**Detecting Toxic Language in Online Multiplayer Computer Games**Sine Gov

**Introduction**

Toxic language, which is defined as language that is unnecessarily aggressive or inflammatory, is a problem that plagues online communities, especially in online multiplayer computer games. Aside from causing a bad experience while in game, toxicity has a financial impact on companies with products that suffer the problem. Players who experience toxic behavior often report these incidents to customer service representatives, which must then be reviewed. There are methods that are currently employed to combat false reports: thresholds are put into place so customer service representatives are not inundated with reports, however if an online multiplayer game is plagued by toxic behavior, there will still be several reports that must be reviewed. This becomes not only costly from a customer service perspective, but now PR must be involved to help address the perceived toxicity issue in the game. This negative image could also mean loss of current players and prevention of new players from playing the game.

In an ideal situation, we could identify and address the root causes of negativity in these games, however it is a complicated problem with no obvious trend; people will become toxic in game for various reasons and it may well be due to external factors outside of the game itself and the game only serves as a catalyst to trigger the behavior. Until the root cause can be identified, an intermediate solution is necessary to help alleviate the strain that toxicity puts on customer service departments and the online game community.

An automated solution could be designed to identify toxic behavior and action players before a customer service representative needs to address the issue. Using machine learning and data collected from various sources, an algorithm that will identify whether language used could be considered as negative or positive could be created. This algorithm could detect toxicity before a player reports the behavior thus saving the individuals effort for needing to fill out a report. In addition, by not requiring a report to detect toxic language, customer service representatives are free to address other issues. If this algorithm were to have a high accuracy, it could virtually eliminate the need for human intervention for toxic behavior, saving money and time for the game developers and improving the online game experience and possibly the online community.

For this project, the game Overwatch, an competitive online first person shooter, will be examined. For more information about Overwatch, visit the official Overwatch product page: <https://playoverwatch.com/en-us>

**Data Set**

The data was gathered from three online communities relating to Overwatch: Twitter, the official Overwatch forums, and the Overwatch subreddit on Reddit. A script was written for each site that scraped text comments from each site respectively. Additional data such as user name, time of post, link to post was gathered for each text comment, however these were not used in the development of the algorithm. All data was gathered on the same day: December 22, 2017.

**Limitations of Data Set**

This data set does not contain any actual in game chat text, which is where toxicity would be reported. Additionally, this data does not contain human ratings for the sentiment of the text; the bing lexicon in R was used to determine the sentiment scores. This poses a problem since the language of gamers has specific nuances that may not be caught by this library. Additionally, words like ‘kill’ or ‘Mercy’ are commonly used in Overwatch to describe actions in the game, but carry a neutral sentiment. Since the bing lexicon was used, it may be inaccurate in identifying the true sentiment of a statement.

**Data Preparation**

Each text comment was first tokenized and analyzed for sentiment using the bing lexicon from the tidytext library. The score from each of the words was used to calculate the overall sentiment of a given text comment using the formula:

Since each comment may have a varying amount of words, this formula was used to normalize the values between comments so direct comparisons could be done.

**Preliminary Findings in Data**

To determine how toxic a community was in comparison to others, an average of the sentiment scores for each comment was calculated. By doing this, we can also validate the notion that Overwatch communities may suffer from a toxicity problem. Below is a table of the scores:

|  |  |
| --- | --- |
| **Online Community/Website** | **Sentiment Score** |
| Twitter | 0.277143 |
| Reddit | 0.138521 |
| Official Overwatch Forums | 0.115945 |
| Combined (All) | 0.220478 |

These scores were loosely validated by myself through reading through comments for each of these communities. Twitter seemed the most positive and people posted pictures and to alert others when they start streaming. Reddit had several negative comments, but not nearly as much as the official forums for Overwatch.

**Approach**

The approach I will be taking for this project is different from my original proposal. I originally wanted to identify a way to see what items/topics were possibly causing negativity. This information could then be used to inform the designers on how they should proceed in designing their games. I decided against this for a few reasons. First, game design is often full of nuance and intuition and while people may complain on internet forums, the vast majority may actually agree with or like the changes made to a game; the vocal minority does not represent the vast majority. Second, I did not quite have the tools available to perform this kind of analysis. Lastly, this algorithm to detect negative language will provide a a more immediate solution to a complicated problem.