in *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 51–59. doi: 10.18653/v1/W18-5107.