Analyzing Toxicity in Online Gaming Communities

Ayushi Ghosh

Maulana Abul Kalam Azad University of Technology, Kolkata, West Bengal 700064, India

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract:.

The video game culture resulting from the massive consumption of games has built various gaming communities online. The controversies in these groups, especially, the infamous Gamergate controversy indicates that sexist behaviour is prevalent in this circle. The social dynamics of these communities are closely examined in this research. In particular, posts from Twitter and Reddit are analysed to determine the racism, sexism, and political affiliation persisting in these groups. Twitter and Reddit posts related to 13 popular games which have been discussed below are analysed to find answers to the research question. NLP tools such as Bag of Words, Sentiment Analysis, and Word Embeddings are applied to analyse the posts. A formula is developed based on these tools to determine the extent of racism, sexism, and Trump-hate associated with each of the gaming communities. The results help in capturing and evaluating the emotional and linguistic properties of the conversational language that these communities engage in.

Keywords: Online Games, Verbal Violence, Natural Language Processing.

1. Introduction

Online multiplayer games have been the cornerstone for bringing together thousands of gamers together from all over the world. It provides an excellent networking opportunity for people to share cheats, walkthroughs, guides, reviews, and much more to connect with other players. On a different note, the social dynamics of these communicates also indicate racism, sexism, and political affiliation. An exploratory study of sexism in online gaming revealed ways gender-based harassment has been framed and debated within gaming community contexts and the barriers for the recognition of women as 'legitimate' gamers that delimit full gaming citizenship[1]. The infamous Gamergate controversy also highlights the intensity of sexism and misogyny within the video game world. Anita Sarkeesian, a Canadian-American feminist, who started a nonprofit to make movies and video games better for women and girls, has been a prime victim of sexist behavior in online gaming communities. She receives a volume of rape and death threats inconceivable to most of the population, including tweets fantasizing about raping her, telling her to kill herself, or calling her a litany of sexist and racist slurs[2].

42 percent of players are female but still a sexist culture thrives online and offline the video game world[3]. Apart from gender being a salient aspect within harassing interactions,[2] racism and political affiliation are two more stereotypical dimensions in these communities. Now, with technological advances in online multiplayer games and video gaming's increased prevalence worldwide, a growing percentage of the population is becoming unwittingly exposed to a slew of abusive acts that are only becoming more visible[4].

This research paper takes an in-depth look into these aspects to analyze the toxicity in online gaming communities. Data is collected from Twitter and Reddit with a special focus on games covering a broad demographic spectrum of participants. NLP based technologies such a s Bag-of-Words, Sentiment Analysis, and Word Embedding are applied to analyze the data collected. The linguistic properties of the gaming communities are further evaluated to address the research question.

1.1Literature Study

Notable work on this has been previously done by Dr Shane Murnion, Prof William J Buchanan, Adrian Smales, and Dr Gordon Russell where they did a paper "Machine Learning and Semantic Analysis of IngameChat for Cyber Bullying".an automatic data collection system is presented that continuously collects in-game chat data from one of the most popular online multi-player games: World of Tanks. The data was collected and combined with other information about the players from available online data services. Classification of the collected data was carried out using simple feature detection with SQL database queries and compared to classification from AI-based sentiment text analysis services that have recently become available and further against manually classified data using a custom-built classification client built for this paper.The simple SQL classification proved to be quite useful at identifying some features of toxic chat such as the use of bad language or racist sentiments, however the classification by the more sophisticated online sentiment analysis services proved to be disappointing. This paper have used Text Blob Library instead of SQL Classification to give more accurate result.

2. Methodology

2.1 Data Collection

To understand the social dynamics of gaming communities, those games were examined that are accessible to a wide audience. The table below represents the games that were chosen for this research.

Table 1. Games chosen for this research

Game	Description
PlayerUnknown's Battlegrounds (PUBG)	An immensely popular online multiplayer battle royale game that features mass-scale combat
Fortnite	Another massively online multiplayer battle royale game
FIFA	FIFA is among the most celebrated soccer games
World of Warcraft	A leading multiplayer online role-playing game
Dota 2	A game renowned for multiplayer online role-playing. It is the most-played game on the popular gaming platform, Steam
League of Legends	A popular multiplayer online brawl arena game that follows a freemium model
Guild Wars 2	Another popular multiplayer online brawl arena game featuring a storyline that is responsive to player actions
Heroes of the Storm	A crossover multiplayer online battle arena video game
Hearthstone	Digital collectible card game
Overwatch	A team-oriented shooter
Bloodstained	A hack and slash game
The Sims	Life simulation game which is considered to have a female majority of players
Minecraft	Open-world survival game with an emphasis on crafting

Twitter and Reddit were targeted for gathering information shared in relation to the above games.

The tweet-preprocessor library was used to retrieve tweets relevant to the games present in Table 1 and the ones tweeted by gamer profiles.

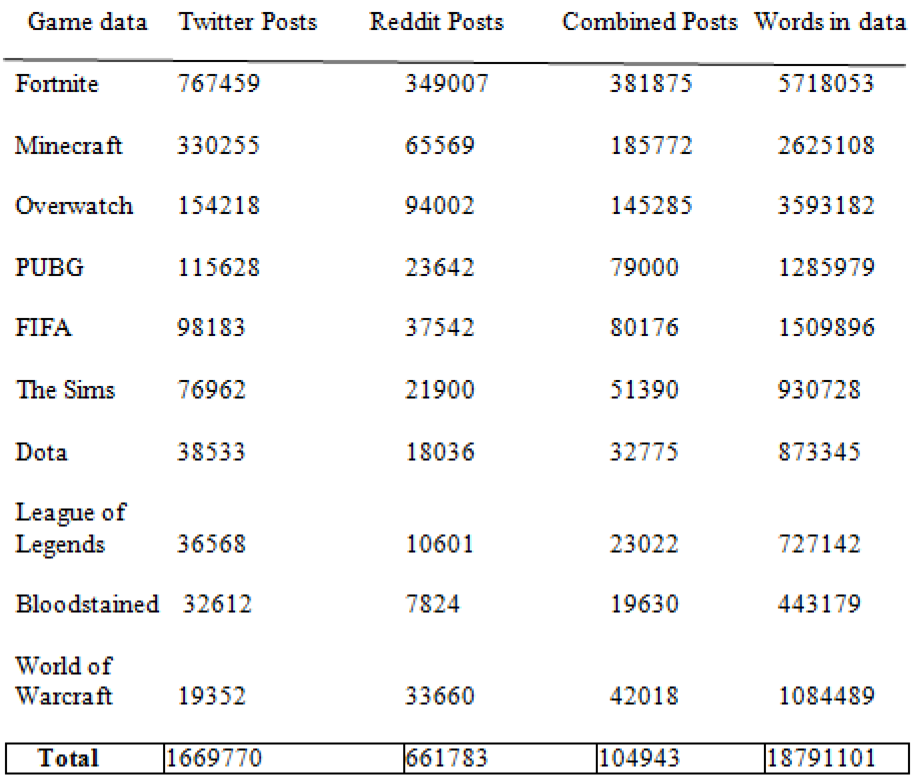
On the other hand, Python Reddit API Wrapper (PRAW) was used to scrape data from Reddit. The topics, subreddits, comments, and replies were carefully examined.

2.2 Data Cleaning

Textual data coming from online environments tend to have misspelled words, special characters, URLs, and emojis. Therefore, this data is restructured using NLP based technologies before converting it into a machine-readable format.

To move forward with this task, data in English language was chosen first. The emojis, mentions, special characters, URLs, and emoticons were removed after that including duplicate posts. The words that convey little meaning on their own, commonly referred to as "stopwords", were also removed. The following table represents the statistics of the data collected before and after cleaning.

Table 2. Statistics of the data collected before and after data cleaning



Game data	Twitter Posts	Reddit Posts	Combined Posts	Words in data
Fortnite	767459	349007	381875	5718053
Minecraft	330255	65569	185772	2625108
Overwatch	154218	94002	145285	3593182
PUBG	115628	23642	79000	1285979
FIFA	98183	37542	80176	1509896
The Sims	76962	21900	51390	930728
Dota	38533	18036	32775	873345
League of Legends	36568	10601	23022	727142
Bloodstained	32612	7824	19630	443179
World of Warcraft	19352	33660	42018	1084489
Total	1669770	661783	104943	18791101

2.3 Data Manipulation

The preprocessed data is further manipulated using the following NLP techniques to derive deeper insights from the information collected.

Bag of Words The bag-of-words model is commonly used in methods of document classification where the (frequency of) occurrence of each word is used as a feature for training a classifier [5].

The bag-of-words model is applied in this research to ascertain the term frequency in order to measure the importance of a particular term in a document. A document in this case, implies a game's entire data set comprising Reddit threads and tweets. The words are further grouped into neutral sentiment words and negative sentiment words.

Sentiment Analysis Sentiment analysis, also called opinion mining, is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes [6].

In this research, the Text Blob library is used for the purpose of sentiment analysis. This resulted in scores that can be represented as – {polarity, subjectivity}. The polarity score can be described as a float value [-1.0, 1.0]. A negative emotion is indicated with a negative score and it is the otherwise for a positive emotion. However, the subjectivity score is dismissed considering that all the posts are subjective in nature.

Word Embedding The vector representation of words was utilized to acquire quantitative attributes in relation to the content in the tweets and Reddit threads. The famous deep learning model, Google's Word2Vec, was used to compute the same in order to capture the semantic meaning of the words in context of a particular game's community.

The Gensim library Word2Vec implementation was used to train the Word2Vec model for this research. Each unique word in the document is converted to a high dimensional vector with this model which moves forward to become a part of the training corpus.

This technology is employed to model the relation of the gaming communities to the aspects of racism, sexism, and political affiliation. A set of pre-defined vector word arithmetics was used to calculate this relation. One such example can be described below for the game, PUBG.

man is to woman like gamer is to: [’origin’, ’gamers’, ’siblings’, ’adult’, ’battleground’, ’print’, ’str’, ’fortnitememes’, ’playerunknownsbat’, ’freefire’] man is to woman like pro is to: [’gamestop’, ’claw’, ’handcam’, ’gay’, ’mdiscrazy’, ’realme’, ’aka’, ’finger’, ’hein’, ’roz’] man is to streamer like girl is to: [’viewers’, ’gamer’, ’youtuber’, ’nimo’, ’youtube’, ’doc’, ’channels’, ’savage’, ’twitchsquads’, ’adult’]

trump is to muslim like man is to: [’lego’, ’rising’, ’enemy’, ’belt’, ’issued’, ’holes’, ’racing’, ’bat’, ’star’, ’epsoide’] most similar words to ’mother’ are [’fuckers’, ’ass’, ’sister’, ’committed’, ’fucker’, ’edgy’, ’ryan’, ’giant’, ’stupid’, ’eye’] most similar words to ’girlfriend’ are [’ajak’, ’shes’, ’whore’, ’horses’, ’cute’, ’chor’, ’istg’, ’bye’, ’college’, ’sis’] most similar words to ’trump’ are [’donald’, ’shithole’, ’wallah’, ’mr’, ’wat’, ’cringe’, ’hello’, ’iran’, ’nonsense’, ’act’]

Empirical Evaluation A formula is developed to extract factual information in relation to the aspects of sexism, racism, and political affiliation which can be described below:

$$\alpha * X_{term-freq} + \beta * X_{sentiment} + \gamma * X_{embedding}$$

This formula can be applied to calculate the aspects of our research interest but before that, the inner terms also need to be determined. Therefore, two sets of words are defined for each of these aspects – Neutral Words and Negative Words.

The calculation of the inner terms can be described as follows

$$X_{term-freq} = \frac{Negative\ Posts}{Total\ Number\ of\ Posts} \qquad \ldots (1)$$

Where Negative Posts are the posts containing one or more words from the Negative Words set for a given property (Racism, Sexism, Trump-hate).

$$X_{sentiment} = Avg_{Polarity}[Neutral\ Posts] \qquad \ldots (2)$$

Where Neutral Posts are the posts containing one or more words from the Neutral Words set for a given property. As we wish to evaluate how negative is the sentiment in posts concerning our topics.

$$X_{embedding} = \frac{1}{N}\sum_{n=1}^{N}\frac{Most\ similar\ negative\ words}{Most\ similar\ words} \qquad \ldots (3)$$

Where Most Similar Negative Words are the total number of Negative Words that appeared in the Most-similar results for a given neutral word in the trained embedding model per game. And N represents the number of Neutral Words.

3. Results

3.1 Bag of Words

The bag of words model helped in determining the term frequency in the content that was gathered. The highest term frequency was achieved by the game’s name for every game which is obvious.

Figure 1 represents a comparison of sexist, racist, and Trump related words. Racist words are found to be the highest in FIFA’s community apart from significant usage of Trump related words. On a different note, Dota 2 and League of Legends have a low score in terms of Trump hate.

The Sims community dominated by women, used the most sexist words, which is an interesting finding.

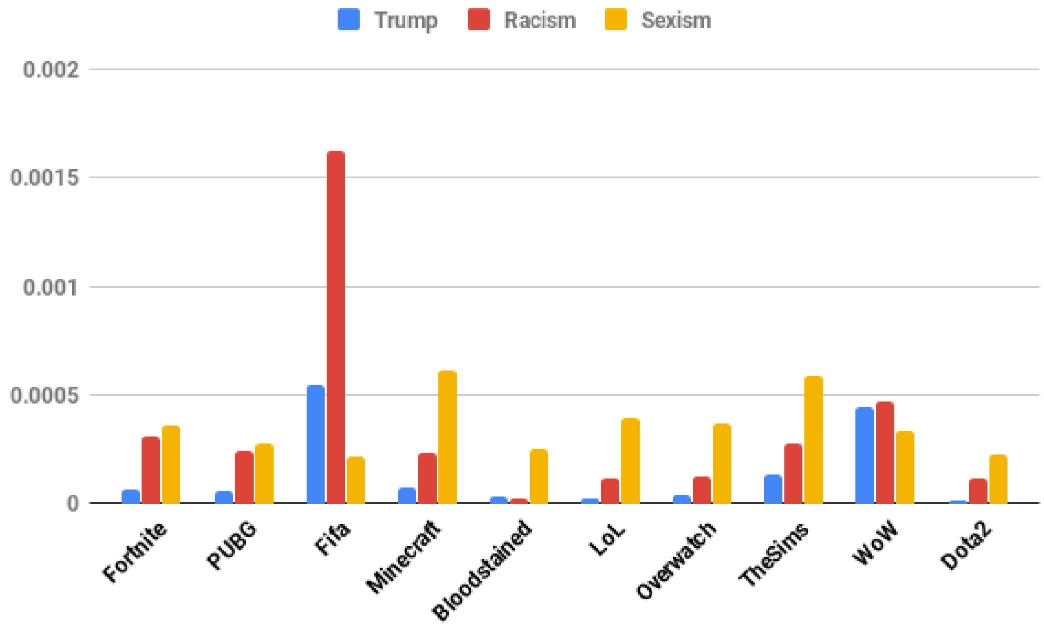


Figure1. Percentage of sexist, racist and Trump related words in each game's community

3.2 Sentiment Analysis

The results of Sentiment Analysis can be demonstrated in Figure 2 and Figure 3.

Figure 2 shows the percentage of positive, neutral, and negative posts per game community. It can be observed that most of the posts are emotional which can be deduced from the fact that the posts are either positive or negative in nature.

Bloodstained being a violent game, is fairly negative whereas, World of Warcraft and League of Legends are positive. Also, Minecraft is found to be neutral.

To determine the strength of polarity of each of the posts, average positive and negative polarity is calculated for every community. Figure 3 is a representation of the same. It is observed that Bloodstained is strongly negative and Minecraft is found to be positive to a great extent.

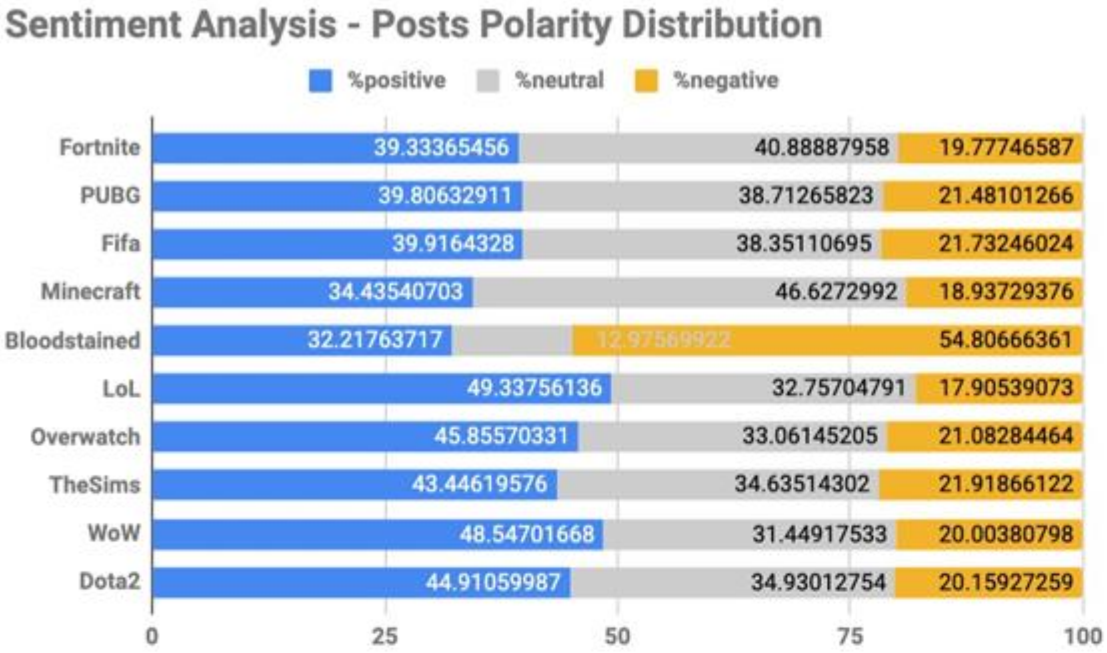


Figure 2. Blue bars represent percentage of Positive posts per game community, grey for Neutral and yellow for Negative sentiment.

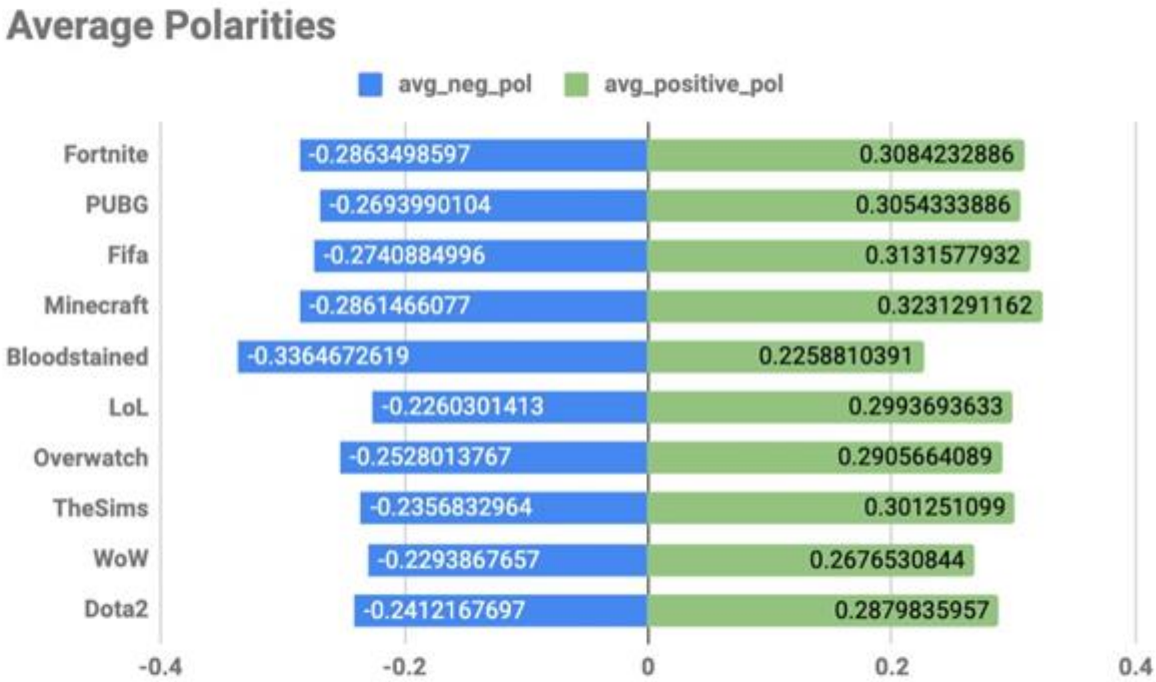


Figure 3. Blue bars represent the avg. value of all the negative polarity posts for that community while the green is avg of all positive polarity posts.

3.3 Word Embeddings

Upon achieving the embedding models that resemble the gaming language, word triplets calculation was performed. Some interesting results are presented below.

When referring to women gamers in the game Fortnite these triplets were found

Man → Woman like Gamer → [gamergirls, egirl, greatgame] while when performing this on FIFA we got Man → Woman like Gamer → [meaningless, exclude, prize] that show a much more discriminating attitude towards female players.

When reviewing the different communities' politic affiliation, for the game Fortnite it was found Trump → Muslim like Man → [islamophobic, huzzah, recruits] and that the most similar words to Trump are [toddler, donald, rat, president, tik, harass] while in the game PUBG the triplet Trump → Clinton like Man → [dumbest, mess, alves] and the most similar words to Trump are [donald, cringe, iran, nonsense, act] these can indicate that PUBG players are more right-winged than Fortnite players that present a more liberal line.

When reviewing similar words in FIFA the most similar words to Arab are [slave, gulf, exploitation, immigrants, qatar] and most similar words to Africans are [slaves, insult, slave, labor, nigga] which can indicate a culture of racism.

3.4 Empirical Evaluation

Different coefficients were considering while performing the empirical evaluation for racism, sexism, and Trump hate in each game. This can be described below

$\gamma = 0.6$ - the coefficient for the word embedding term, which holds the most valuable contribution to this metric, hence the highest portion is granted to it.

$\alpha = 0.3$ - the coefficient for the term frequency, which holds insight to community's linguistic choices, and therefore, is the second most important.

$\beta = -0.1$ - the coefficient for the sentiment, when experimenting with the results from the sentiment analysis model was found to be inaccurate in some cases which puts its reliability in question so it was set as the lowest contributing part. It has a negative value as the polarity score is negative.

Figure 4 represents results similar to that of the bag of words results. It marks FIFA as the most racist gaming community with a score of 0.0472. However, this is not a very high score for our evaluation bracket [0,1].

Dota 2 achieved the highest score, 0.0577, in terms of sexism. The least sexist community is The Sims as opposed to the bag of words results.

Trump-hate scores were observed to be fairly high in Bloodstained. However, the reason behind this was that the word "president" is related to the game but it was not used in context of Trump here. Interesting, removing this resulted in Trump-hate scores to fall to 0 for Bloodstained.

Fortnite achieved the highest score in terms of Trump-hate, 0.007. It is still significantly low indicating that the gaming communities are least interested in Trump related aspects.

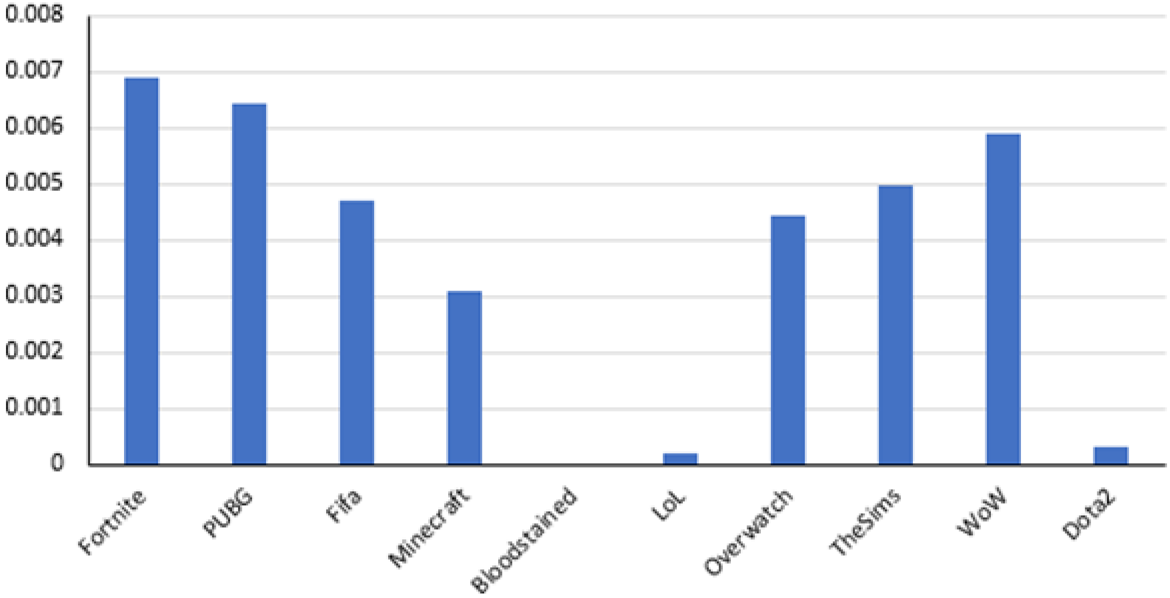


Figure 4. Values for Racism property per game community

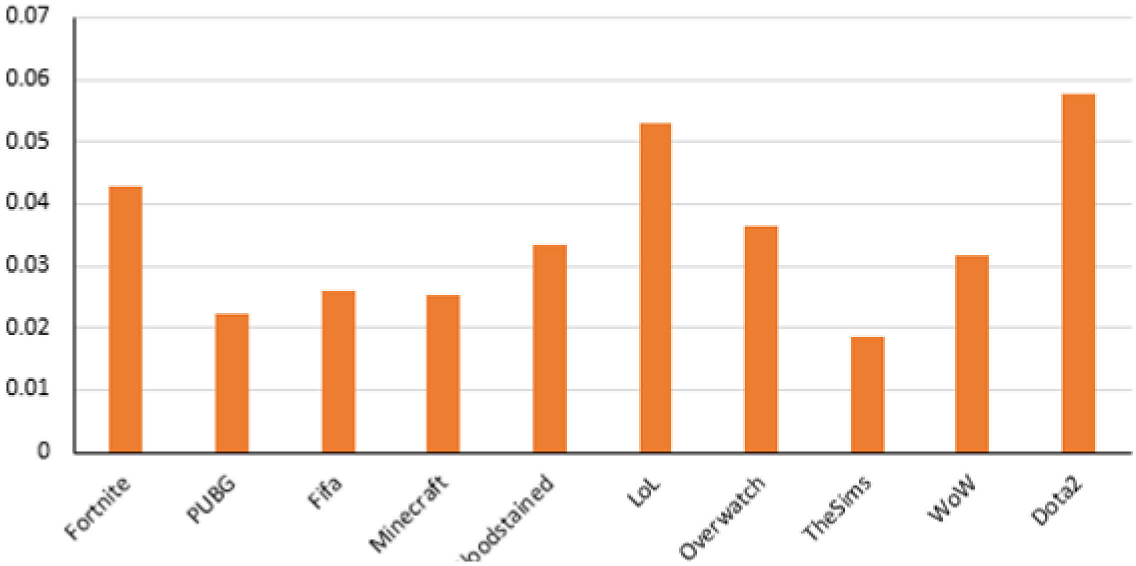


Figure 5. Values for Sexism property per game community

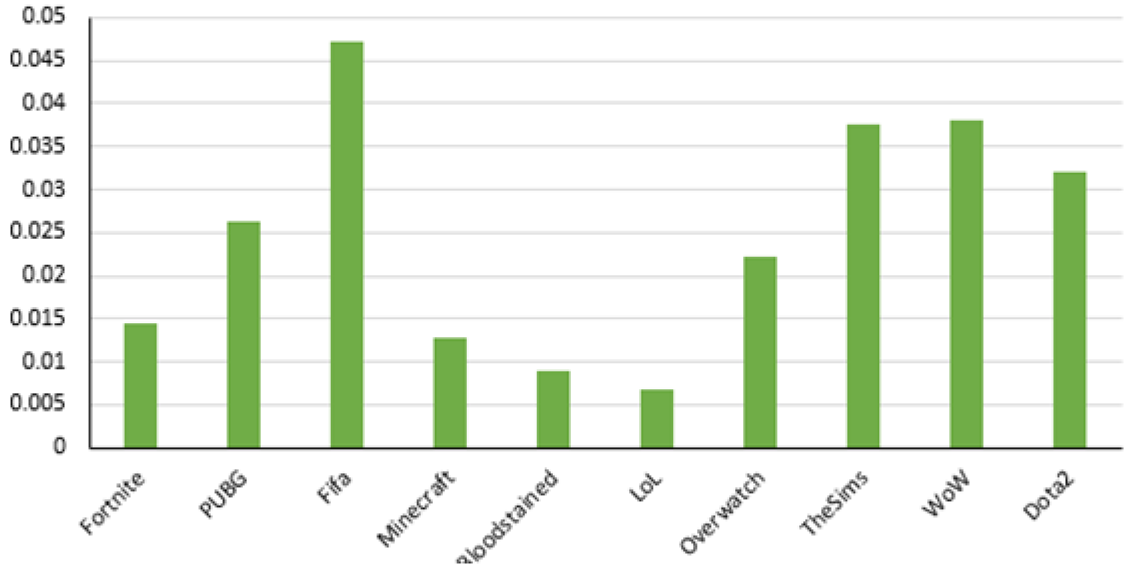


Figure 6. Values for Trump-hate property per game community after re-running empirical evaluation on Bloodstained without the word president

4. Conclusion and Future Work

The results of this research indicate racism and sexism truly intoxicate the gaming communities to a great extent. Racism is observed to be the highest in the community of FIFA and the Bloodstained community expresses strong negative emotions. The scores for sexist behavior are interestingly very high for the Sims community for the term frequency results as opposed to the empirical evaluation results. Apart from that, the findings reflect that the gaming communities are least interested in politics. In conclusion, this novel method helps to measure the linguistic and emotional properties deciphered from the language that a community uses.

However, our work is limited to gaming communities and analysis of the toxic behavior pertinent to the language that these communities use. This work can also be extended to perform content moderation. Using GlobeVe instead of Word2Vec, even though Word2Vec works on the pure co-occurrence probabilities so that the probability of the words surrounding the target word is maximized. Word2Vec is a predictive model while Glove is a count-based model. However, our work is limited to gaming communities and analysis of the toxic behaviour pertinent to the language that these communities use. This work can also be extended to perform content moderation. Our work is only limited to identifying the negativity metric in terms of sexism, racism, and political affiliation. Other factors that lead to marginalization online and affect the sentiments of minority groups were not taken into consideration. Thus, our future work also involves working on this area..

References

1. Nic Giolla Easpaig, B. (2018). An exploratory study of sexism in online gaming communities: Mapping contested digital terrain. Community Psychology in Global Perspective, 4(2), 119-135.
2. Fighting, F. J. A. S. I. to Make the Web Less Awful for Women–And Getting Death Threats in the Process. URL: https://www. cosmopolitan. com/career/a39908/anita-sarkeesian-internetsmost-fascinating/(дата звернення: 31.10. 2019).
3. Rivas, J. Watch Anita Sarkeesian Deconstruct Sexism in Gaming. URL: https://www. colorlines. com/articles/watch-anita-sarkeesian-deconstruct-sexism-gaming (дата звернення: 31.10. 2019).
4. Smith, N. (2019, February 26). Racism, misogyny, death threats: Why can't the booming video-game industry curb toxicity? Retrieved August 26, 2020, from https://www.washingtonpost.com/technology/2019/02/26/racism-misogyny-death-threats-why-cant-booming-video-game-industry-curb-toxicity/
5. McTear, M. F., Callejas, Z., & Griol, D. (2016). The conversational interface (Vol. 6, No. 94, p. 102). Cham: Springer.
6. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167