DEVELOPING AN IMPROVED CHAT FILTERING SYSTEM

IDENTIFYING MOTIVATIONS BEHIND PLAYER TOXICITY IN COMPETITVE SETTINGS

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Parth Raut

November 1, 2021

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Kathryn A. Neeley, Department of Engineering and Society

Daniel G. Graham, Department of Computer Science

Introduction

Chat filtering is a common practice in video games to try to combat toxicity and harassment and protect the youth from harmful language online. As of now, the method of filter messages in game chat is just searching for specific words/phrases in each chat message. There are other unused methods for chat filtering such as Rough Set Theory (Roy, 2013, n.p.), a mathematical approach to filtering text, or pattern and list recognition (Martens, 2015, p. 2), but they are more complex implementations, so developers tend to shy away from using them in their games. However, the problem with the current implementation of chat filtering is that it does not take into account the intent of player messages. As a result, game developers are not able to accurately identify toxic behavior in game chat through the chat filter and some players go unpunished while others are falsely punished. Unfortunately, there are some consequences to this problem. Toxicity from players causes other players to feel frustrated and have a bad experience in game (Neto, 2017, p. 1). These players who have become frustrated and unhappy with their experience in the certain games may leave them and thus reduce the quality of other players' experiences in those games because of the lack of players who are left playing the game (Neto, 2017, p. 1). To resolve this, I will develop and analyze a chat filtering system that takes into account the intent of player messages and appropriately identifies toxic and non-toxic behavior as part of my technical topic. Additionally, I will conduct research on player behavior and why players behave in a toxic manner as part of my STS topic.

Technical Topic: Developing an Improved Chat Filtering System

Currently, video games use a chat filtering system that looks for specific words and phrases (without considering the context of the message) in chat messages to flag them for toxic behavior and/or censor them. When this happens, the developers can choose what happens to the

player. One option is to just censor the message, warn the player of their inappropriate language, and if the problem persists, temporarily or permanently ban the player's account. The second option is flagging the message for review and a developer will later review the message to determine if the message was toxic or not. In each case, the existing chat filter system attempts to reduce toxicity by taking action against the offending player. It is important to ensure that players have a good experience while playing the game.

Unfortunately, while the current chat filtering system does reduce toxicity, it also flags messages that may contain a banned word or phrase but not have any toxic intention, such self-deprecating remarks (Martens, 2015, p. 3). Because of this, messages are incorrectly flagged for inappropriate wording even if the message as a whole was not intended to be that way. For example, Martens (2015, p. 3) states that messages containing words such as "crap" have a negative connotation but are not considered an insult, so filtering for messages that have toxic meaning is important. Furthermore, players who have no intention of sending a toxic message in the game chat are punished as a result of this incorrect flagging of messages. In the case of the first option, players are wrongfully punished and may lose access to their account if the chat filter deems their message to be toxic. In the case of the second option, developers would need to spend extra time and resources to review falsely-flagged messages that are not actually toxic. This can be seen in Kwak's research, where he uses 590,311 Tribunal cases from the game League of

Legends, where each Tribunal case is a single player who was accused of toxicity in up to five matches (Kwak, 2014, p. 2). This number is only a subset of the total number of Tribunal cases and doesn't include other games.

category	description	rules	examples	precedence	unique count
nonlatin	special character, foreign language	pattern	文章	500	20133
praise	acts of courtesy, kindness, sport spirit or gratitude	list	gj, gg, thx, hf	100	295
bad	profanity, swear words, inappropriate language	list, letterset	noob, idiot, f*	90	4881
laughter	acronyms expressing laughter	letterset	HAHAHAHA, lol, ROFL	60	2158
smiley	emoticons, symbols resembling faces or emotions	pattern, list	:D, ;), oO, -_-	50	1110
symbol	symbols or numbers	pattern	?, 1, ..., ??!?, /	40	3181
slang	DotA-specific game-technical terms, used to coordinate with team	list	ursa, mid, back, farm, bkb	30	10046
command	in-game commands, control words to trigger certain effects	pattern	!ff, !pause, -swap	20	2513
stop	English stop words	list	was, i, it, can, you	10	1322
timemark	automatically generated time-stamps, prepended in pause-mode	pattern	[00:05], [01:23]	5	223

Figure 1: Word/Phrase Annotation Categories and their Frequency (Martens, 2015, p. 3)

Using the work of previous research projects, I will develop a chat filter that analyzes player messages and accounts for player intention when filtering through the messages. Martens et. al. developed model for pattern, list, and letterset recognition from chat logs from the game DotA as seen in Figure 1 and categorizes the messages into categories such as "praise" (acts of courtesy, kindness, etc.) or "bad" (profanity, swear words, etc.). Through their research, Martens et. al. found that their letterset recognition analysis, where the "set of letters of the word equals the set of letters of a word from a pre-defined list" introduced a negligible amount of false positives from their dataset. Additionally, Kwak et. al. analyzed when certain words or phrases were said during a match and which types of words were used by toxic players. This method of analyzing messages adds more "context" to what the player meant when they sent the message. I will combine the two methods presented by Martens et. al. and Kwak et. al. to develop a chat filter that can recognize when a player is being toxic and when they are not.

STS Topic: Identifying Motivations Behind Player Toxicity in Competitive Contexts

Toxic players are present in every game, regardless of the actions taken to prevent them from ruining other players' experiences. About 25% of customer support calls to video game companies from players of their respective games are on the topic of toxicity (Blackburn, 2014,

p. 1). Usually, the punishments of the toxic players are carried out without much question of why they were behaving in the way that they were. While it is important to take preventive measures and actions against toxic players, it is equally important to understand why they behave in the way that they do. It is better to understand the root of the problem, which is why players behave in a toxic manner, than to blindly come up with solutions to preventing toxicity. Unfortunately, the research that has been done on analyzing player toxicity doesn't completely encompass the reason for their behavior (Neto, 2017, p. 1). Classification for the reasons that players play games and the types of interactions between players have been identified in previous works, but the psychology behind toxic player intentions is not concrete. For example, player motivation for video games can be classified as play as power, play as progress, play as fantasy, and play as self (Lin, 2005, p. 3). Some toxic players fall into the category of play as power, where the player does whatever they want but at the expense of other players (Lin, 2005, p. 3), but this doesn't explain the reasoning behind toxic players in a competitive setting (Neto, 2017, p. 1). Fully understanding why players are toxic is essential because, if that understanding is not developed, it may be difficult to reduce toxicity in all games, decreasing the quality of the average player's experience. As it is, toxic players are difficult to identify because the intent behind player actions and player messages are often unknown and taken out of context (Lin, 2005, p. 3). It is difficult to distinguish players who are just having fun in the game with other players who reciprocate that fun and players who are playing the game at the expense of other players. Finding out the motivation behind toxic players can help identify them in game and address the problem of reducing toxicity.

I will do research on the mindset and motivations of toxic players in video games and why they behave in a toxic manner. This will be done by analyzing literature on player behavior

in video games and how toxicity plays a role in it. Specifically, I will be reviewing the work of Lin and expanding on the problem definition of the motivation behind toxic behavior that currently is categorized in a non-competitive context. In order to completely understand why toxicity exists, there is a need to understand it both in a competitive and non-competitive context since different games exist in both settings. This understanding will hopefully facilitate better preventive measures against toxicity and get to the root of the problem so players can enjoy their experience in video games.

Conclusion

Chat filters currently only provide a basic level of toxicity detection because of their inability to account for intent and context of player messages. I will be designing a chat filter that accomplishes this when analyzing and filtering through player messages. This design will be building on the work of Martens et. al. with their letterset recognition analysis and Kwak et. al. with their analysis on the types of words and phrases used by toxic players at certain points of a competitive match. With the success of this design of a chat filter, it will be possible to decrease the number of players who are incorrectly flagged as toxic in video games while still correctly identifying those who are actually toxic.

In terms of understanding why toxic players behave in the way that they do, I will be gaining a better understanding of this in the context of competitive video games. This will be building on the work of Lin et. al., who have defined the problem of the motivation of toxic behavior in the context of only non-competitive video games. Being able to achieve this complete understanding will be able to facilitate better toxicity preventive measures and allow players to enjoy the games they play.

References

Du, Y., Grace, T. D., Jagannath, K., & Salen-Tekinbas, K. (2021). Connected Play in Virtual Worlds: Communication and Control Mechanisms in Virtual Worlds for Children and Adolescents. Multimodal Technologies and Interaction, 5(5), 27. https://doi.org/10.3390/mti5050027

Roy, S. S., Charaborty, S., Sourav, S., & Abraham, A. (2013). Rough Set Theory Approach for Filtering Spams from boundary messages in a Chat System. 2013 13th International Conference on Intelligent Systems Design and Applications (Isda), 28–34. http://isda.softcomputing.net/isdapaper9.pdf

Kuzu, R. S., & Salah, A. A. (2018). Chat biometrics. IET Biometrics, 7(5), 454–466. https://doi.org/10.1049/iet-bmt.2017.0121

A Systematic Review of Literature on User Behavior in Video Game Live Streaming. (n.d.). Retrieved October 5, 2021, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7246545/

Kwak H., Blackburn J. (2015) Linguistic Analysis of Toxic Behavior in an Online Video Game. In: Aiello L., McFarland D. (eds) Social Informatics. SocInfo 2014. Lecture Notes in Computer Science, vol 8852. Springer, Cham. https://doi.org/10.1007/978-3-319-151687_26

Hirata, K., Shimokawara, E., Takatani, T., & Yamaguchi, T. (2017). Filtering method for chat logs toward construction of chat robot. 2017 IEEE/SICE International Symposium on System Integration (SII), 974–979. https://doi.org/10.1109/SII.2017.8279349

Blackburn, J., & Kwak, H. (2014, April 23). STFU noob! predicting crowdsourced decisions on toxic behavior in online games. arXiv.org. Retrieved October 6, 2021, from https://arxiv.org/abs/1404.5905.

Neto, J. A. M., Yokoyama, K. M., & Becker, K. (2017, August 1). Studying toxic behavior influence and player chat in an online video game. Studying toxic behavior influence and player chat in an online video game | Proceedings of the International Conference on Web Intelligence. Retrieved October 6, 2021, from https://dl.acm.org/doi/abs/10.1145/3106426.3106452.

Märtens, M., Shen, S., Iosup, A., & Kuipers, F. (2015). Toxicity detection in multiplayer online games. 2015 International Workshop on Network and Systems Support for Games (NetGames), 1–6. https://doi.org/10.1109/NetGames.2015.7382991

Lin, H., & Sun, C.-T. (2005, January 1). The “White-Eyed” Player Culture: Grief Play and Construction of Deviance in MMORPGs. Proceedings of DiGRA 2005 Conference: Changing Views - Worlds in Play.