Analysis of Social Behavior vs Team Performance in DotA 2

- Garrett St. Amand
- Matthew Korsa
- Michael Ayres

Introduction:

Dota 2 is a free to play Multiplayer Online Battle Arena (MOBA) where two teams of 5 compete against each other to try to destroy each others' Ancient. This task is accomplished through a long series of decisions made by each player, as the teams (and sometimes the players within the teams) fight over resources, mostly in the form of gold and experience points (XP), as well as prerequisite objectives (structures which must be destroyed before an ancient can be attacked). Teams have several forms of communication available within the game client to help with coordinating purpose and relaying information about the other team. It is actually suggested that how team players communicate with one another and the quality of collaboration is more important than individual player skill[1].

Despite the importance of good team communication and collaboration, in-game communications are all too frequently used to communicate in an abusive or otherwise unhelpful manner, especially to one's own team. Such behavior, which often includes homophobic, sexist or racist language, is often classified as "toxic"[2], and such messaging often leads to lessening enjoyment from the game by one's team members. The toxic culture that has become the norm in MOBA games has become so prominent, game publishers have had to introduce multiple approaches to combat it ranging from in-game hindrances to offenders to outright bans from competitive play[3], to bans from the game as a whole. One publisher even went so far as to hire a psychologist to concoct a way to curb toxicity[4]. Regardless of steps taken to reduce toxicity in MOBA games, it still remains a problem and negatively affects the games. Some players have compared the use of toxicity in these games as being a type of addiction with its use giving the offending player the same psychological release as someone who indulges in giving in to an addiction[5]. Some blame the toxicity on the game design itself

[1] "Analyzing Factors Contributing to the Success of a Team in DOTA 2 - DiVA portal." http://www.diva-portal.org/smash/get/diva2:1216849/FULLTEXT01.pdf. Accessed 2 May. 2019.

[2] "MOBA Monday: Toxicity in MOBAs - MMOGames.com." 20 Jul. 2015, https://www.mmogames.com/gamearticles/moba-monday-toxicity-mobas/. Accessed 2 May. 2019.

[3] "Is MOBA culture helplessly toxic? This classy ban acceptance says no - VG247." 3 Jun. 2014, https://www.vg247.com/2014/06/03/is-moba-culture-helplessly-toxic-this-classy-ban-acceptance-says-no/. Accessed 2 May. 2019.

[4] "How Valve Tricked Players into Being Less Toxic - Motherboard - Vice." 6 Nov. 2016, https://motherboard.vice.com/en_us/article/qkjned/how-valve-tricked-players-into-being-less-toxic. Accessed 2 May. 2019.

[5] "A Daily Dosage of Toxicity in Dota2 – HCI Games Group – Medium." 5 Jun. 2018, https://medium.com/@hcigamesgroup/a-daily-dosage-of-toxicity-in-dota2-16488fbf70b6. Accessed 2 May. 2019.

which breeds toxicity as soon as players have to choose their characters[6]. Some studies even suggest toxic player behavior has more than just an emotional impact, finding links between toxic player behavior and reduced average win rates[7] which was our focus.

Our project sought to further test the effects of toxic player behavior on gameplay performance, but with a focus on measurable differences for discrete time intervals. The idea with this decision was to see the more immediate effects of toxic behavior, rather than the already-studied long term impacts. Further, chat messages which might be considered toxic tend to occur after a minor loss within a match (when one's character dies being a common time to start angrily messaging allies), which leaves the question of cause unanswered (did the toxic messaging cause the team to lose, or does the team losing induce toxicity among its members?). Our analysis was built to give insight to this topic, as we measured team toxicity and performance at the minute-by-minute level, rather than comparing completed match results, thus giving an opportunity to measure any short term impacts of toxic player behavior while maintaining the sequence of events.

Problem Characterization:

In order to accurately determine the effects of toxic behavior on performance, there are a few factors we needed to consider to ensure that our results were not impacted by other aspects of the game. Dota 2 is a complex game with a lot of variables that can help and hinder teams throughout a game, not limited to experience, objectives, gold, and just plain luck. One of the challenges we faced was making sure that the results we were recording were influenced by toxic behavior and not the regular happenings of the game. Another factor we had to account for was that the impact of toxic behavior potentially influences team performance for a certain period of time after. For that reason simply looking at the previous minutes' data was not sufficient. Another challenge we faced was the actual analysis of what makes a sentence toxic and how to check for false positives.

Separating our results from the influence of other game factors was crucial for figuring out what the effects of toxicity had on team performance. One aspect of that separation was using data that by itself wouldn't be affected by which team is currently winning. To achieve this, we decided to use delta values rather than cumulative ones, so each team's performance would be evaluated based on how much gold or experience they earned in a certain period rather than just how much they have earned overall.

Another thing we had to account for was how toxic behavior continues to influence the game even after it has occurred. Our initial plan was to record the number of toxic messages for each one minute period and then compare it to team performance for the minute ahead of it. A fundamental problem with that approach is that a one minute window is not a reasonable timeframe for people to react to and subsequently get over the toxicity. If two players are in a heated argument where a lot of regrettable words are tossed around, the chances are that

[6] "How the average Dota 2 experience breeds toxicity - AFK Gaming." 6 May. 2018, https://www.afkgaming.com/articles/how-the-average-dota-2-experience-breeds-toxicity. Accessed 2 May. 2019.

[7] "The Impact of Toxic Behavior on Match Outcomes in DotA - University of Tilburg." http://arno.uvt.nl/show.cgi?fid=145375. Accessed 2 May. 2019.

interaction is going to be fresh in their minds for at least the next 5 to 10 minutes. For that reason, we instead used the last 5 minutes of our toxic behavior for our analysis. Since Dota 2 games regularly last around 40 minutes, this restriction still allows us to get sufficient data from each match.

Determining which sentences were toxic presented a few of its own challenges. First, we had to determine what makes a sentence toxic. One obvious indicator of toxicity is usually swearing. To record such instances, we used a dataset of over 2000 swear words and compared that against each word in each chat message. Another possible indicator of a toxic message was if a message was typed in all caps, as all caps messages in the context of online chat is usually indicative of yelling. For that reason, each message in all caps that was longer than 5 characters long (in order to not record false positives from acronyms, such as “LOL” and “ROFL”) was recorded as a toxic message. One big issue with analyzing speech is that while it is straightforward enough to check if a certain word is bad or if a phrase is in all caps, determining what a sentence is saying and whether or not the intent was toxic is another thing entirely. For that reason, our analysis will likely under report toxicity.

Dominant approaches to the problem:

There have been several different approaches for trying to determine the impact of toxicity on team work, a game's outcome, and other performance metrics. There tends to be a variance in a few key areas, particularly with what approach is used to detect toxicity and what metrics are being recorded to determine which team is outperforming the other.

One such analysis, "The Impact of Toxic Behavior on Match Outcomes in DotA" by Arjen Traas [7], scans chat logs for toxic phrases and attempts to predict the outcome of the game. Traas' approach to the problem was to use a dataset which had the gold, experience, kill/death ratios, and outcome and weighed it against the total amount of toxicity that was present throughout the game for each team. A disadvantage of this approach is that without access to minute-by-minute metrics it would be more difficult to attribute chat behavior with performance. A reason for this is that games can be quite long, and 40 minutes can be plenty of time for people to clash and recover without letting the toxic behaviour multiply. An advantage of their methodology for detecting toxicity was by letting their toxic word flagger have greater functionality for finding toxic words in different permutations, so it would catch "noob", "noooob", or "n000000000b", etc.

Another report, "Analyzing Factors Contributing to Success of a Team in Dota 2" by John Andersson [1], uses metrics other than toxicity to try to predict match outcomes. Specifically Andersson uses factors such as how well an esports team’s country performed at the olympics, individual performance, and age of their team's organization. The linear model Andersson used had a p-value of .1029, which failed to reject the null hypothesis. This possibly suggests that predicting a game's outcome can not accurately be done by simply analyzing the individual players.

Methodology:

For comparing short term toxic chat behaviors and gameplay performances, we used a dataset which included fifty-thousand matches of Dota 2. These matches were ranked, which basically means players were likely taking these matches very seriously. The data included minute-to-minute performance of each player in each match, along with a list of all chat messages (linked to match, time, and player), a record of every objective taken (primarily noting when a structure preventing a team from winning was destroyed), and two sets of records describing teamfights, which were any instances of at least three player characters dying within a fifteen second window.

The first major task for our analysis was to classify toxic player behavior. There are several types of behavior available to players which could be considered toxic, the most directly visible being abusive or otherwise aggressive chat messages. With the goal now to classify chat messages as toxic or non-toxic, we compiled a blacklist of commonly used swears, slurs, and terms specific to Dota 2 which could be seen as toxic. This list was formed using two existing datasets for swearing alongside some additional terms we came up with from personal experience with the game (ie "reported" is normally a perfectly acceptable term, but in the context of an online game, with moderation built largely around community reporting, can be seen as a threat). We then ran Spark code which looked at each of the 1.3 million valid chat messages (we lost a small percentage to formatting and language issues) and checked their contents against our blacklist of about two thousand terms. If a message contained a word on the list, or the message was sent using all capital letters, we flagged the message as toxic. If the message contained no matches, it was assumed non-toxic for our purposes. The list of flags for toxic and non-toxic messages was then saved as a CSV, keeping columns for matchID, the player who sent the message, and time the message was sent.

Next, we analyzed teamfights to determine the likely winner. The data for teamfights was split between two files, one being metadata about the fight (one row per fight), and the other a more detailed description of each player's stats before and after the fight (ten rows per fight, as 10 players are in a match). After grouping these two files together by matchID and fight number (added column to enable proper grouping), we could calculate the changes in gold and experience for both teams, as well as total character deaths and buybacks used (a player spending buyback is equivalent to dying as far as this analysis was concerned). With these statistics available, we then compared the two teams performances in all four categories; if one team measured better performance in a greater number of categories, the winner was declared accordingly. There were rare instances of a tie (about 9k in 278k or 3.25%), which were dropped before saving the teamfight results data as the match the fight took place in, the time the fight started, and the team measured to be the winner.

With all our preliminary analysis completed, the next step was to group all relevant data into a singular file. The issue at hand being the minute-to-minute player data files are not innately comparable to the data listing the event-driven storage of a chat message, teamfight, or objective taken. The solution was to fix each timestamp of the event-driven files by rounding up to the next minute, then grouping everything which matched on both matchID and minute keys. The resulting data was saved to csv, taking special note to the events columns, as more than one can occur within a minute, which was resolved by a space-delimited list of such events within the appropriate column in the csv output. This csv now contains the data for each minute within a match, listing the gold and experience shifts for either team alongside a dynamic list of any significant events which took place during the minute.

The final step involved running the proposed analysis on the now usable data. This step requires maintaining the entirety of any match on the same machine, as well as reading each minute of the match in chronological order. With that constraint handled, each minute of a match was processed for the purpose of determining which team "won" that minute. Winning in any given minute meant comparing the two team's changes in gold and experience points for the past minute, as well as factoring in the winner of any teamfights which occurred and any objectives destroyed in this timeframe. Whichever team scored the highest on more of these metrics than the other was declared the winner for the given time interval. Of note with this comparison, we were comparing changes in gold and experience, rather than actual gold and experience, in order to measure which team gained the most from this minute, rather than make predictions about the outcome of the game. Alongside the assessment of victory, the chat records for each team for the previous five minutes (excluding the current minute), and tallied for toxic and non-toxic messages sent. Each team was then assigned a toxicity score, which was equal to the percent of toxic messages sent from the given team out of the total messages sent by the team during this five minute interval. The two teams toxicity scores were compared, and the team with at least 5% greater toxicity was declared the more toxic team for the minute in question. The two "winner" values for toxicity and actual victory were then stored as a tuple to be counted up after all other minutes were processed. These tuple counts are the final data of the program, which was then run through a one-population z-test for statistical analysis.

All four of these steps were performed using Spark, with the toxic classifier written in Scala, the teamfight assessment and final analysis written in Python, and the data mapping-objective tracker written in Java.

Benchmarks:

The dataset our project used contained match information for fifty-thousand ranked games of Dota 2[8]. After filtering out bad chat message data (some usernames had commas) and foreign character chats (our blacklist was exclusively English terms and a couple of numbers), we classified 1,336,891 chat messages, averaging 26.73 messages per match. The data overall described 2,209,778 minutes of gameplay, which sets the average match length at 44.2 minutes.

After tallying up the total of minutes in the dataset, we checked each tuple for matching pairs and neutral outcomes. Neutral outcomes made a majority of the dataset, mostly due to the chosen time interval of 5 minutes and the frequency at which players use the in-game text chat. This meant we found a very large number of minutes which did not have a more toxic team, approximately 78% of the data. While this percentage is certainly noteworthy, it is within reason for a Dota 2 match, especially given our dataset averaged one chat message every 1.65 minutes of gameplay. After filtering out neutral pairs, we counted the number of times the tuple values matched. If the team measured as more toxic in the previous five minutes was also found to be the winner at the present minute, we noted a match. Of the 395,236 minutes available at this step, we found 186,815 of such matches; meaning that in 47.267% of minutes

[8] Devinanzelmo. (2016, October 24). Dota 2 Matches. Retrieved April 23, 2019, from https://www.kaggle.com/devinanzelmo/dota-2-matches

where one team was measurably more toxic, the more toxic team performed better. In a world where toxicity in the past five minutes has no impact on the performance of the team in the present, this value should make up approximately 50% of results (this is our null hypothesis). Running these numbers through a binomial z-test using the sample size of 395 thousand minutes and 186 thousand matches, we found a z-score of approximately -34, which is incredibly low. This z-score equates to a p-value of < 0.0001, leaving us with 99.99% confidence in rejecting the null hypothesis. These results strongly suggest we reject the null hypothesis, with data making it fairly clear there is a link between team toxicity and future game performance.

Worth noting, while these results strongly suggest a link between toxicity and game performance, the strength of this effect is hard to measure. Our measured 47.267% winrate for toxic teams does not itself directly say the cause, only the fact that the sequence of toxic play was often followed by slightly worse game performance, not necessarily as a result of the toxic play, but that is potentially the case (there may be some other variable which causes both toxic messaging and future performance drops).

Statistics were calculated using a web tool[9], and all data processing steps were run through Spark. Each individual step took under a few minutes to process, so in theory we could include additional data without much issue (granted, the means of getting more data is another programming step in itself).

Insights Gleaned:

This project had a few noteworthy pitfalls that we ended up hitting which required some pivoting and improvising. One such obstacle was foreign languages appearing in the dataset. While our program could identify certain toxic words and mannerisms in English, we were not well equipped to handle other languages. Another issue we encountered was that toxicity was lower than expected. This created issues because when neither team is being toxic at all, we can't do comparisons to see whether toxicity has an impact at all.

Foreign languages present problems to our analysis because we have no way of determining what another culture deems toxic. Even translating our "Bad Words" dataset to any other languages we encountered would be a half baked solution at best, as we don't know if these words are necessarily toxic in other languages. Another issue besides word translation is in mannerisms. When you talk in all caps in English it is generally accepted that you are yelling, which in an online game is usually toxic. Is there an equivalent to caps lock in Chinese? Japanese? Russian? Probably, but ultimately determining what is considered toxic in every language wasn't feasible. Our work around was simply to disregard chat that was in a foreign language by looking at their alphabet.

Toxicity being lower than we expected was probably a symptom of rudimentary toxicity detection. While it is easy to recognize toxicity when one person calls another person a very rude, flagged term, more subtle toxicity will largely go ignored. Simple sentences like "are you going to do anything this game?" or "our AM has an extra chromosome", while both being very

[9] Z-test for One Population Proportion. (2019, February 21). Retrieved May 2, 2019, from https://mathcracker.com/z-test-for-one-proportion.php

toxic, would not be flagged as such. In order for a program to recognize such statements as toxic it would need to have a strong understanding of the English language and perhaps Dota 2 as well. Although our analysis inevitably misses some toxic statements, since we are comparing two teams it can be assumed that over the entire dataset we will miss equally as much from either team. Because of this, our results should still be significant.

Further, while our approach sought to get closer to the cause and effects involved with toxic player behavior, we still could not draw exact conclusions about causality. Further analysis which include more contextual data, such as the history of the team's performance, could lead to a more meaningful conclusion, as there are still several confounding variables which prevent a firm conclusion as to the exact impact of toxic behavior.

Transformation of the problem space in the future:

At its roots this project's purpose was to determine the effects of communication on how well a team performs. Although we chose a specific game to analyze, the results of our project are easily applicable to similar games and at least somewhat applicable to teamwork related activities in other aspects of society. Whether it's people arguing over whose fault it is that their project is not meeting its deadlines or players on a sports team trying to assign blame for a missed goal, toxic communication typically doesn't produce desired results. When it comes down to it we are trying to study human behavior using technology, but in a very controlled environment.

One can only speculate as to the future of gaming, but current trends suggest that people like realism in video games. In less than thirty years the first person shooter genre has evolved from Wolfenstein 3D, considered the first 3D shooter, to games like the modern Call of Duty franchise which are getting closer and closer to photorealism. As the technology to support it came about, gaming as a whole has made gradual, definitive strides towards a more realistic experience. While graphics obviously come to mind when thinking about realism, a more subtle improvement could be in the artificial intelligence of the NPCs (non-player characters) in the game. Though artificial intelligence has certainly come a long way in gaming, right an in game conversation with an NPC is little more than a web of premeditated dialogue you traverse through. Although today there aren't any mainstream instances of a game where a player truly communicates with the NPCs, as machine learning and artificial intelligence truly take off in the next decades I think it will be commonplace within and outside of gaming.

With improvements to technology for understanding human speech, our capability to analyze chat logs for toxicity would expand immensely. Rather than just throwing a sentence into a few functions to see if they raise any appropriate flags we could truly analyze player interaction to find out not only who is being toxic, but towards who and for what reason. We could even determine which players are being targeted with toxic statements to see how it affects them on an individual level. On the flip side we could also analyze chat for other behaviors, such as good sportsmanship or positive teamwork, to see how they benefit a team. In the future humans will use technology to tell them more about their own nature.

Conclusions:

The reduction and elimination of toxicity in MOBA games has long been the goal of game publishers with some obvious and some not so obvious results in its reduction. The first and most obvious result is more enjoyable gameplay. This is mainly due to a lowering of stress and pressure while playing the game. If a player doesn't constantly have to read negative text or listen to negative speech, they can focus on the game enjoyably which often boosts productivity which is reflected as better gameplay.

Another result is a greater sense of psychological safety for players. Psychological safety was identified as the most important factor for team success by the Google's Project Aristotle study[10] and is defined as the freedom to be able to take risks on a team without feeling insecure or embarrassed. If a player doesn't reside within a toxic culture, they feel more secure about their abilities and perform tasks better which advances the goals of a team.

A third result, and one that is attractive to game publishers, is the retention of current players and the expansion of the player base by attracting new players[11]. The bottom line is that games are made to make money. If players are constantly met with toxic attitudes and a toxic culture, they will either leave if they are a veteran player or decide not to play if they are a new player. Lack of toxicity and a sense of acceptance may make veteran players wish to play more and new players wish to continue playing the game. This means more revenue for the game publisher.

Finally, the last result, and one that is less obvious is team and player success. This was the focus of our project and we were tasked with asking ourselves several questions. Does toxic behavior actually affect success? Is there an advantage to being a toxic player? Is it better to focus more on gameplay and less on berating one's fellow players?

As shown from the data analysis performed by our team, it is far better to focus on just playing the game, placing a higher emphasis on collaboration and positive communication, and less on being toxic, be it in the form of aggressive micromanagement or malicious intent. The data showed a correlation between presence of toxicity and success rate in terms of win percentage. This revealed a definite advantage to win percentage for non-toxic players as opposed to toxic ones. This is most likely because time spent by a player berating his or her teammates for decisions that he or she may feel they need to make or decisions teammates are currently making could be better spent on on focusing on gameplay and how well the player works with said teammates.

It should be the goal of all players to reduce toxicity. This means that if a player is toxic they should work to reduce their toxicity, and if a player is encounters toxicity in the game, they should help to combat it. In helping to reduce toxicity, players will experience more enjoyment from the game, better psychological health, a bigger pool of prospective opponents to play against, and as show from our data, success in playing the game.

[10] "Analyzing Factors Contributing to the Success of a Team in DOTA 2 - DiVA portal." http://www.diva-portal.org/smash/get/diva2:1216849/FULLTEXT01.pdf. Accessed 2 May. 2019.

[11] "The Ongoing Issue of Toxicity in DOTA 2 | EsportsTalk.com." 10 Oct. 2018, https://www.esportstalk.com/blog/toxicity-in-dota-2-8480/. Accessed 2 May. 2019.

Bibliography:

"Analyzing Factors Contributing to the Success of a Team in DOTA 2 - DiVA portal."
http://www.diva-portal.org/smash/get/diva2:1216849/FULLTEXT01.pdf. Accessed 2 May. 2019.

"MOBA Monday: Toxicity in MOBAs - MMOGames.com." 20 Jul. 2015,
https://www.mmogames.com/gamearticles/moba-monday-toxicity-mobas/. Accessed 2 May. 2019.

"Is MOBA culture helplessly toxic? This classy ban acceptance says no - VG247." 3 Jun. 2014,
https://www.vg247.com/2014/06/03/is-moba-culture-helplessly-toxic-this-classy-ban-acceptance-says-no/. Accessed 2 May. 2019.

How Valve Tricked Players into Being Less Toxic - Motherboard - Vice." 6 Nov. 2016,
https://motherboard.vice.com/en_us/article/qkjned/how-valve-tricked-players-into-being-less-toxic. Accessed 2 May. 2019.

"A Daily Dosage of Toxicity in Dota2 – HCI Games Group – Medium." 5 Jun. 2018,
https://medium.com/@hcigamesgroup/a-daily-dosage-of-toxicity-in-dota2-16488fbf70b6. Accessed 2 May. 2019.

"How the average Dota 2 experience breeds toxicity - AFK Gaming." 6 May. 2018,
https://www.afkgaming.com/articles/how-the-average-dota-2-experience-breeds-toxicity. Accessed 2 May. 2019.

"The Impact of Toxic Behavior on Match Outcomes in DotA - University of Tilburg."
http://arno.uvt.nl/show.cgi?fid=145375. Accessed 2 May. 2019

"The Ongoing Issue of Toxicity in DOTA 2 | EsportsTalk.com." 10 Oct. 2018,
https://www.esportstalk.com/blog/toxicity-in-dota-2-8480/. Accessed 3 May. 2019.

“Dota 2 Matches” (dataset)
Devinanzelmo. (2016, October 24). Dota 2 Matches. Retrieved April 23, 2019, from
https://www.kaggle.com/devinanzelmo/dota-2-matches

Z-test for One Population Proportion. (2019, February 21). Retrieved May 2, 2019, from
https://mathcracker.com/z-test-for-one-proportion.php