Journal of International Women's Studies

Volume 24
Issue 3 *Revisiting Women's and Gender Studies across Interdisciplinary Research*

Article 7

July 2022

The Bullying Game: Sexism Based Toxic Language Analysis on Online Games Chat Logs by Text Mining

Nihan Yıldırım

et al

The Bullying Game: Sexism Based Toxic Language Analysis on Online Games Chat Logs by Text Mining

By Aslı Ekiciler[1], İmran Ahioğlu[1], Nihan Yıldırım[1*] and İpek İlkkaracan Ajas[1], Tolga Kaya[1]

Abstract

As a unique type of social network, the online gaming industry is a fast-growing, changing, and men-dominated field which attracts diverse backgrounds. Being dominated by male users, game developers, game players, game investors, the non-inclusiveness and gender inequality reside as salient problems in the community. In the online gaming communities, most women players report toxic and offensive language or experiences of verbal abuse. Symbolic interactionists and feminists assume that words matter since the use of particular language and terms can dehumanize and harm particular groups such as women. Identifying and reporting the toxic behavior, sexism, and harassment that occur in online games is a critical need in preventing cyberbullying, and it will help gender diversity and equality grow in the online gaming industry. However, the research on this topic is still rare, except for some milestone studies. This paper aims to contribute to the theory and practice of sexist toxic language detection in the online gaming community, through the automatic detection and analysis of toxic comments in online games chat logs. We adopted the MaXQDA tool as a data visualization technique to reveal the most frequently used toxic words used against women in online gaming communities. We also applied the Naïve Bayes Classifier for text mining to classify if a chat log content is sexist and toxic. We also refined the text mining model Laplace estimator and re-tested the model's accuracy. The study also revealed that the accuracy of the Naïve Bayes Classifier did not change by the Laplace estimator. The findings of the study are expected to raise awareness about the use of gender-based toxic language in the online gaming community. Moreover, the proposed mining model can inspire similar research on practical tools to help moderate the use of sexist toxic language and disinfect these communities from gender-based discrimination and sexist bullying.

Keywords: Verbal Harassment, Multiplayer Online Games, Chat Logs, Toxic language, Sexism, gender bias, Text Mining, Naïve Bayes Classifier, Women Players

Introduction

While gender equality is one of the United Nation's 17 Sustainable Development Goals (UN 2020), verbal harassment and language toxicity remains a dark reality in physical and digital communities, directed to women and non-binary individuals and all minorities. A Sexual harassment is characterized as unwanted sexual progress or any other activity that targets someone based on their gender, including sexual harassment, lewd or discriminatory remarks, or sexual coercion, pressuring someone to perform sexual actions (Pina et al., 2009; Tang, Reer and Quandt, 2020. Symbolic interactionists (Blumer 1969; Mead 1934) and feminists assume that words matter, since any terms that dehumanize and discriminate against others can make it easier to harm them (Schwalbe 2008; Kleinman et al., 2009). Whittaker and Kowalski (2015) revealed that texting and social media are the most commonly used venues for cyberbullying. At the same time there are numerous studies that

[1] Department of Management Engineering, Istanbul Technical University, 34367 Istanbul, Turkey
{ekiciler15, ahioglu15, yildirimni, ilkkaracan, kayatolga}@itu.edu.tr
* Corresponding author

explore the topic of hate speech (Leite et al., 2020; Waseem and Hovy, 2016; Chung et al., 2019; Basile et al., 2019).

Social network platform moderators and specific content selection systems require analytical tools and embedded models which automatically identify toxic comments. However, developers of such intelligent models should consider some significant challenges such as the lack of explicity of toxic language and the large spectrum of types of toxicity (e.g. sexism, racism, insults). They also have to be aware of the fact that toxic comments correspond to a minority of comments which threatens the availability of highly unbalanced data needed for automatic data-driven approaches (Leite et al., 2020). However, identifying cases in social media and social network platforms such as online gamer communities is a challenging task that requires dealing with massive amounts of data. Therefore, intelligent systems with automatic approaches for detecting online hate speech have received significant attention in recent years (Leite et al., 2020). Many studies also focused on social media for exploring toxicity through applying learning algorithms to user-generated content (Risch et al., 2018; Subramani and et al., 2018, 2019, Mai et al., 2018). Recently, Fan et al. (2021) suggest a model to detect and classify toxicity in social media from user-generated content using the Bidirectional Encoder Representations from Transformers (BERT).

As a unique type of social network, the online gaming industry is a fast-growing, changing and male-dominated field which attracts players from diverse backgrounds. The online gaming industry is generally dominated by male players, game developers, game investors, creating a non-inclusive community prone to gender-based hostility, which resides as a salient problem in the community. According to the Female Gamer Survey by Bryter (2019), one in three women gamers report being exposed to harassment or discrimination by men gamers. In the online gaming communities, most women players report encountering offensive language or verbal abuse against them. Abusive language is the most reported type of toxic behaviour, with a rate of almost 85 percent in one million reports (Kwak et al., 2015). Identifying and reporting toxic behaviour, sexism, harassment in on-line gaming is a critical need in preventing cyberbullying and helping gender diversity and equality grow in the online gaming industry. However, the research on this topic is still rare, except for some milestone works like Blackburn and Kwak (2014), Kwak et al., (2015) and Martens et al. (2015).

This paper aims to contribute to the theory and practice of sexist toxic language detection in social networks, with a focus on analysing and automatically detecting toxic comments in the context of online gaming communities. We propose an analytical system that reveals verbal abuse in online gaming platforms, raising awareness about use of gender-based toxic language and contrubuting to efforts to prevent discrimination in digital communities. The proposed system analyzes the chat messages sent during online games to highlight the toxic and sexist language by utilizing text mining methods and publicly reports the results of this analysis to all shareholders.

The intention here is to create a social consciousness, which we hope, will create its sanction power to silence the harassers. To this end, we first retrieved an online game chat log data on which to conduct our analysis. We labelled the content in a randomly generated sample from this database as sexist versus non-sexist, using the keywords we acquired from our literature review. Based on this classification method, the machine-learning model examines the data and learns from it. Consequently, the system becomes able to report on new data (chats) concerning the sexist discourse it contains according to the model it learned. We ran the Naïve Bayes Classification method with an R coding language, classifying new data according to the labelled training data. While online game developers and publishers naively try to address this problém, there are still many in the community who advocate sexist and obscene language as part of the online gaming culture and say, "*women cannot*

complain about it" (Fletcher, 2012). With the help of text mining tools, we want to unveil how inappropriate and degrading the online chat messages can be and raise awareness in the on-line gaming community by developing a generic detection tool to identify toxic and discriminative language.

Research on Toxicity and Sexist Toxic Language in Social Networks

Language Toxicity in Online Games:

With the participation of billions of players, the multiplayer online gaming community is a social networking platform that provides entertainment, enjoyment, and engagement for gamers globally (Tyack et al., 2016). Due to its high interactivity, competitiveness and stressing nature, online gaming communities are classified as highly risky in terms of destructive interactions among gamers (Griffiths, 2010; Yousefi and Emmanouilidou, 2021). Yousefi and Emmanouilidou (2021) defined the negative online behavior by referring to cyberbullying (Salawu et al., 2017), cyber-harassment (Marwa et al., 2018), abuse (Founta et al., 2019), hate speech, and toxic language on different social networking platforms. Jigsaw, a unit within Google, that explores threats to open societies, defines toxicity as "rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion". Jigsaw (2021) recommends using this definition of 'toxicity' as opposed to concepts like 'abuse' or 'hate speech' for both practical and scientific reasons since toxicity refers to a broader category of language, subject to individual interpretation.
Toxic behaviour is a common problem, and multiplayer online video games are based on teamwork. Typically lousy behaviour is considered toxic in multiplayer games because multiple players may be exposed to such behaviour through player interactions, harming the entire group (Blackburn and Kwak, 2014). Toxic language is widespread in online games as a means of releasing anger and disappointment in a destructive manner (Kordyaka et al., 2019). Previous studies also focused on the controversial issue of whether violent video games lead to aggressive behavior in real life (Lemercier-Dugarin et al., 2021)

Several indicators can be defined as toxic behaviour in the online video gaming community. Harassment is also referred to as toxic behaviour in the context of an online game, which involves offensive words, verbal harassment, derogatory behaviours, unacceptable titles, spamming, and failure to communicate (Kwak, Blackburn and Han cited in Lapolla, 2020). Toxic behaviours negatively affect game players and online video game platforms, video game developers, etc. Detecting toxic behaviour is valuable, yet forestalling it or decreasing its effect is even more crucial (Blackburn & Kwak, 2014). To understand and detect the toxic language, all stages of online video games and behavioural patterns of online game players should be monitored and analyzed. The video game The League of Legends seems to be a in a transformation phase (Blackburn and Kwak, 2014). During the early stage of the match, toxic players act the same as regular players. However, they change their actions at some point. They do not apologize or compliment for sharing their feelings and avoid strategic contact with team members. Online video game players' toxic behaviours occur relative to game performances, desire to win the match, and communication patterns.

In their recent research, Lemercier-Dugarin et al. (2021) examined the relationship between toxicity (a form of verbally aggressive behaviour directed against other players) in multiplayer online video games and several potential predictors such as personality traits and emotional reactivity. They concluded that younger age, being male, spending a lot of time playing per week, and being highly achieving increased the likelihood of reporting toxicity and behavioural change in the game.

Sexism Based Toxicity and Female Gamer Experiences.

On-line gaming industry is a growing industry, streotyped as a men-dominated community, yet this identification changes with increasing numbers of women game players. In fact, it is estimated that 41 percent of computer gamers and video gamers in the USA are women (Statista.com 2019). While 79.4 percent of women gamers agree that video games can be inspiring, 75.9 percent report that they are abused or discriminated against during online games (McDaniel, 2016). Girl gamers experience additional toxicity aimed at them because of gender in addition to the general toxicity that video game players experience (Khandaker, 2019). Sexual harassment, discrimination, and verbal abuse are the main salient negative experiences of toxicity that women game players face. Ballard and Welch (2017) found that the groups most likely to receive adverse reactions from male players were women and lower-performing online game players, and male gamers are more likely than females to commit cyberbullying.

We used our literature review to identify some keywords in toxic sexist (gender discriminative) language used in on-line games. Table 1 shows the most common words used in online game platforms and other online platforms to degrade females.

Table 1: Sexism Based Toxic Keywords

Keyword	**Way of use**	**Reference**
bitch	By seeking to shame women with labels that counter these normative standards, harassers perpetuate traditional assumptions, intentionally or not.	Felmlee, Rodis, & Zhang (2019).
whore		
cunt		
plague	Hateful molestation	Johnson (2014).
cancer		
fat bitch		
obese cunt		
Get the fuck out	Attacks by misogynist	Johnson (2014).
boobs	Objectification of women and harassment	Ciampaglia 2019
pussy		
fat		
Slut.	Insulting women based on their sexuality	Rose 2018
Like a girl	Minimizing women's achievement	Rose 2018
Tits/Titty	Asserting a female's body part	Pkwzimm (2018)

The word “fuck” is widely discussed for being a sexist word in English language. Hobbs (2013) argued that the word fuck is a sexist word since it functions as a metaphor for male sexual aggression. Notwithstanding its increasing public use, or enduring cultural models that inform our beliefs about the nature of sexuality and sexual acts preserve its status as a vile utterance that continues to inspire moral outrage” (Hobbs, 2013). Based on Hobbs‘ (2013) discussion on “the directness and specificity of its reference to bodily penetration“, we classified this word as “toxic” on the ground that it reflects patterns of masculinity as an expression of gender-based discrimination.

Kleinman et al. (2009) defined the term „bitch“ as one that reproduces sexism. Despite sometimes also being used in daily language in a way to express “strong” or “fancy”

females, and despite the fact that i tis also used by women,“bitch“ i salso term that reinforces the idea that women are essentially different from men and in a negative way. Hence it can be classified as a sexist hence a toxic word.

Past and Current Solutions

As online gaming becomes increasingly social and interaction between the players' increases, it becomes inevitable for gaming companies to recognise toxicity as a concrete problem. Hence the need arises to detect toxicity in online games, and impose sanctions. Riot Games, owner of one of the most popular video games, League of Legends, launched The Tribunal in 2011 as a crowdsourcing platform to let “expert” players decide whether a reported player should be punished or not (Kwak et al., 2015). They cancelled it circa 2018 due to being inefficient and slow (Nexus, 2018). Kou and Nardi (2014) reported that when the Tribunal was formed, over 47 million votes were cast in its first year evaluating toxic behavior (as cited in Lapolla, 2020). 74 percent of players who faced punitive toxicity interventions subsequently changed their in-game behavior. Blizzard, the owner of Overwatch, tried a different approach to stop toxic behaviour: It launched an endorsement system to promote positive behavior via getting endorsed for such behaviors by other players, shown on a badge next to their name. Blizzard announced that this system had decreased the overall toxicity by 40 percent in 2019 (Ziegelman, 2020). CS:GO’s producer Valve has also taken some steps to fight toxicity during online games. One of them is to stop verbal abusers by auto-muting them if they get too many behavioural in-game reports (Ziegelman, 2020).

Another example is an AI-based system called “Minerva” that allows in-game reporting of a toxic player at the exact time the toxic behaviour happens (Scuri, 2019). In December 2020, the official Twitter account of FaceIt, the developer of the Minerva system, shared a tweet with a visual explaining the updates of the system: “Minerva can now detect toxic behaviour in voice chat, voice detection works in all languages, the system detected more than 1.9 million toxic messages since launch, and enabled a 62 percent reduction in seriously offensive messages” (FaceIt, 2020). Moreover, this system is expected to learn from the reports and make its own decisions and predictions on toxic behaviour during games soon (Scuri, 2019).

Text Mining Models for Online Game Content

Manual and human moderators remain sufficient for real-time identification and in-depth analysis of negative online behavior and prevent toxicity in online platforms and social networks (Blackburn and Kwak, 2015). Therefore, there has been an urge to develop methods to automatically detect toxic content (Blackburn and Kwak, 2014; Martens et al., 2015; Mayr et al., 2016; Singh et al., 2017; Chen et al., 2017; Gamback and Sikdar, 2017; Liu et al., 2018; Gomez et al., 2020; Yousefi and Emmanouilidou, 2021). Recently, Yousefi and Emmanouilidou (2021) developed an audio-based toxic language classifier for English language using Microsoft's self-attentive Convolutional Neural Networks (CNNs).

Jigsaw (2021) suggests that tagging comments in training AI is easier to understand for annotators. Many of the studies on language toxicity in social networks are primarily for the English language (Davidson et al., 2017; Wulczyn et al., 2017; Founta et al. 2019; Mandl et al., 2019; Zampieri et al., 2019b; Leite et al., 2020). Hosam (2019) identified toxic comments in Arabic social media by machine learning and mainly by gradient boosting technique (XGBoost algorithm).

Though it seems like a relatively niche area, there are in fact quite a few studies which explore the ways of automating and speeding up the detecting of toxicity in on-line games. Table 2 presents a summary of the studies which proposed a text mining modelling method

for detecting online game toxicity. We use this background in order to develop a specific model for detecting sexist toxicity in on-line gaming.

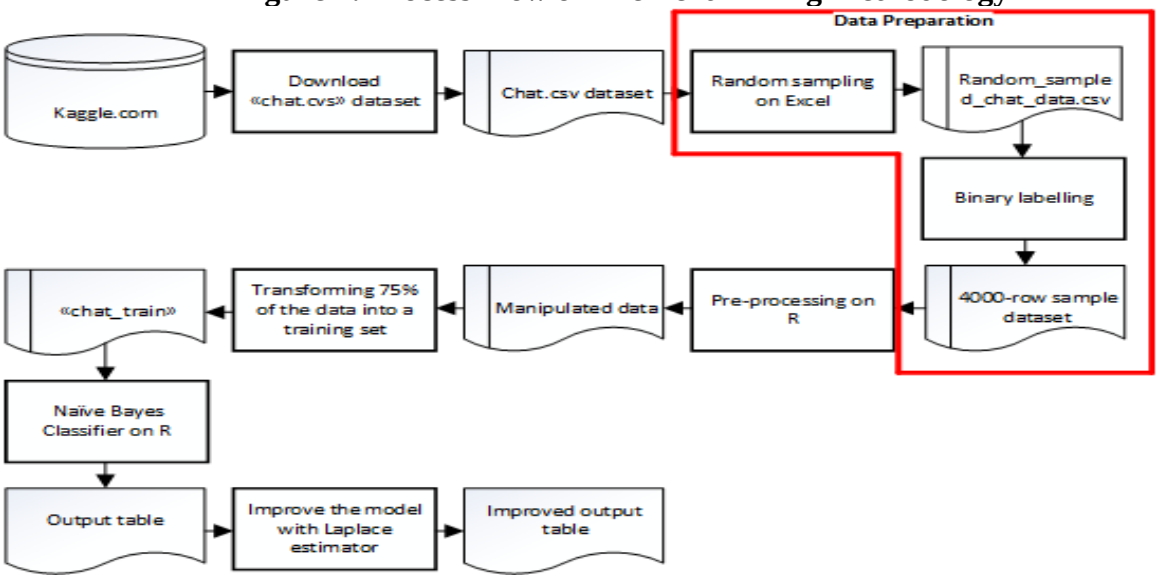
Table 2: Studies on Text Mining for Toxicity and Sexism

Authors	Application Area
Martens et al. (2015).	This paper suggests an annotation system to classify the most used words. Using a novel natural language processing framework, the model detects profanity in chat logs of a Multiplayer Online Battle Arena game and develops a technique to classify toxic judgements and n-gram to gather the semiotic toxicity
Blackburn and Kwak, (2014).	This research proposes a supervised learning method to predict crowdsourced decisions on toxic behaviours in one of the most popular online games- The League of Legends.
Murnion et al. (2018).	Classification of chat log data of an online game carried out using simple SQL query, AI-based sentiment text analysis and custom-built classification client.
Grosz and, Conde-Cespedes, (2020)	Natural language processing and deep learning-based model automatically detect if the statements are sexist or not. The suggested model has seven versions by different combinations of NLP and deep learning techniques.
Hui et al. (2008)	"IMAnalysis" that enables intelligent hat message analysis using text mining techniques k-Nearest Neighbours, Naive Bayes and Linear Support Vector Machine
Chen et al. (2017)	Text mining for the detection of abusive content in social media

Methodology

We use a two-step methodology. First, we use MAXQDA, a tool for data visualization, in order to present the frequency and types of toxic sexist terminology used in the game in a visual manner. In the second step, we use Naïve Bayes classifier in order to detect sexist toxic language in on-line gaming chat logs. This methodology is shown in Figure 1. The procedure starts with downloading the raw data from Kaggle.com in .cvs form, creating a “chat.cvs” dataset consisting of 1,048,576 rows. We drew a randomly sampled dataset from this chat.cvs data on Excel. A clustering technique whereby a label is assigned to each observation in the sample and the dataset is divided into disjoint clusters is called ‘a naïve' approach to the labelling problem (Plessis, Niu, and Sugiyama, 2013). Following this approach, each observation in the sample dataset was binary-labelled manually by the co-authors as “sexist” or “non-sexist” using the sexist toxic language identified in the literature as discussed above. The number of rows in the sample dataset was decided heuristically. After manipulating this sample dataset with pre-processing, 75 percent of it was transformed into the training dataset. A Naïve Bayes Classifier was used on our training dataset; the result was improved by adding the Laplace estimator to our model. The Laplace estimator adds a small number to each count in the frequency table, ensuring that each feature has a nonzero probability of occurring with each class. The Laplace estimator is typically set to 1, ensuring that each class-feature combination appears at least once in the data (Lantz, 2013) so that it eliminates the problem of zero probability in any class.

Figure 1: Process Flow of The Text Mining Methodology



Key performance Indicators – Keywords

KPIs (Key Performance Indicators) that this study utilized are derived from the literature presented in Table 3. The categorization of the words are filtered by these keywords (Lewis et al., 2004).

Table 3: Key Performance Indicators for Toxic Language Detection

KPI (Key Performance Indicators)	**Conclusion**	**Reference**
Number and rate of swear and offensive words used among the total chat data	According to the research of Murnion et al. (2018), approximately 13 per cent of the 126,000 messages collected for the research were found to contain swear words.	Murnion S., Buchanan, W., Smales, A. & Russell, G. (2018).
2. Emphasizing opponent's gender without reason (gender-oriented psychological pressure, harassment)	Most female gamers who present 75 women were very aware of the need to remain anonymous and conceal their identity and spoke about their recognition of the need to do so or suggested it as a tactic to cope with negative online behaviour	McLean and Griffiths (2019).
3. Ratio of leaving game earlier due to being subject of verbal/abuse/toxicity	According to INAP's survey (2019) involving more than 200 gamers and game developers, 20% of gamers and 31%of developers think that abucive game messages are the reason players stop playing multiple online games.	INAP (2019). The No. 1 Reason Gamers Will Quit Playing an Online Multiplayer Game (
4. The same X% of the players generates Y% of total verbal toxicity	For community analytics, ratios are far more helpful than just looking at absolute numbers.	Saclmans (2020). Why are ratios important

Our Data

Our data is a 1.048.576 row CSV file of chat logs of the online game called "Dota 2". Features of our dataset are "match_id", "key", "slot", "time", and "unit" which we only used "key", the chat entries of players in each row. The chat entries belong to 36,655 matches in total. This dataset is downloaded from Kaggle.com (Anzelmo, Dota 2 Matches).

Figure 2: Dataset downloaded from Kaggle.com

	A	B	C	D	E
1	match_id	key	slot	time	unit
2	0	force it	6	-8	6k Slayer
3	0	space created	1	5	Monkey
4	0	hah	1	6	Monkey
5	0	ez 500	6	9	6k Slayer
6	0	mvp ulti	4	934	Kira
7	0	bye	6	1486	6k Slayer
8	0	hah	1	1488	Monkey
9	0	fate	6	1496	6k Slayer
10	0	is cruel	6	1502	6k Slayer
11	0	fuck my ass	0	1524	Double T
12	0	ka bu tooooooooooooo	0	1721	Double T
13	0	wtf	1	1854	Monkey
14	0	TA?	1	1855	Monkey

Random Sampling

We first needed to create a sub-sample of our dataset in order to manually label each row whether the content entailed sexist toxic language. Each row of the dataset was assigned a random number with the Excel formula "RAND()". After freezing randomly assigned numbers to each row, we sorted the rows by those numbers from smallest to largest, so the rows were shuffled, and we sampled the first 4000 rows. We named this sample "random_sampled_chat_data.csv".

Labelling the Sample Dataset

After creating a 4000-row sample dataset, each row of this dataset was labelled by a binary coding dependent on the content. We referred to the toxicity definition of Jigsaw (2021) in this study. If the row contained sexist toxic language, the statement was labelled "1" and "0" if they are not. This evaluation is done by the authors separately, dividing the random sampled set into two.

Figure 3: Random sampled, labelled data (=RAND()Binary labelling)

	A	B	C	D	E	F	G
1	match_id	Binary	key	slot	time	unit	Random
2	24743	1	why r u running bitch??	9	2237	[STANDIN	0,001932
3	10466	0	take	2	1141	Realize	0,001933
4	24188	0	ZZZZ	7	256	Dog	0,001934
5	20146	0	gg	9	4200	kenshi	0,001935
6	34777	0	vote baltar!!!	3	2649	Gaius "Go	0,001937
7	1872	0	WOW	6	1166	Ondoy	0,001937
8	30701	0	cant even steal bh once	4	3826	Dr. Naifu	0,001938
9	27543	0	lol nabsters	6	2179	synerg.un	0,001939
10	13759	0	this slardar is talking lol	9	2562	BÂµRr420	0,00194
11	18530	0	?????	3	1259	serenity	0,001942
12	19035	0	we had a toilet break since u	0	2976	yormis	0,001943
13	607	0	4k ping	3	1946	1inchWon	0,001943

Data Pre-processing

A text corpus is created using the "key" feature of our dataset. Then, our data had to be cleaned before text-mining techniques were applied. The cleaning process steps remove non-Latin character rows, lower all the cases, remove the stop words, remove the numbers, remove the punctuations, and remove the whitespaces used unnecessarily.

Data Transformation

75 percent of the cleaned data is assigned to a "chat_train" variable to train our model and learn the patterns. The rest is assigned to a "chat_test" variable to test the accuracy of our model. Those test and train sets include the same ratios of binary entries with the primary dataset.

Text Mining Application by Naïve Bayes Classification Method

Classification can be simply defined as the process in which the output variables of relevant records are given to an algorithm to learn from those outputs to predict the unknown outputs of new records (Wikarsa et al., 2015). In that sense, classification is a supervised learning method.

Naïve Bayes algorithm is a simple application of Bayes theorem for classification where it has become a de-facto standard for text classification (Lantz, 2013). Naïve Bayes Classification's basic formulation is as below (Ren et al., 2009) ;

$$P(C_K \backslash x) = \frac{P(x \backslash C_K) * P(C_k)}{\sum_{k'} P(C_K \backslash x) * P(C_k)} \quad (1)$$

Where;

x = (x1, x2 , . . . , xd) is a d-dimensional instance which has no class label,

C = {C1, C2, . . . , Ck} is the set of the class labels,

P(Ck) is the prior probability of Ck (k = 1, 2, ..., k) that are deduced before new evidence, P(x\Ck) is the conditional probability of seeing the evidence x if the hypothesis Ck is true

A Naïve Bayes classifier assumes that the value of a specific feature of a class is unrelated to the value of any other feature;

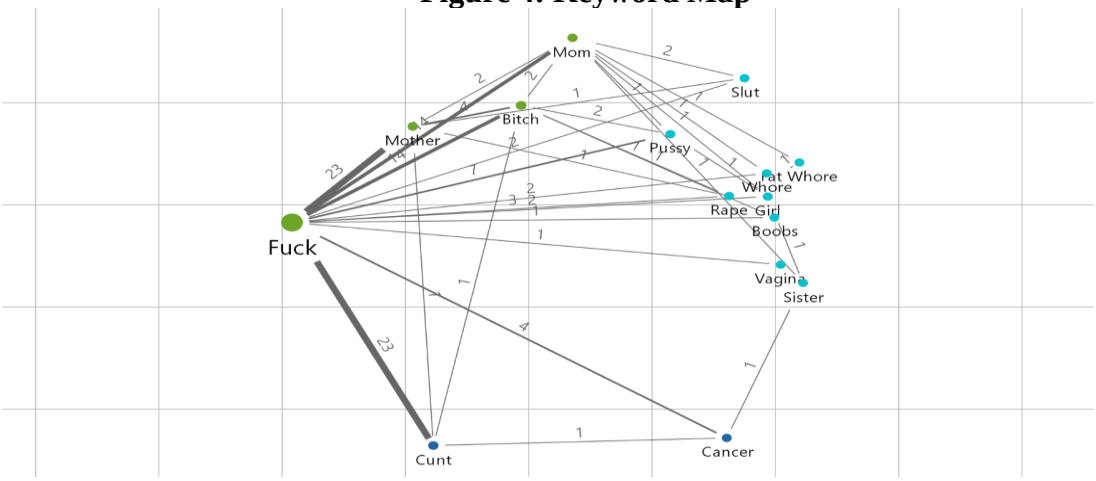
$$P(x \backslash C_K) = \prod_{j=1}^{d} P(x^j \backslash C^k) \quad (2)$$

We first run the model with Laplace=0. Then, to improve the results, we rerun the model with Laplace=1. Laplace estimator is used to add a small number to each frequency table, ensuring each feature would have a non-zero probability of occurring with each class (Lantz, 2013).

Findings from Data Visualization Analysis

As a result of our qualitative content analysis research and work on the data, we developed the Keyword Map in Figure 4. This illustration was created with the help of MAXQDA Tool. One thousand online game chat messages were scrutinized, and toxic sexist words were selected as keywords. The most used keywords are visualized according to their frequency and collocation in the game chat messages illustrated above. The display with three different colours shows that there are three different clusters according to their relationship. The words "Fuck", "Mother", "Bitch", "Mom" formed one cluster, while the words "Cunt" and "Cancer" formed another cluster. Also, the words "Slut", " Pussy ", "Fat Whore", "Whore", "Rape", " Girl ", " Boobs ", " Vagina ", " Sister " make up another cluster.

Figure 4: Keyword Map



Their proximity indicates that those words are built on top of each other, mainly used together. The lines between the words representa visualization of the relationships of the different words with each other. The dots, the sizes of the dots, and the frequency of the words in the text are parallel. For example, the term “Fuck” is shown with a big green dot as it is the most used word in the game chat messages. A number on the line indicates the frequency of use of the two words together. For instance, the words “Fuck” and “Mother” were used together 23 times in in-game chat messages. The thickness of the line between the words varies according to the use of the two words together. Due to the words “Fuck” and” Cunt” being used together 23 times, there is a thick line between them. 23 times represents the highest number that two words were used together.

Figure 5: Keyword Percentages

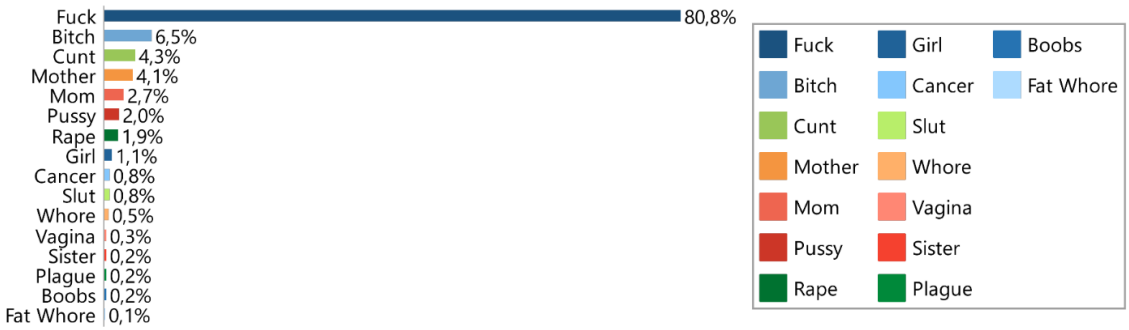


Figure 5 shows the frequency of the mostly used sexist words. Using the MAXQDA Tool, we sifted through 1000 online game chat messages to select the most frequently used sexist words, which served as keywords. This figure shows the percentage of each specified sexist keyword in total use of chosen sexist keywords. Unique colours are assigned to the keywords to explain the keywords more clearly. Accordingly the word “Fuck” is the most commonly used toxic sexist keyword with a rate of 80.8 percent of all usage.

Findings from Naïve Bayes Classifier Text Mining

Table 6 shows the accuracy rates of the Naïve Bayes Classifier for the text mining process by different Laplace estimators. 1 means “includes toxic content,” and 0 means “does not include toxic content”. As described above, after manually labelling almost 4000 rows according to their toxic content in accordance with the coding scheme from literature given in Table 1, we preprocessed this data. We removed numbers, punctuations, non-Latin characters

and the like. 75 percent of the cleaned data was used as the training set, and the rest was used as the testing set.

In our first attempt, the model predicted the non-toxic chat logs correctly with a rate of 99.7 percent, but the correct prediction rate for the chat logs including toxic, sexist was lower at 62.5 percent out of 657 meaningful rows. This result was achieved by a model with a Laplace estimator equals to 0. To improve the results, we ran the model once again with a Laplace estimator equal to 1. The model has been enhanced in terms of correct prediction of toxic content with a rate of 67.5 percent. However, its performance in terms of accurate prediction of non-toxic content was relatively worse with a rate of 94.5 percent. Running the model with a Laplace estimator to 2, predicted the non-toxic content with an accuracy rate of 73.6 percent and toxic content with an accuracy rate of 77.5 percent. As we inserted a higher degree for the Laplace estimator, the prediction accuracy for non-toxic language has declined, while the prediction accuracy for toxic language has increased.

Table 6: Accuracy Rates with and without Laplace Estimators

	Laplace 0			**Laplace 1**			**Laplace 2**		
n=167	***Actual***			***Actual***			***Actual***		
Predicted	***0***	***1***	***Row total***	***0***	***1***	***Row total***	***0***	***1***	***Row total***
0	615	15	630	583	13	596	454	9	463
	0,976	0,024	0,959	0,945	0,325	0,959	0,736	0,225	0,959
	0,997	0,375							
	2	25	27	34	27	61	163	31	194
1	0,074	0,926	0,041	0,055	0,675	0,041	0,264	0,775	0,041
	0,003	0,625							
	617	40	657	617	40	657	617	40	657
Column Total	0,939	0,061		0,939	0,061		0,939	0,061	

A summary of the results of the toxic chat log classification model is shown in Table 7. While the true negative ratio constantly decreased, the true positive has increased with every step of Laplace.

Table 7: Chat log classification model results

Laplace Estimator	***P(0/0)***	***P(1/1)***
0	99,70%	62,50%
1	94,50%	67,50%
2	73,60%	77,50%

Conclusion

Differentiating from the previous studies, this article aimed to propose an analytical model that identifies particularly sexist toxic language by online game players via using a hybrid method composed of text mining and qualitative data analysis. Results from both methods proved that the gender discriminative toxic language is evident in online gaming chat logs. However, there are no existing measures in on-line gaming to prevent or at least moderate toxic and sexist language. As revealed in previous studies, the toxic language has a correlation with behavioral patterns, and majority of toxic language speakers are young males and intense players (Lemercier-Dugarin et al., 2021). Toxic language analytics, in this context, offer valuable tools not only for the non-toxification of digital communities but also for preventing the bullying games, if accompained by responsible and active moderation.

In the prediction model, we ran the Naïve Bayes Classification method with an R coding language, classifying new data according to the labelled training data. Additionally, we added the Laplace estimator as recommended by Lantz (2013) to increase the accuracy of our model, which can be a methodological contribution to similar practices. Our model predicted non-toxic and toxic sexist content with an accuracy rate of 99,7% and 62,5 % in the test data set, respectively, with a Laplace estimator equals 0, 94,5% and 67,5% respectively with Laplace estimator equals 1 and lastly 73,6% and 77,5% respectively with Laplace estimator equals 2. These results reveal that the Laplace estimator did not correspond to the expectation that it would increase accuracy, i.e. Laplace did not correct the model as expected. While the true negative ratio constantly decreased, the true positive has increased with every step of Laplace. This finding should also be considered in such applications and in future research to find the most suitable Laplace number for the case in hand. In addition to that, we ran MAXQDA tool to visualize the relations of the toxic keywords considering the usage of them in this context. MAXQDA revealed 3 major Keyword Clusters with Fuck, Cunt and Whore as the lead keywords indicating the dose of sexism in male dominated e-gaming culture.

Though the above analysis provided some interesting insights, it would have been possible to achieve more accurate and more grounded outcomes if our database was more recent and had extensive features. It would be possible to improve the analysis by cooperating with online game developers, online game publishers, women's associations, NGOs, e-sports sponsor brands, and even governments and international organizations. The proposed analytical model can be adapted also to identify and prevent racist and other similar discriminative contexts in future research on toxicity in online games.

References

Anzelmo, D. (2016). "Dota 2 Matches". [Data file]. Retrieved at August 17, 2020 from https://www.kaggle.com/devinanzelmo/dota-2-matches/metadatafat

Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F. M., Rosso, P., & Sanguinetti, M. (2019). SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. *In Proceedings of the 13th International Workshop on Semantic Evaluation. Proceedings of the 13th International Workshop on Semantic Evaluation. Association for Computational Linguistics.* https://doi.org/10.18653/v1/s19-2007

Ballard, M. E., & Welch, K. M. (2017). Virtual warfare: Cyberbullying and cyber-victimization in MMOG play. *Games and Culture: A Journal of Interactive Media, 12*(5), 466–491. https://doi.org/10.1177/1555412015592473

Blackburn J. ; and H. Kwak, (2014). "Stfu noob! predicting crowdsourced decisionson toxic behavior in online games," in Proceedings of the 23rd international conference on World wide web, pp. 877–888.

Blumer, H. (1969). *Symbolic interactionism: Perspective and methods.* Englewood Cliffs, NJ: Prentice Hall.

Bryter (2019). *Female gamers survey 2019*. Bryter-Research. www.bryter-research.co.uk

Chen, H. ; S. Mckeever, and S. J. Delany (2017). "Harnessing the power of text mining for the detection of abusive content in social media," in *Advances in Computational Intelligence Systems*. Springer, pp. 187–205.

Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., & Guerini, M. (2019). CONAN - COunter NArratives through Nichesourcing: a Multilingual Dataset of Responses to Fight Online Hate Speech. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Proceedings of the 57th Annual Meeting of the*

Association for Computational Linguistics. Association for Computational Linguistics. https://doi.org/10.18653/v1/p19-1271

Ciampaglia, G. L. (2019). Can online gaming ditch its sexist ways?, The Conversation, August 27, Retrieved at December 27, 2020, from https://theconversation.com/can-online-gaming-ditch-its-sexist-ways-74493

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017, May). Automated hate speech detection and the problem of offensive language. *In Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 11, No. 1, pp. 512-515).

Faceit. [@faceit](December, 2020). Product Update Summary 👇 👂 Minerva can now detect toxic behaviour in voice chat, Voice detection works in all languages. Detected more than 1.9 million toxic messages since launch Observed a 62% reduction in seriously offensive messages being sent pic.twitter.com/HWfnoXODs8. Twitter. https://twitter.com/FACEIT/status/1338878980586352640?s=20

Fan, H., Du, W.; Dahou, A., Ewees, A.A.; Yousri, D.; Elaziz, M.A., Elsheikh, A.H.; Abualigah, L., Al-qaness, M.A.A (2021). Social MediaToxicity Classification Using Deep Learning: Real-World Application UK Brexit. Electronics, Vol. 10, 1332. https://doi.org/10.3390/ electronics10111332

Felmlee, D., Rodis, P. I., & Zhang, A. (2019). Sexist Slurs: Reinforcing Feminine Stereotypes Online. *Sex Roles*, Vol. 83 Issue 1-2, pp.16-28. doi:10.1007/s11199-019-01095-z

Fletcher, J. (2012). Sexual harassment in the world of video gaming, BBC World Service

Founta, A.M.; D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis. (2019). "A unified deep learning architecture for abuse detection," in *Proceedings of the 10th ACM Conference on Web Science*, pp.105–114.

Gamb¨ack, B. ; and U. K. Sikdar. (2017). "Using convolutional neural networks to classify hate-speech," in Proceedings of the 1st Workshop on Abusive Language Online, pp. 85–90.

Griffiths, M. (2010). "Gaming in social networking sites: a growing concern?" *World Online Gambling Law Report*, vol. 9, no. 5, pp. 12–13, 2010.

Grosz, D., & Conde-Cespedes, P. (2020). Automatic Detection of Sexist Statements Commonly Used at the Workplace. In Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD), LDRC Workshop.

Hobbs, P. (2013). Fuck as a metaphor for male sexual aggression. *Gender and Language*. . 10.1558/genl.v7i2.149

Hosam, O. (2019). Toxic comments identification in arabic social media. *Int. J. Comput. Inf. Syst. Ind. Manage. Appl*, 219-226.

Hui, S. C., He, Y., Dong, H. (2008). Text mining for chat message analysis. 2008 *IEEE Conference on Cybernetics and Intelligent Systems*, Chengdu, , pp. 411-416, doi: 10.1109/ICCIS.2008.4670827.

INAP (2019). The No. 1 Reason Gamers Will Quit Playing an Online Multiplayer Game (2019, April 09). Retrieved at December 20, 2020 from https://www.inap.com/blog/top-reason-gamers-quit-playing-online-multiplayer-game/

Jigsaw, 2021. No. 003: Toxicity, https://jigsaw.google.com/the-current/toxicity/

Johnson, K. (2014). Overt and Inferential Sexist Language in the Video Game Industry (Doctoral dissertation, University of Oregon).

Khandaker, J. (2019). Girl Gamers and Toxicity (Doctoral dissertation). The Faculty of the Department of Sociology University of Houston.

Kleinman, S., Ezzell, M.B., Frost, A. C. (2009). "Reclaiming Critical Analysis: The Social Harms of 'Bitch.'" *Sociological Analysis*, Vol: 3, Issue: 1, pp:46–68.

Kordyaka, B., Klesel, M., & Jahn, K. (2019). Perpetrators in League of Legends: Scale Development and Validation of Toxic Behavior. Proceedings of the *52nd Hawaii International Conference on System Sciences.* doi:10.24251/hicss.2019.299

Kwak, H., & Blackburn, J. (2015). Linguistic Analysis of Toxic Behavior in an Online Video Game. *Lecture Notes in Computer Science Social Informatics*, pp: 209-217. doi:10.1007/978-3-319-15168-7_26

Kwak, H., Blackburn, J., & Han, S. (2015). Exploring Cyberbullying and Other Toxic Behavior in Team Competition Online Games. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. doi:10.1145/2702123.2702529

Lantz, B. (2013). Chapter 4: Probabilistic Learning – Classification Using Naive Bayes. In B. Lantz, Machine Learning with R (pp. 89 - 117). Birmingham, UK.: Packt Publishing Ltd.

Lapolla, M. (2020). Tackling Toxicity: Identifying and Addressing Toxic Behavior in Online Video Games. Master Project, Master of Arts in Public Relations Seton Hall University

Leite, J.A., Silva, D.F., Bontcheva, K. Scarton, C. (2020) Toxic language detection in social media for Brazilian Portuguese : new dataset and multilingual analysis. *In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing.* 10th International Joint Conference on Natural Language Processing - AACL-IJCNLP 2020, 04-07 Dec 2020, Suzhou, China (online). Association for Computational Linguistics (ACL) , pp. 914-924. ISBN 9781952148910

Lemercier-Dugarin, M., Romo, L., Tijus, C. and Zerhouni, O. (2021). ‘‘Who Are the Cyka Blyat?’’ How Empathy, Impulsivity, and Motivations to Play Predict Aggressive Behaviors in Multiplayer Online Games, *Cyberpsychology, Behavior, And Social Networking*, Vol:24 Issue: 1, DOI: 10.1089/cyber.2020.0041

Lewis, D. D., Yang, Y., Rose, T., & and Li, F. (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research*, Vol: 5, pp: 361-397

Liu,P, Guberman, J, Hemphill, L, and A. Culotta, “Forecasting the presence and intensity of hostility on instagram using linguistic and social features,” arXiv preprint arXiv:1804.06759, 2018.

M¨artens, M. ; S. Shen, A. Iosup, and F. Kuipers (2015). “Toxicity detection in multiplayer online games,” in *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pp. 1–6.

Mai, I.; Marwan, T.; Nagwa, E.M. (2018). Imbalanced Toxic Comments Classification Using Data Augmentation and Deep Learning. In *Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications* (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 875–878.

Mandl, T., Modha, S., Majumder, P., Patel, D., Dave, M., Mandlia, C., & Patel, A. (2019, December). Overview of the hasoc track at fire 2019: Hate speech and offensive content identification in indo-european languages. *In Proceedings of the 11th forum for information retrieval evaluation* (pp. 14-17).

Marwa, T.; O. Salima, and M. Souham (2018). “Deep learning for online harassment detection in tweets,” in 2018 *3rd International Conference on Pattern Analysis and Intelligent Systems (PAIS*). IEEE, pp. 1–5.

Mayr, A; G. Klambauer, T. Unterthiner, and S. Hochreiter (2016). “Deeptox: toxicity prediction using deep learning,” *Frontiers in Environm. Science*, Vol: 3, pp. 80-92.

McDaniel, M A (2016). "Women in Gaming: A Study of Female Players' Experiences in Online FPS Games" (2016). *Honors Theses*. 427. https://aquila.usm.edu/honors_theses/427

Mclean, L., & Griffiths, M. D. (2018). Female Gamers' Experience of Online Harassment and Social Support in Online Gaming: A Qualitative Study. *International Journal of Mental Health and Addiction*, Vol: 17, Issue: 4, pp: 970-994. doi:10.1007/s11469-018-9962-0

Mead, G.H. (1934). *Mind, Self, and Society from the Standpoint of a Social Behaviorist.* University of Chicago Press: Chicago.

Murnion, S., Buchanan, W. J., Smales, A., Russel, G. (2018). Machine learning and semantic analysis of in-game chat of cyberbullying. *Computers and Security*, Vol: 76, pp.197-213.

Nexus. (2018). *Ask Riot: Will Tribunal Return?* . https://nexus.leagueoflegends.com/en-us/2018/08/ask-riot-will-tribunal-return/

Pina, A., Gannon, T. A., Saunders, B. (2009) An overview of the literature on sexual harassment: *Perpetrator, theory, andtreatment issues. Aggression and Violent Behavior 14* (2009) 126–138

Plessis, M. C. d., Niu, G., and Sugiyama, M. (2013). Clustering unclustered data: Unsupervised Binary Labeling of Two Datasets Having Different Class Balances, 2013 *Conference on Technologies and Applications of Artificial Intelligence*, 2013, pp. 1-6, doi: 10.1109/TAAI.2013.15.

Pkwzimm (2018). Sexism in Twitch chat: Comparing audience language for male and female streamers. Retrieved at December 27, 2020, from https://principallyuncertain.com/2018/03/06/sexism-twitch-chat/

Ren, J., Lee, S., Chen, X., Kao, B., Cheng, R., & Cheung, D. (2009). Naive Bayes Classification of Uncertain Data. *2009 9th IEEE International Conference On Data Mining.* https://doi.org/10.1109/icdm.2009.90

Risch, J.; Krestel, R. (2018). Aggression identification using deep learning and data augmentation. In P*roceedings of the First Workshop on Trolling, Aggression and Cyberbullying* (TRAC-2018), Santa Fe, NM, USA, August 25, 2018; pp: 150–158.

Rose, K. (April, 2018). Everyday Misogyny: 122 Subtly Sexist Words about Women (and what to do about them), April 12, 2018. Retrieved at December 27, 2020, from http://sacraparental.com/2016/05/14/everyday-misogyny-122-subtly-sexist-words-women/

Sackmans Co., (2020). Why are ratios important? (Retrieved at December 20, 2020 from https://www.sackmans.co.uk/why-are-ratios-important/)

Salawu, S.; Y. He, and J. Lumsden. (2017). "Approaches to automated detection of cyberbullying: A survey," *IEEE Transactions on Affective Computing*, Vol: 11, Issue:1, pp: 3-24.

Scuri, M. (2019). Introducing the in-game reporting feature to enable our Community to work with Minerva. Retrieved December 29, 2020, from https://blog.faceit.com/introducing-the-in-game-reporting-feature-to-enable-our-community-to-work-with-minerva-566439f0727

Schwalbe, M. 2008. *The Sociologically Examined Life: Pieces of the Conversation.* Fourth Edition. New York: McGraw-Hill.

Singh, V. K., S. Ghosh, and C. Jose (2017). "Toward multimodal cyberbullying detection," in Proceedings of the *2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, pp. 2090–2099.

Statista, (2019). U.S. video gamer gender statistics (2019) | Retrieved December 29, 2020, from https://www.statista.com/statistics/232383/gender-split-of-us-computer-and-video-gamers/

Subramani, S.; Michalska, S.;Wang, H.; Du, J.; Zhang, Y.; Shakeel, H. (2019). Deep Learning for Multi-Class Identification From Domestic Violence Online Posts. *IEEE Access*, Vol: 7, pp: 46210–46224.

Subramani, S., Wang, H., Vu, H.Q.; Li, G. (2018). Domestic violence crisis identification from facebook posts based on deep learning. *IEEE Access* 2018, Vol: 6, pp: 54075–54085.

Tang, W. Y., Reer, F., & Quandt, T. (2020). Investigating sexual harassment in online video games: How personality and context factors are related to toxic sexual behaviors against fellow players. *Aggressive Behavior*, Vol: 46, Issue:1, pp: 127-135. doi:10.1002/ab.21873

Tyack, P. Wyeth, and D. Johnson (2016). "The appeal of moba games: What makes people start, stay, and stop," in Proceedings of the 2016 *Annual Symposium on Computer-Human Interaction in Play*, 2016, pp. 313–325.

UN (2020). THE 17 GOALS of Sustainable development. Retrieved February 12, 2021, from https://sdgs.un.org/goals

Waseem, Z., & Hovy, D. (2016, June). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. *In Proceedings of the NAACL student research workshop* (pp. 88-93).

Whittaker, E.; Kowalski, R.M. (2015). Cyberbullying via social media. *Journal of School Violence*, Vol: 14, pp: 11-29.

Wikarsa L. and Thahir, S. N. (2015) "A text mining application of emotion classifications of Twitter's users using Naïve Bayes method," *2015 1st International Conference on Wireless and Telematics (ICWT)*, pp. 1-6, doi: 10.1109/ICWT.2015.7449218.

Wulczyn, E., Thain, N., & Dixon, L. (2017, April). Ex machina: Personal attacks seen at scale. *In Proceedings of the 26th international conference on world wide web* (pp. 1391-1399).

Yousefi, M. & 1:57Emmanouilidou, D. (2021). Audio-based Toxic Language Classification using Self-attentive Convolutional Neural Network. 2021 *29th European Signal Processing Conference (EUSIPCO)* | August 2021

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., & Kumar, R. (2019). Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). *arXiv preprint arXiv:1903.08983.*

Ziegelman, K. (2020). How Major Gaming Companies Are Combatting Toxicity in 2020. Retrieved December 29, 2020, from https://www.spectrumlabsai.com/the-blog/how-major-gaming-companies-are-combatting-toxicity-in-2020