AGGRESSION IN DIGITAL INTERACTIONS: THE EFFECT OF TOXICITY IN ONLINE GAMING COMMUNICATION

Iuliia Naidenova
HSE University, Russian Federation Perm

Petr Parshakov
HSE University, Russian Federation, Perm

Nikita Matkin*
HSE University, Russian Federation, Perm
namatkin@hse.ru

Funding: This article is an output of a research project implemented as part of the Basic Research Program at the HSE University.

ABSTRACT

This study analyzes the computer mediated text communication of non-professional video game players. The purpose of this study is to identify the impact of player communication toxicity on team performance. The dataset comprises 42,720 matches played between November 5 and November 18, 2015, including game statistics and chat messages. We use a BERT-model to classify toxicity of messages. Then regression analysis was used to estimate the impact of toxic behavior on winning probability and game satisfaction. The results indicate that the highest winning probability of a team corresponds to a low number of toxic messages in team communication, low average toxicity of communication, but high variation in toxicity levels. Furthermore, low levels of toxicity in communication can enhance overall game satisfaction.

Keywords: toxic behavior, team performance, logistic regression, Dota 2, esport.

1. INTRODUCTION

Communication plays an important role in teamwork. The dependence of teamwork on the communication methods employed by its members is undeniable: "Communication is the cornerstone of team interaction, without which teams would not be able to share information and knowledge, discuss and debate issues or strategies, or develop solutions to problems" (Hassal 2009, p. 16); "Issues such as leadership, collective efficacy, team cohesion, and group goal setting undoubtedly have great theoretical and practical value within sport teams. These issues all rely on one social process that may be the most important component of intra-team interactions. This process is communication" (Sullivan & Feltz 2003, p. 1693). In Letsky (2008), communication is regarded as team-level cognitive processing. Thus, the investigation of communication and its relationship with team performance is an essential area of study.

Globalization trends in business lead to distributed multinational teams that rely significantly on remote communication. Collaboration among team members across different locations is generally facilitated by digital tools such as email and chat. Chang, Chuang, & Chao (2011) argue that such virtual teams significantly differ from other small groups in terms of communication. Team sports, among many other activities, highlight the inherent human necessity for belonging to a larger whole. Similar to other activities, team sports have recently undergone a digital transformation. Competitive computer gaming has become popular worldwide. Moreover, it has now become a recognized professional sport (e-sports or cybersports) (Leavitt, Keegan, & Clark, 2016).

This study examines the role of aggression in amateur esports. By utilizing open data, this research analyzes the emergence of aggression

during gameplay. The anonymity inherent in online gaming environments eliminates the influence of sociocultural and individual factors on aggression, thereby allowing the analysis to focus exclusively on contextual and situational determinants (Spaaij & Schaillée, 2018). Yet the field of communication within computer gaming teams remains largely understudied.

Communication within computer game teams falls under the category of computer-mediated communication, defined as "communication that takes place through a variety of media and provides distributed group members with video, audio, and text-based messaging capabilities" (Harris & Sherblom, 2008, p. 296). Players utilize audio and text chats; however, the present research will focus solely on the latter. Computer-mediated communication enables messages anonymization, which can reduce players' sense of responsibility for their toxic behavior (Christopherson, 2007). Walther (2022) describes hate speech in social media as a prosocial phenomenon, motivated by "moral grandstanding, political derision as fun and peer support for interpersonal violence" (p. 1).

This behavior creates conditions of inequality such as disadvantages for certain players. Toxic communication refers to a special type of antisocial behavior, including harassment, griefing, cyberbullying and cheating (Kwak et. Al., 2015). This type of communication is especially common in online games. Despite the importance of teamwork in online games, some players engage in various strategies, including offensive language and bullying. Toxic communication creates an uncomfortable environment for players, but it can also be normalized. «Antisocial behaviors in online play directly harm player wellbeing, enjoyment, and retention—but research has also revealed that some players normalize toxicity as an inextricable and acceptable element of the competitive video game experience» (Beres et. al., 2021).

Computer gaming enables us to analyze toxic communication patterns and identify the role of toxic communication in team performance. The purpose of the current study is to investigate toxic team communication and its relationship with team performance, using data from non-professional computer game teams in Dota 2. The main research question of this study is what the relationship between toxic communication and game performance is. Unlike professional teams, these teams typically experience a high player rotation. Thus, player coordination relies mainly on in-game communication, rather than on formal training or previous experience. However, communication within computer game teams provides a valuable setting to investigate the effectiveness of text-based communication in the context of urgent decision-making.

The rest of the paper is organized as follows. The first section presents a literature review on the features of virtual teams, computer-mediated communication and the effects of toxicity in online games on team performance. Next, the data is described, along with relevant industry statistics. Following that, the methodology employed is explained. Finally, the empirical results, discussion, and conclusion sections complete the paper.

2. LITERATURE REVIEW

The phenomenon of virtual teams, in which members work together from a distance and primarily communicate via specific tools, has become very widespread across many spheres (Chang et al., 2011). Such teams can be multinational, culturally diverse, and work together for a short period of time for a particular project. The authors mention the necessity of adapting communication in virtual teams (Chang et al., 2011; Newlands, Anderson, & Mullin, 2003). Adachi, Hodson, and Hoffarth (2015) describe the relationship between intergroup competition and its association with heightened levels of prejudice and discrimination in online video games. Conversely, intergroup cooperation may contribute to a reduction in bias. Their findings suggest that cooperative interactions within online gaming environments serve as an effective mechanism for fostering both intergroup competition and cooperation in real-world contexts.

There is no single, clear definition of toxic behavior in the literature, as different communities within various games may define such behavior differently (Grandprey-Shores et. al., 2014). Online game studies typically focus on specific types of toxic behavior, including harassment, griefing, cyberbullying, and cheating. The authors of the Jigsaw Unintended Bias Toxicity Classification

dataset, which was used to train the model in this study, define toxicity as "*rude, disrespectful or otherwise likely to make someone leave a discussion*" (Jigsaw Unintended Bias in Toxicity Classification, n.d.). In verbal communication, toxicity manifests as hostile comments and harassment (Sun, Yu, & Chen, 2024). In the game context, the main purpose of toxic behavior may be the desire to force a player to leave the match.

Toxic communication in online games is widely studied in various disciplines, including Psychology (Liu & Agur, 2023), Cultural Studies (Bush et. al., 2015), Economy (Huston et. al., 2023; Nexø, 2024), Linguistics (Kwak & Blackburn, 2015). These diverse approaches employ a wide range of methods, including qualitative methods such as interviews and surveys, and quantitative methods like crowdsourced data analysis and statistical analysis of gaming activity.

The growing interest in this topic is evident from the increasing number of publications in recent years. For example, toxic behavior in online games could be a predictor of mental toughness (González & Resett, 2023). Kordyaka, Park, Krath and Laato (2023) examine toxic behavior in online games based on the individuals' social status. Wijkstra, Rogers, Mandryk, Veltkamp and Frommel (2023) provide systematic literature review, which states that "most systems intervene only after toxicity occurs with few interventions that act before toxicity".

The correlation between toxic behavior and team performance has been studied in various games, such as League of Legends (Monge & O'Brien, 2022; Neto, Yokoyama & Becker, 2017), For Honor (Canossa et. al., 2021), and Fortnite (Gandolfi et. al., 2023). These studies have established a negative correlation between toxic behavior and performance Although League of Legends and Dota 2 are quite similar, researchers highlight differences in moderation system (Wang, 2023) and game dynamics (Winn, 2015). However, no study has yet investigated the relationship between toxic behavior and performance specifically in Dota 2.

Monge & O'Brien (2022) indicates two approaches to this relationship. First, self-determination theory posits that poor performance worsens behavior (Ryan & Deci, 2000). Deci, Przybylski, Rigby and Ryan (2014) point out that when individuals' needs for competence, relatedness, and autonomy are threatened, it leads to frustration and aggression. If teammates perform poorly, it is expected to evoke anger and other destructive emotions in other team members (Carver & Harmon-Jones, 2009).

The second approach is the "bad apple" perspective which posits that online toxic behavior worsens performance (Byington, Felps, & Mitchel, 2006). Felps et al. suggest that even a single group member engaging in toxic behavior can disrupt the performance of the entire group or team. These undesirable behaviors include withholding effort, expressing negative emotions, and violating interpersonal norms. These behaviors create a negative environment within the team, such as feelings of inequity, negativity, and reduced trust. These negative states disrupt performance by prompting withdrawal and behavioral retaliation.

The third group of theories describes the normalization of toxic behavior in an online environment. Cook, Schaafsma and Antheunis (2018) suggested that negative behavior has become a normative part of gaming culture. "Trolling is, in fact, a normal, expected event, sometimes even described as a rite of passage. No one escapes it, and it thus becomes a shared, common experience between gamers, cementing the community" (p. 3337). Adinolf and Turkay (2018) established that collegiate esports clubs recognized negative behavior as endemic and promoted the development of a competitive culture. Beres et al. (2021) produced evidence that moral disengagement and toxic online disinhibition both contribute to a reduced perception of toxicity of such behaviors. As players learn to normalize negative behavior, its correlation with performance may weaken.

3. Methodology

3.1 Context and data

Dota 2, despite being released in 2013, remains the second most played multiplayer game on Steam, with 330,210 players per day according to SteamDB (Most Played Multi-Player Games, 2024). Dota 2 is a multiplayer online battle arena (MOBA) video game. In Dota 2, two teams of five players compete in matches, with

each team protecting its own base on the map. Each player controls a distinct "hero" character with special abilities and play styles, adding variety and strategy to the game. There are two opposing factions in Dota2: the Radiant and the Dire. There is a set of hero characters affiliated with each faction. Additionally, specific areas of the map with distinct landscapes and features are designated for each faction: the Radiant controls the bottom left half of the map, while the Dire claims the top right portion. In games involving non-professional players, teams form randomly, but players who queue with a group of friends will be placed on the same team. There are two types of matches. Ranked matches influence players' matchmaking rating (MMR), which indicates their skill level. Non-ranked matches do not influence the rating. During the game, players can send messages to their team or to all players in chat. All messages are saved in a game data dump. The game is available as free-to-play.

Our dataset comprises 50000 ranked ladder matches from the Dota 2 data dump, alongside 1439488 chat messages (provided by OpenDota) representing player communications within each match.. The dataset contains games from November 5 to November 18, 2015[1]. We utilize information on match results and players' intra-team text chat. Neither voice chat nor inter-team text chat is available. For each match we have all symbols each team member types in text chat. We measure overall communication intensity by the total number of messages in chat, while "side intensity" is defined as the number of messages sent by players on each respective team. Subsequently, the data was aggregated at the team-match level. Toxic intensity is defined as the number of toxic messages sent per match per team during the game. The mean toxicity per team was calculated as the number of toxic messages per match divided by the total number of messages sent by that team in that match.

Table 1 reports summary statistics of our sample, which consists of 42,720 Dota 2 matches, encompassing both Dire and Radiant sides. Matches played with non-English textual communication were excluded from the dataset. The win metric, representing the binary outcome of matches, has a mean value of 0.50, indicating an equal distribution of wins and losses across the dataset. The average number of messages per team is 16. The average number of words per message is 2.4.

Toxic intensity has a mean of 0.93 with a relatively high standard deviation of 1.82, indicating significant variability in the levels of toxicity across matches. This metric's values range from 0 to 66, highlighting extreme cases of in-game toxicity. Conversely, the toxic mean, reflecting the average level of toxicity per match, is 0.05 (5%). Duration, quantifying match lengths in seconds, averages at 2,486.12 seconds (approximately 41 minutes). Additionally, we created a dummy variable for long matches, which is equal to one for matches longer than the median match duration. Side intensity is a metric representing the number of messages sent by each team.

The mean of the toxic standard deviation metric, illustrating the variability in toxicity levels within matches, is 0.13, indicating overall fluctuations in toxicity levels across matches. The votes metric, representing players' game satisfaction, exhibits a mean close to zero (0.02) with a substantial standard deviation of 0.69. Game satisfaction is measured by votes: a player's negative vote is -1, and a positive vote is +1. This metric's range from -10 to 66 reflects a broad spectrum of player experiences and opinions. Toxicity at the start of the match is a binary variable, assigned a value 1 if toxic behavior occurred within the first 120 seconds of the match. Toxic mean after first blood indicates the average level of toxicity during the 60 seconds following the first kill in the game.

Table 2 presents the correlation matrix of the main variables. Toxic intensity is moderately correlated with communication intensity, toxic mean, and toxic standard deviation, as these metrics are derivatives of one another. Toxicity at the start of the match has a weak positive correlation with toxic mean after first blood, which may indicate a relationship between these two metrics.

[1] It was inspired by the Dota 2 Matches data published here by Joe Ramir. This dataset is available on Kaggle: https://www.kaggle.com/devinanzelmo/dota-2-matches/home.

Table 1. Summary statistics of the main variables

Statistic	N	Mean	St. Dev.	Min	Max
win	85,440	0.50	0.50	0	1
Toxic intensity	85,440	0.93	1.82	0	66
Duration (sec.)	85,440	2,486.12	638.99	59	16,037
Communication intensity	85,440	15.97	15.76	2	363
Toxic mean (%)	85,440	0.05	0.08	0.00	1.00
Toxic std	85,440	0.13	0.17	0.00	0.71
Long match	85,440	0.50	0.50	0	1
Votes	85,440	0.02	0.69	-10	66
Toxicity at the start of the match	85,440	0.05	0.22	0	1
Toxic mean after first blood	85,440	0.007	0.07	0	1

Table 2. Correlation matrix of the main variables

	Win	Toxic intensity	Duration	Communi-cation intensity	Toxic Mean	Toxic std	Votes	Toxicity at the start of the match	Toxic mean after first blood
Win	1								
Toxic intensity	-0.035	1							
Duration	0	0.043	1						
Communi-cation intensity	0	0.538	0.068	1					
Toxic Mean	-0.067	0.593	0.015	0.101	1				
Toxic std	-0.05	0.638	0.024	0.23	0.909	1			
Votes	0	0	0	0.003	-0.002	-0.001	1		
Toxicity at the start of the match	0.001	0.276	-0.01	0.175	0.225	0.27	0	1	
Toxic mean after first blood	-0.003	0.136	-0.005	0.056	0.124	0.129	-0.002	0.3	1

3.2 Research Methods

Games generate rich digital data, which allows for automated analysis of social dynamics within a party. Identifying toxic patterns within a game is necessary to further improve performance. Therefore, further research is needed on developing complex automated methodologies that apply neural networks and statistical analysis to big data. The novelty of the approach lies in establishing the relationship between toxic behavior and performance, based on statistics from Dota 2 in-game data.

We used DistilBERT, pre-trained by Martin Pan, to classify toxic comments. Messages in Dota 2 often consist of short phrases and slang, similar to comments on Social Networks. This model is available on HuggingFace: https://huggingface.co/martin-ha/toxic-comment-model. The model was trained and evaluated on Jigsaw Unintended Bias in Toxicity Classification dataset from an archive of the Civil Comments platform (Jigsaw Unintended Bias in Toxicity Classification, n.d.). This dataset contains 500000 labeled online comments from 50 English-language news sites between 2015 and 2017. Given that the model was trained on online communication data, it is suitable for classifying chat within a computer game. The model classified 82641 toxic messages (approximately 5%) and 1356847 non-toxic messages (approximately 95%) within our dataset. Subsequently, we filtered the data to include only messages containing Latin characters. The most frequent non-toxic comments include "gg" (81859), "lol" (36949), "?" (19695). The most frequent toxic comments are "wtf" (4564), "fuck" (1335), "fuck you" (786), "idiot" (765).

We address the issue of toxic behavior and its relationship to team performance. For this purpose, we employ logistic regression with the match outcome (win/loss) as the dependent variable. Therefore, our results

should be interpret in terms of correlation, not causation.

$$p(win_{ij}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot toxicity_{ij} + CV_{ij})}}$$

where *win* is a dummy indicator of a win by team *i* in match *j*. *Toxicity* represents a set of measures related to a team's toxic communication, specifically including: *toxic intensity* defined as the number of toxic messages sent by the team; *toxic mean*, calculated as the number of toxic messages per match divided by the total messages for that team in the match (as described above); *toxic standard deviation (toxic std)*, which quantifies the variability in toxicity levels among team's messages. *CV* comprises control variables such as *Communication Intensity, Team Side, Duration,* and *Long match. Communication Intensity* quantifies the number of messages exchanged per team. *Team Side* is a binary variable that takes the value of 1 for teams on the Radiant side and 0 for those on the Dire side. *Duration* represents the match length in seconds. *Long Match* is a dummy variable for matches that exceeded the median duration.

To analyze the impact of communication toxicity on game satisfaction, we c onstructed a linear regression model incorporating a similar set of variables of interest and control variables. The dependent variable, *vote,* represents the difference between the number of positive votes and negative votes cast by team *i* in match *j*. Thus, we have the following equation:

$$votes_{ij} = \beta_0 + \beta_1 \cdot toxicity_{ij} + CV_{ij} + \varepsilon_{ij}$$

4. RESULTS

4.1 Empirical Results

Table 3 presents the results of logit models analyzing the impact of different toxicity metrics on the likelihood of winning in Dota 2 matches. Two models are presented, each with *win* as the dependent variable, but incorporating different toxicity metrics to assess their influence on match outcomes.

Model 1 explores the effects of toxic intensity. Toxic intensity negatively affects the win probability, with a coefficient of -0.09 and average marginal effect of -0.0216, suggesting that higher toxicity levels within a team are associated with lower chances of winning. Figure 1 illustrates the marginal effect of toxicity on winning probability. Regarding control variables, side intensity shows a positive relationship with winning, while the

Table 3. Logit regression results of match winning probability

	Dependent variable:		
	win		
	(1)	**(2)**	**(3)**
Toxic intensity	-0.09***		
	(0.01)		
Toxic mean		-3.13***	-1.90***
		(0.23)	(0.12)
Toxic std		0.73***	
		(0.11)	
Long match			-0.01
			(0.02)
Duration	-0.0000	-0.0000	-0.0000
	(0.0000)	(0.0000)	(0.0000)
Communication intensity	0.01***	0.002***	0.004***
	(0.001)	(0.0005)	(0.0004)
Team side	0.16***	0.16***	0.16***
	(0.01)	(0.01)	(0.01)
Toxic mean: Long match			0.31*
			(0.17)
Constant	-0.13***	-0.05*	-0.04
	(0.03)	(0.03)	(0.04)
Observations	85,440	85,440	85,440
Log Likelihood	-58,984.89	-58,907.30	-58,926.82
Akaike Inf. Crit.	117,979.80	117,826.60	117,867.60
Note:			*p<0.1; **p<0.05; ***p<0.01

duration of the match appears to have a negligible effect. Being on the Radiant side is associated with a higher probability of winning, highlighting a potential side advantage in the game.

Model 2 includes toxic mean and toxic standard deviation instead of toxic intensity. Here, *toxic* mean exhibits a significant negative association with winning, while toxic standard deviation shows a positive relationship. This suggests that not just the average level of toxicity but its variability within the team influences match outcomes, with greater variability being positively associated with winning. The effects of *side intensity* and being on the Radiant side remain positive and significant, albeit with a slight adjustment in the coefficients for side intensity compared to Model 1. Model 3 shows that the size of the toxicity effect differs for *long matches*, Figure 2 illustrates this.

We further analyze how toxicity affects player perception of the game. Table 4 presents result for the *votes metric* as the dependent variable. In the first model, all factors except constant term show a negligible impact on *votes*, with coefficients close to zero and not statistically significant. In the second nodel, we found significant relationships for *toxic mean* and *toxic standard deviation* with number of *votes*, both at the 5% significance level. The *toxic mean* negatively influences *votes*, while *toxic standard deviation* has a positive effect. These findings suggest that while higher average toxicity levels are associated with fewer votes, greater variability in toxicity levels within matches correlates with an increase in votes.

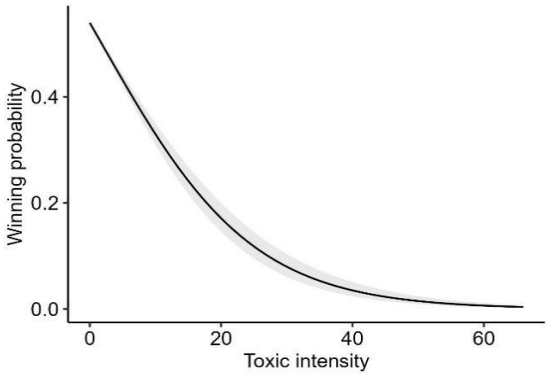

Figure 1. Toxic intensity and predictions of winning probability

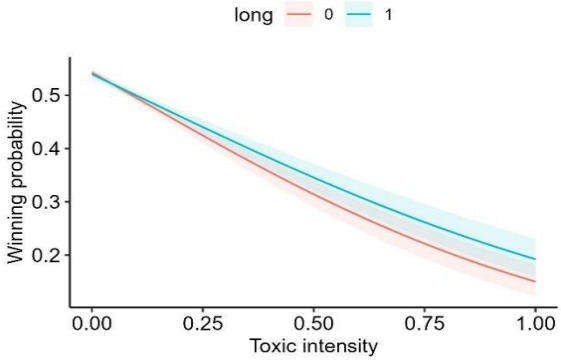

Figure 2. Toxic intensity and predictions of winning probability by match duration

Table 4. OLS results for votes metric as the dependent variable

	Dependent variable:	
	votes	
	(1)	(2)
Toxic intensity	-0.001	
	(0.001)	
Toxic mean		-0.05**
		(0.02)
Toxic std		0.03**
		(0.01)
Duration	0.0000	0.0000
	(0.0000)	(0.0000)
Communication intensity	0.0001	-0.0000
	(0.0001)	(0.0001)
Team side	0.0000	0.0000
	(0.002)	(0.002)
Constant	0.01*	0.01*
	(0.003)	(0.003)
Observations	85,396	85,396
R^2	0.0000	
Adjusted R^2	-0.0000	
Log Likelihood		1,504.77
Akaike Inf. Crit.		-2,997.53
Residual Std. Error	0.24 (df = 85391)	
F Statistic	0.73 (df = 4; 85391)	
Note:	*p<0.1; **p<0.05; ***p<0.01	

4.2 Robustness check

In this section, we conduct a robustness check to ensure that our results are not influenced by endogeneity, where the toxicity observed during the game could be a consequence rather than a cause of game dynamics. To address this concern, we re-estimate our regression model by focusing only on toxicity at the start of the game (within the first 5 minutes). By doing so, we aim to capture toxicity that is more likely to reflect inherent

player personality traits rather than reactions to in-game events.

As shown in Table 5, the coefficient for the toxicity at the start of the game is statistically significant and negative, consistent with our main results. This indicates that early-game toxicity has a detrimental effect on the probability of winning, supporting our hypothesis that toxic behavior is harmful to team performance.

Additionally, we include a variable representing toxicity after the "first blood" event (the first kill in the game). The results show that this measure of toxicity is not statistically significant, suggesting that the toxic behavior following significant game events does not significantly affect the match outcome. This further reinforces the conclusion that our primary findings are robust and not driven by game events, thereby reducing concerns about endogeneity in our analysis.

These findings confirm that the negative impact of toxicity on team performance is not merely a reflection of how the game evolves but is linked to the players' intrinsic behaviors from the start of the match.

Table 5. Logit regression results for the starting toxicity

	Dependent variable:	
	win	
	(1)	(2)
Toxicity at the start	-0.07**	
	(0.03)	
toxic mean after first blood		-0.13
		(0.09)
Duration	-0.0000	-0.0000
	(0.0000)	(0.0000)
Communication intensity	0.003***	0.003***
	(0.0004)	(0.0004)
Team side	0.15***	0.15***
	(0.01)	(0.01)
Constant	-0.11***	-0.11***
	(0.03)	(0.03)
Observations	85,440	85,440
Log Likelihood	-59,138.81	-59,140.96
Akaike Inf. Crit.	118,287.60	118,291.90
Note:	*p<0.1; **p<0.05; ***p<0.01	

5. CONCLUSION

This study focuses on the analysis of text-based communication among members of virtual teams within the context of online Dota2. We consider non-professional players.

The results show that the intensity of toxic communication negatively affects the likelihood of winning. Conversely, the intensity of communication shows a positive relationship with winning. We can conclude that communication should be active, but not toxic. This result confirms the results of Monge & O'Brien (2022). This result corresponds to the hypotheses and findings of other authors who posit a negative relationship between toxic behavior and win rate. However, the variability in communication toxicity has a positive effect on the likelihood of winning. Thus, certain toxic messages have the potential to either inspire team members or discourage opponents, thereby helping the team win.

Toxic behavior also influences match satisfaction. There is no discernible effect of the number of toxic messages, but the average level of toxicity has a positive impact and variability in toxicity has a negative impact on match satisfaction, mirroring the effects observed for winning. Thus, toxic communication predominantly harms the overall playing experience.

In addition, other factors influencing the probability of winning were identified, such as communication intensity and playing on the Radiant side. These factors show a positive correlation with the probability of winning. Intense communication can contribute to team coordination or might indicate a relatively easy game where players have more time to communicate via text.. The significant effect of playing on a particular side requires further investigation as it potentially indicates imbalance in the game.

However, it is important to note that our analysis is subject to certain limitations. A significant limitation of our study is the AI classifier used to detect toxicity. As noted by the model's author, it performs poorly for some comments that mention a specific identity subgroup. The potential bias is evident in the fact that the model is more likely to interpret messages containing words like "jew", "muslim", "black", "white", or

"gay" as toxic messages. Our analysis showed a relatively low frequency of these specific words in our dataset, including "jew" (24 messages), "muslim" (75), "white" (55), and "black" (214). The word "gay" is used 1297 times among toxic messages; however, these messages are often genuinely toxic, for example, "stop gay", "hey gays" or "this is a gay hero". Despite this fact, this inherent model bias could have affected the toxicity classification and, consequently, the results of our regression analysis. Furthermore, while the model demonstrates high overall accuracy (94%), its f1-score is 0.59. In the context of imbalanced datasets such as ours (~5% of toxic messages), an f1-score of 0.59 suggest that model can produce a notable number of false positive or false negative mistakes. This bias can lead to a misestimation the true toxicity level of communication within team interactions and potentially distort the identified relationship between toxicity and performance.

Specifically, since we are focusing on non-professional level matches, there is limited player-specific data available. As a result, our ability to identify and control for potential factors that may impact team performance is restricted. Consequently, it is advisable not to draw definitive conclusions about causal relationships based on our findings. Moreover, our sample consists of non-professional matches; therefore, the results can be consciously generalized on professional esports players.

6. REFERENCES

Adachi, P. J., Hodson, G., & Hoffarth, M. R. (2015). Video game play and intergroup relations: Real world implications for prejudice and discrimination. *Aggression and violent behavior*, *25*, 227-236.

Adinolf, S., & Turkay, S. (2018, October). 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* (pp. 365-372).

Beres, N. A., Frommel, J., Reid, E., Mandryk, R. L., & Klarkowski, M. (2021, May). Don't you know that you're toxic: Normalization of toxicity in online gaming. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-15).

Busch, T., Boudreau, K., & Consalvo, M. (2015). Toxic gamer culture, corporate regulation, and standards of behavior among players of online games. In *Video game policy*, 176-190. Routledge.

Canossa, A., Salimov, D., Azadvar, A., Harteveld, C., & Yannakakis, G. (2021). For honor, for toxicity: Detecting toxic behavior through gameplay. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CHI PLAY), 1-29.

Carver, C. S., & Harmon-Jones, E. (2009). Anger is an approach-related affect: evidence and implications. *Psychological bulletin*, 135(2), 183.

Chang, H. H., Chuang, S.-S., & Chao, S. H. (2011). Determinants of cultural adaptation, communication quality, and trust in virtual teams' performance. *Total Quality Management & Business Excellence*, 22(3), 305–329. https://doi.org/10.1080/14783363.2010.532319

Christopherson, K. M. (2007). The positive and negative implications of anonymity in Internet social interactions:"On the Internet, Nobody Knows You're a Dog". *Computers in Human Behavior*, *23*(6), 3038-3056.

Cook, C., Schaafsma, J., & Antheunis, M. (2018). Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, 20(9), 3323–3340. https://doi.org/10.1177/1461444817748578

Felps, W., Mitchell, T. R., & Byington, E. (2006). How, When, and Why Bad Apples Spoil the Barrel: Negative Group Members and Dysfunctional Groups. *Research in Organizational Behavior*, 27, 175–222. https://doi.org/10.1016/S0191-3085(06)27005-9

Gandolfi, E., Ferdig, R. E., Krause, K., Copus, A., Ostrowski-Delahanty, S., & Alemagno, S. (2023). Problematic Gaming at a Crossroad: Exploring the Interplay Between Internet Gaming Disorder, Toxic Attitudes, and Empathy in Digital Entertainment. *Games and Culture*, 15554120231211991.

Huston, C. Y., Cruz, A. G. B., & Zoppos, E. (2023). Normalizing the Toxic Consumer Subject: Sustaining Neoliberal Logics Within Online Gaming. *Journal of Macromarketing*, 43(4), 447-459.

González Caino P and Resett S. (2024). Toxic Behavior and Tilt as Predictors of Mental Toughness in League of Legends Players of Argentina. *HCI International 2023 – Late Breaking Posters*. 10.1007/978-3-031-49215-0_55. (464-468).

Grandprey-Shores, K., He, Y., Swanenburg, K. L., Kraut, R., & Riedl, J. (2014, February). The identification of deviance and its impact on retention in a multiplayer game. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 1356-1365).

Jigsaw Unintended Bias in Toxicity Classification. (n.d.). Kaggle. https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification

Kordyaka, B., Park, S., Krath, J., & Laato, S. (2023). Exploring the relationship between offline cultural environments and toxic behavior tendencies in multiplayer online games. *ACM Transactions on Social Computing*, 6(1-2), 1-20.

Kwak, H., Blackburn, J., & Han, S. (2015, April). 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* (pp. 3739-3748).

Kwak, H., & Blackburn, J. (2015). Linguistic analysis of toxic behavior in an online video game. In Social Informatics: SocInfo 2014 International Workshops, Barcelona, Spain, November 11, 2014, Revised Selected Papers 6 (pp. 209-217). Springer International Publishing.

Nexø, L. A. (2024). Toxic Behaviors in Esport: A Review of Data-Collection Methods Applied in Studying Toxic In-Gaming Behaviors. International Journal of Esports, 3(3).

Leavitt, A., Keegan, B. C., & Clark, J. (2016). Ping to Win?: Non-Verbal Communication and Team Performance in Competitive Online Multiplayer Games. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16* (pp. 4337–4350). Santa Clara, California, USA: ACM Press. https://doi.org/10.1145/2858036.2858132

Letsky, M. P. (Ed.). (2008). Macrocognition in teams: theories and methodologies. Aldershot, Hants, England ; Burlington, VT: Ashgate.

Liu, Y., & Agur, C. (2023). "After all, they don't know me" Exploring the psychological mechanisms of toxic behavior in online games. *Games and Culture*, 18(5), 598-621.

Monge, C. K., & O'Brien, T. C. (2022). Effects of individual toxic behavior on team performance in League of Legends. *Media Psychology*, 25(1), 82-105.

Most played Multi-player games. (2024, August 20). SteamDB. Retrieved August 20, 2024, from https://steamdb.info/charts/?category=1

Neto, J. A., Yokoyama, K. M., & Becker, K. (2017, August). Studying toxic behavior influence and player chat in an online video game. In *Proceedings of the international conference on web intelligence* (pp. 26-33).

Przybylski, A. K., Deci, E. L., Rigby, C. S., & Ryan, R. M. (2014). Competence-impeding electronic games and players' aggressive feelings, thoughts, and behaviors. *Journal of Personality and Social Psychology*, 106(3), 441–457. https://doi.org/10.1037/a0034820

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Spaaij, R., & Schaillée, H. (2019). Unsanctioned aggression and violence in amateur sport: A multidisciplinary synthesis. *Aggression and violent behavior*, *44*, 36-46.

Sullivan, P. J., & Short, S. (2011). Further Operationalization of Intra-Team Communication in Sports: An Updated Version of the Scale of Effective

Communication in Team Sports (SECTS-2): EFFECTIVE COMMUNICATION IN SPORTS TEAMS. *Journal of Applied Social Psychology*, 41(2), 471–487. https://doi.org/10.1111/j.1559-1816.2010.00722.x

Sun, X., Yu, V., & Chen, V. H. H. (2024). Toxic behavior in multiplayer online games: the role of witnessed verbal aggression, game engagement intensity, and social self-efficacy. *Chinese Journal of Communication*, 1-19.

Walther, J. B. (2022). Social media and online hate. *Current Opinion in Psychology*, *45*, 101298.

Wang, Q. (2023). A Comparison of Moderation Systems in DOTA2 and League of Legends from a Player Perspective.

Wijkstra, M., Rogers, K., Mandryk, R. L., Veltkamp, R. C., & Frommel, J. (2023, October). Help, My Game Is Toxic! First Insights from a Systematic Literature Review on Intervention Systems for Toxic Behaviors in Online Video Games. In *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play* (pp. 3-9)

Winn, C. (2015, January). The well-played moba: How dota 2 and league of legends use dramatic dynamics. In *Proceedings of DiGRA 2015 Conference*.

***