Effective Toxicity Prediction in Online Multiplayer Gaming: Four Obstacles to Making Approaches Usable

Julian Frommel
Utrecht University
Utrecht, Netherlands
j.frommel@uu.nl

Regan L. Mandryk
University of Saskatchewan
Saskatoon, Canada
regan@cs.usask.ca

ABSTRACT

Toxicity represents a threat to the safety and health of online multiplayer gaming communities. This has been recognized by industry, academia, and players and led to efforts for combating toxicity, including different approaches for predicting toxicity from behaviour. Despite promising results, such approaches have not yet been able to meaningfully combat toxicity at scale. In this position paper, we describe four obstacles that impede usable applied toxicity prediction in multiplayer games that could help to combat harm. We want to foster a discussion about how user-centered artificial intelligence approaches may help solve these obstacles.

KEYWORDS

toxicity, reporting, prediction, classification, multiplayer, esports, competitive, game, gaming

1 INTRODUCTION

There is wide agreement among developers [33], researchers [4], and players [28] that toxic behaviour in online games represents a danger for the safety and health of game communities. The term toxicity refers to a wide variety of negative and harmful behaviours, including harassment or abusive communication (e.g., see [1, 2, 4, 14, 24, 27, 37, 40]).

Toxic behaviours have negative ramifications, such as a negative influence on player experience [3, 14, 26, 40] and mood repair [7], causing psychological distress, rumination, and social withdrawal [15, 16, 28, 32, 35], decreased individual and team performance [31, 40], negative impact on game developer revenue [4, 20], and the potential to inflict gendered or racial trauma [16, 25]. Further, there is a cyclical nature to toxicity [20, 21], leading to a normalization of toxicity in gaming communities [2, 4] and highlighting the need to break this cycle through interventions.

One of the approaches essential for combating toxicity within online games is the prediction of toxicity, i.e., detecting whether behaviour is toxic or represents harassment. While academic research [5, 12, 29, 30, 34, 38, 39, 41] and industry stakeholders [8, 18, 36] have been tackling this challenge, it is as of yet unsolved. In this position paper, we will provide a short overview of toxicity prediction and four obstacles to solving the problem of enabling effective toxicity prediction, which is one aspect of a larger set of challenges we need to solve to combat toxicity at large.

We aim to foster a discussion about how *user-centered artificial intelligence* methods may contribute to valid and effective methods for predicting toxicity, which are essential for ensuring safe and healthy gaming environments, and may also provide benefits within other digital spaces, such as social media platforms.

2 PREDICTING TOXICITY

Many approaches have been proposed for predicting toxicity in digital spaces. Frequently, this task is tackled with supervised learning approaches that use text as input and predict whether a message is toxic or represents harassment. This has been a challenge for different digital spaces, such as Wikipedia talk page messages [12, 19], social media platforms [9, 29], and a variety of different multiplayer gaming environments [5, 17, 18, 36, 38]. Generally, artificial intelligence approaches are well-suited to deal with such tasks.

Previous work on predicting toxic behaviour in digital spaces has suggested the potential of different approaches to varying degrees. In non-game contexts, Liu et al. [29] were able to predict hostile *Instagram* comments with high performance (up to 84% AUC) and Zhang et al. [41] were able to predict personal attacks between *Wikipedia* editors with 64.9% accuracy. Dessi et al. [12] used deep learning approaches with long short-term memory models predicting if *Wikipedia* comments were toxic with an F1 score of up to 95.7%. In games, Blackburn and Kwak [5] achieved AUC scores of up to 79.9% for predicting if reported players in *League of Legends* matches were ultimately punished or pardoned. Stoop et al. [38] detected toxicity in *League of Legends* conversations, with F1 scores of up to 60.0%. Reid et al. [34] achieved were able to predict if Overwatch matches were toxic with up to 86.3% using features from in-game communication.

This is also tackled in commercial gaming environments. Machine-learning based systems are used in commercial games to detect and combat toxicity, such as in *Overwatch* [6, 8, 18] and *League of Legends* [22, 23], and on competition platforms like FACEIT [36]. Such prediction approaches are used for different meaningful interventions, e.g., to automatically detect and block slurs or to help detect and sanction offenders.

However, despite these encouraging results, toxicity and harassment remain a problem within online gaming, as is evident by recent statistics in 2021 showing that 83% of adult gamers experienced harassment in multiplayer games [28]. Thus, it is apparent that the promising research has not been applied at scale to practical solutions that lead to meaningful improvements.

We argue that toxicity prediction in online contexts is generally challenging, with additional complications arising within a gaming

Veröffentlicht durch die Gesellschaft für Informatik e.V.
in K. Marky, U. Grünefeld & T. Kosch (Hrsg.):
Mensch und Computer 2022 – Workshopband, 04.-07. September 2022, Darmstadt

https://doi.org/10.18420/muc2022-mci-ws12-315

context. In the next section, we propose four obstacles that we need to overcome if we want to enable valid and effective toxicity prediction. We aim to foster discussion around how user-centered artificial intelligence approaches may help address these obstacles.

3 OBSTACLES TO TOXICITY PREDICTION

While there are undoubtedly myriad challenges, we present four obstacles here that are essential to overcome for enabling valid and effective prediction methods in gaming environments.

3.1 Individual Differences

Individual differences in perceiving something as toxic or not represent an obstacle for training data that is used in supervised learning approaches. On one hand, it is challenging to implement models that predict toxicity at an individual level because it is difficult to assess information about how a specific individual perceives situations. Further, not enough information about individuals is accessible in an applied gaming context, e.g., because of privacy considerations in applied at-home settings. On the other hand, it is likely that generalized models that predict at a population level are not ideal for predicting on an individual level, because of a large variance in what different players consider toxic. This also applies to raters who often generate the ground truth labels used for training data. As such, it is unclear if generalized models are able to predict at an individual level. This can be quite problematic in this context.

For example, a model trained with one individual's labels may output predictions that differ from an individual's subjective assessment of a message, interaction, or situation. Thus, model predictions may sometimes consider data more toxic than a player, resulting in situation, in which an intervention that is triggered surprises a player who may even be opposed. In contrast, this can lead to an underestimation, i.e., a model predicts that a message is not toxic (based on average ground truth) but an individual may perceive it as problematic, which can lead to harm and distress.

Therefore, we need to account for individual differences in the perception of toxicity to implement valid and effective prediction methods.

3.2 Context

Context matters! This argument, which is widely accepted in HCI and related fields like ubiquitous computing, applies to the prediction of toxicity. This relates to subjectivity (e.g., what one player considers harmless could be offensive to another player), but also extends to the context in which the interaction happens. Two messages with the same content may be sent and perceived very differently based on the context, for example, someone being called a "dumbass", which may be quite offensive in many cases but a joking and snarky remark when uttered in a group of friends, who have known each other and think such talk is a hilarious and integral part of their interaction [17]. In the same way, a message may be perceived differently based on one's own mood (e.g., on a bad day) or the in-game context (e.g., after a mistake that leads to losing a game vs a funny but ultimately meaningless mishap).

Prediction of toxicity that can account for context should use methods that are more complex than simple keyword-based approaches because it is necessary to identify context, intent, and meaning [10]. The gaming context further complicates prediction because some toxic behaviours are less overt, such as subtle behaviours [4] or in-game behaviours like afking [11]. This means that no single approach may be easily used to account for all different types of toxicity.

3.3 Privacy

Privacy considerations are essential because most toxicity happens in communication between players, resulting in interpersonal communication data that is monitored and analyzed for the prediction of toxicity. This applies to communication in semi-public channels (e.g., team or match chats) but is even more problematic if analysis approaches are applied to non-public channels like direct messages, where players may have higher expectations of privacy and not think about others accessing and analyzing their communication. While it may appear unnecessary to apply toxicity prediction to non-public channels, it is evident that toxicity can also happen in such channels, suggesting that interventions and prediction methods also have benefits for such communication.

Thus, there is an evident conflict between the need to monitor and analyze communication to ensure safe environments and the privacy considerations. For example, we suggest that open and ethical use is essential, e.g., to communicate to players that data is analyzed and that data is not stored beyond the actual need for addressing toxicity.

3.4 Practical Application

Practical application is essential when considering the goal of ensuring safe and healthy environments, in which the majority of players is not regularly confronted with toxicity. This is not a methodological challenge but one of applying methods into the real world, associated with multiple aspects.

First, prediction methods are only useful if they are used and translate into meaningful action. This raises questions around how to enable this in an effective and practical way. For example, we need more research to understand what to do with predicted toxicity, e.g., which sanctions are suitable to effectively combat toxicity for good and at scale.

Second, it is difficult to translate academic findings into practice. Even the best methods are useless if they are not used, suggesting the need for increased collaboration efforts between academia and industry to make methods easily usable and applicable in popular and affected gaming environments.

Third, several approaches are used in commercial settings [8, 18, 36] but rarely there is more information available allowing these findings or methods to be used in other games. This is part of a larger problem, in which most of toxic behaviour happens on digital platforms of private companies (including games but also social media), which represent effectively private settings with considerations of intellectual property and competition [13] that lay in stark contrast with collaboration between developers, which would be necessary to combat toxicity. We cannot hope to combat toxicity at large if game developers consider 'toxicity' just one more factor, on which they can outperform their competitors.

Thus, we argue that further work is necessary to tear down these walls and to build bridges between all stakeholders who want to combat toxicity, including industry, academia, and game communities.

4 CONCLUSION AND OUTLOOK

We think that user-centered artificial intelligence methods may be useful for the context of predicting toxicity in gaming environments, due to its focus on integrating users into the process, which may be useful to account for *individual differences* (e.g., by better integrating players into the model creation), *context* (e.g., by incorporating context information into the models), considerations of *privacy* (e.g., by considering conservative data usage and non-permanent storage), and challenges of *practical application* (e.g., by participative development together with stakeholders). We hope to stimulate a discussion about how we can apply user-centered artificial intelligence approaches help solve these obstacles.

REFERENCES

[1] Leigh Achterbosch, Charlynn Miller, and Peter Vamplew. 2017. A taxonomy of griefer type by motivation in massively multiplayer online role-playing games. *Behaviour & Information Technology* 36, 8 (2017), 846–860. https://doi.org/10.1080/0144929X.2017.1306109

[2] Sonam Adinolf and Selen Turkay. 2018. Toxic Behaviors in Esports Games: Player Perceptions and Coping Strategies. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts (CHI PLAY '18 Extended Abstracts)*. Association for Computing Machinery, Melbourne, VIC, Australia, 365–372. https://doi.org/10.1145/3270316.3271545

[3] Jane Barnett, Mark Coulson, and Nigel Foreman. 2010. Examining Player Anger in World of Warcraft. In *Online Worlds: Convergence of the Real and the Virtual*. Springer, London, 147.

[4] Nicole A Beres, Julian Frommel, Elizabeth Reid, Regan L Mandryk, and Madison Klarkowski. 2021. Don't You Know That You're Toxic: Normalization of Toxicity in Online Gaming. In *Proceedings of CHI '21* (Virtual). ACM, 1–15.

[5] Jeremy Blackburn and Haewoon Kwak. 2014. STFU NOOB! predicting crowd-sourced decisions on toxic behavior in online games. In *Proceedings of WWW '14*. ACM, Seoul, Korea, 877–888. https://doi.org/10.1145/2566486.2567987

[6] Blizzard Entertainment. 2015. *Overwatch*. Game [PC]. Irvine, USA.

[7] Nicholas D Bowman and Ron Tamborini. 2012. Task demand and mood repair: The intervention potential of computer games. *New Media & Society* 14, 8 (2012), 1339–1357.

[8] J. Allen Brack and Blizzard Entertainment. 2020. 2020 Blizzard Fireside Chat. https://www.youtube.com/watch?v=CqPdqli0jTs.

[9] Margarita Bugueño and Marcelo Mendoza. 2019. Learning to detect online harassment on Twitter with the transformer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 298–306.

[10] Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.

[11] Joel Chapman and FACEIT. 2021. Fighting Abusive Behaviour — Product Update. https://medium.com/faceit-blog/fighting-abusive-behaviour-product-update-d39e490c00ce

[12] Danilo Dessì, Diego Reforgiato Recupero, and Harald Sack. 2021. An Assessment of Deep Learning Models and Word Embeddings for Toxicity Detection within Online Textual Comments. *Electronics* 10, 7 (March 2021), 779. https://doi.org/10.3390/electronics10070779 Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

[13] European Parliament Committee on Culture and Education Rapporteur: Laurence Farreng. 2022. Draft Report on E-sport and videogames, (2022/2027(INI)).

[14] Chek Yang Foo and Elina M. I. Koivisto. 2004. Defining grief play in MMORPGs: player and developer perceptions. In *Proceedings of the 2004 ACM SIGCHI International Conference on Advances in computer entertainment technology (ACE '04)*. Association for Computing Machinery, Singapore, 245–250. https://doi.org/10.1145/1067343.1067375

[15] Jesse Fox, Michael Gilbert, and Wai Yen Tang. 2018. Player experiences in a massively multiplayer online game: A diary study of performance, motivation, and social interaction. *New Media & Society* 20, 11 (2018), 4056–4073.

[16] Jesse Fox and Wai Yen Tang. 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society* 19, 8 (2017), 1290–1307. https://doi.org/10.1177/1461444816635778

[17] Julian Frommel, Valentin Sagl, Ansgar E. Depping, Colby Johanson, Matthew K. Miller, and Regan L. Mandryk. 2020. Recognizing Affiliation: Using Behavioural Traces to Predict the Quality of Social Interactions in Online Games. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–16.

[18] Iain Harris. 2020. Toxicity in Overwatch has seen an "incredible decrease" due to machine learning. https://www.pcgamesn.com/overwatch/toxic-behaviour-machine-learning.

[19] Jigsaw. 2017. Toxic Comment Classification Challenge. https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data.

[20] Bastian Kordyaka, Katharina Jahn, and Bjoern Niehaves. 2020. Towards a unified theory of toxic behavior in video games. *Internet Research* (2020).

[21] Yubo Kou. 2020. Toxic Behaviors in Team-Based Competitive Gaming: The Case of League of Legends. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*. Association for Computing Machinery, New York, NY, USA, 81–92. https://doi.org/10.1145/3410404.3414243

[22] Yubo Kou and Xinning Gui. 2021. Flag and Flaggability in Automated Moderation: The Case of Reporting Toxic Behavior in an Online Game Community. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 437, 12 pages. https://doi.org/10.1145/3411764.3445279

[23] Yubo Kou, Xinning Gui, Shaozeng Zhang, and Bonnie Nardi. 2017. Managing Disruptive Behavior through Non-Hierarchical Governance: Crowdsourcing in League of Legends and Weibo. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 62 (dec 2017), 17 pages. https://doi.org/10.1145/3134697

[24] Rachel Kowert. 2020. Dark Participation in Games. *Frontiers in Psychology* 11 (2020), 2969. https://doi.org/10.3389/fpsyg.2020.598947

[25] Jeffrey H. Kuznekoff and Lindsey M. Rose. 2013. Communication in multiplayer gaming: Examining player responses to gender cues. *New Media & Society* 15, 4 (June 2013), 541–556.

[26] Haewoon Kwak, Jeremy Blackburn, and Seungyeop Han. 2015. Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, Seoul, Republic of Korea, 3739–3748. https://doi.org/10.1145/2702123.2702529

[27] Noam Lapidot-Lefler and Azy Barak. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Computers in Human Behavior* 28, 2 (March 2012), 434–443. https://doi.org/10.1016/j.chb.2011.10.014

[28] Anti-Defamation League. 2021. Hate is no game: Harassment and positive social experiences in online games 2021. https://www.adl.org/hateisnogame#executive-summary

[29] Ping Liu, Joshua Guberman, Libby Hemphill, and Aron Culotta. 2018. Forecasting the Presence and Intensity of Hostility on Instagram Using Linguistic and Social Features. In *Twelfth International AAAI Conference on Web and Social Media*. Palo Alto, CA, USA, 181–190.

[30] Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *Network and Systems Support for Games (NetGames), 2015 International Workshop on*. IEEE, 1–6. https://doi.org/10.1109/netgames.2015.7382991

[31] C. K. Monge and T. C. O'Brien. 2022. Effects of individual toxic behavior on team performance in League of Legends. *Media Psychology* 25, 1 (2022), 82–105. https://doi.org/10.1080/15213269.2020.1868322

[32] Pew Research Center. 2014. Online Harassment. http://www.pewinternet.org/2014/10/22/online-harassment/.

[33] Shaun Prescott. 2017. Overwatch's Jeff Kaplan on toxic behavior: 'the community needs to take a deep look inwards'. https://www.pcgamer.com/overwatchs-jeff-kaplan-on-toxic-behavior-the-community-needs-to-take-a-deep-look-inwards/.

[34] Elizabeth Reid, Regan Mandryk, Nicole A. Beres, Madison Klarkowski, and Julian Frommel. 2022. "Bad Vibrations": Sensing Toxicity From In-Game Audio Features. *IEEE Transactions on Games* (2022), 1–10. https://doi.org/10.1109/TG.2022.3176849

[35] Kevin C Runions. 2013. Toward a conceptual model of motive and self-control in cyber-aggression: Rage, revenge, reward, and recreation. *Journal of youth and adolescence* 42, 5 (2013), 751–771.

[36] Maria Laura Scuri and FACEIT. 2019. Revealing Minerva and addressing toxicity and abusive behaviour in matches. https://blog.faceit.com/revealing-minerva-and-addressing-toxicity-and-abusive-behavior-in-matches-9073914a51c

[37] Cuihua Shen, Qiusi Sun, Taeyoung Kim, Grace Wolff, Rabindra Ratan, and Dmitri Williams. 2020. Viral vitriol: Predictors and contagion of online toxicity in World of Tanks. *Computers in Human Behavior* 108 (July 2020), 1–6. https://doi.org/10.1016/j.chb.2020.106343

[38] Wessel Stoop, Florian Kunneman, Antal van den Bosch, and Ben Miller. 2019. Detecting harassment in real-time as conversations develop. In *Proceedings of*

the Third Workshop on Abusive Language Online. ACL, Florence, Italy, 19–24. https://doi.org/10.18653/v1/W19-3503

[39] Joseph J Thompson, Betty HM Leung, Mark R Blair, and Maite Taboada. 2017. Sentiment analysis of player chat messaging in the video game StarCraft 2: Extending a lexicon-based model. *Knowledge-Based Systems* (2017). https://doi.org/10.1016/j.knosys.2017.09.022

[40] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–13. https://doi.org/10.1145/3313831.3376191

[41] Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations Gone Awry: Detecting Early Signs of Conversational Failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. ACL, Melbourne, Australia, 1350–1361. https://doi.org/10.18653/v1/P18-1125