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

**Aslı Ekiciler, İmran Ahioğlu, Nihan Yıldırım*, İpek İlkkaracan Ajas, and Tolga Kaya**

Department of Management Engineering, Istanbul Technical University, 34367 Istanbul, Turkey

*Corresponding author

## 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. In the online gaming communities, most women players report toxic and offensive language or verbal abuse against them. Observing and reporting the toxic behavior, sexism, harassment that occur as a critical need in preventing cyberbullying and help gender diversity and equality grow in the online gaming industry. However, the research on this topic is still rare, except for some milestone works. By the aim of contributing to the theory and practice of sexist toxic language detection in the online gaming community, we focus on the analysis and automatic detection of toxic comments in online gamers' communities context. As an analytical system proposal to reveal sexist toxic language in online gaming platforms, we adapted QCA by MaXQDA tool. Also, we applied Naïve Bayes Classifier for text mining to classify if a chat log content is sexist and toxic. We also refined the text mining model with Laplace estimator and re-tested the model's accuracy. Data visualization techniques also provided the most toxic words used against women in online gaming communities. The study also revealed that the NB classifier's accuracy rate did not change by the Laplace estimator. Findings are expected to raise awareness about gender-based toxic language usage. Applying the proposed mining model can inspire similar research and practical immediate solutions on easing the moderation and disinfection of these communities from gender-based discrimination and sexist bullying.

**Keywords:** Online Games Chat Logs, Toxic language, Sexism, Text Mining, Naïve Bayes Classifier

## 1. Introduction

Gender equality is one of the United Nation's 17 Sustainable Development Goals, and the UN (n.d) aims to eliminate all types of discrimination against women around the world. However, the harassment and language toxicity remains a dark reality in physical and digital communities, directed to women and non-binary individuals and all minorities. Among these, sexual harassment is characterized as unwanted sexual progress or other activity that targets someone based on their gender, including sexual harassment, lewd or discriminatory remarks, and sexual coercion, pressuring someone to perform sexual actions (Pina et al., 2009; Tang, Reer and Quandt, 2017). Symbolic interactionists (Blumer 1969; Mead 1934) and feminists assume that words matter, since any terms that dehumanize others can make it easier to harm others (Schwalbe 2008; Kleinman et al., 2009). On the other hand, Whittaker and Kowalski (2014) revealed that texting and social media are the most commonly used venues for cyberbullying. Also, various researchers explored the topic of hate speech (Liete 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 embedding models which automatically identifies 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, insult). 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 (Liete 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) model 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 men-dominated field which attracts diverse backgrounds. The online gaming industry is generally dominated by male users such as game developers, game players, game investors, causing the non-inclusiveness of gender diversity and equality, which resides as a salient problem in the community. According to the Female Gamer Survey of Bryter (2019), one in three women gamers were exposed to harassment or discrimination by male gamers. In the online gaming communities, most women players state offensive language or verbal abuse

against them. Abusive language is the most reported type of toxic behaviour, with a rate of almost 85% in 1 million reports (Kwak et al., 2015). As seen, offensive language, verbal abuse, toxic language including sexism, and harassment are big problems that impact the online gaming industry, especially women game players. From this point of view, observing and reporting the toxic behaviour, sexism, harassment that occur as 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 Kwark (2014), Kwark et al., (2015) and Martens et al. (2015).

Aiming to contribute to the theory and practice of sexist toxic language detection in social networks, we focused on analysing and automatically detecting toxic comments in online gamers communities context. This article aimed to propose an analytical system that reveals verbal abuse in online gaming platforms, raising awareness about gender-based toxic language usage and triggering efforts to prevent discrimination in digital communities. Consequently, we want to address this problem by analyzing the chat messages sent during online games to highlight the toxic and sexist language by utilizing text mining methods to publicly report the results of this analysis to all the shareholders.

The philosophy behind this work is to silence the harassers and create a social consciousness by increasing society's awareness in general and ensuring that this social consciousness creates its sanction power. For this, we first retrieved an online game chat log data that we can conduct our analysis. By applying a classification method, the machine-learning model examines this data, learns from it, and reports new data (chats) concerning the sexist discourse it contains according to the model it learned. We first labelled content with the toxic, sexist language in our sample database as sexist / non-sexist in a randomly generated sample according to the keywords we provided in the literature review. We ran the Naïve Bayes Classification method with an R coding language, classifying new data according to the labelled training data. Online game developers and publishers naively try to address this problem; however, people still advocate sexist and obscene language as a 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 online chat messages can be and raise awareness in those who contribute to this toxic environment by forming a generic type of toxic and discriminative language detection tool.

## 2. Literature Review

### 2.1. Toxicity and Sexist Toxic Language Research in Social Networks: Daring beyond "Negative online Behavior" or not?

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. 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; Martenz 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). Jigsaw's (2021) defines the term of toxicity as a "rude, disrespectful, or unreasonable language that is likely to make someone leave a discussion". Jigsaw recommends using it opposed to words like "abuse" or "hate speech" for both practical and scientific reasons since toxicity refers to a broad category of language, subject to individual interpretation". They also suggests that tagging comments in training AI is easier to understand for annotators. Most of the previous research focused on social media toxicity. Previous work on the language toxicity in social networks are primarily for English (Davidson et al., 2017; Wulczyn et al., 2017; Founta et al., 2018; Mandl et al., 2019; Zampieri et al., 2019b; Leite et al., 2020). Hosam (2019) identified toxic comments in Arabic social media by ML and mainly by gradient boosting technique (XGBoost algorithm). Recently, Yousefi and Emmanouilidou (2021) developed an audio-based toxic language classifier using Microsoft's self-attentive Convolutional Neural Networks (CNNs).

### 2.2. Language Toxicity in Online Games

With the participation of billions of people in it, 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). 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 by using games on player interactions and harming the group (Blackburn and Kwak, 2014). Besides, the effect of toxic players on the gaming industry is problematic. Toxic language is widespread in online games as a means of

releasing anger and disappointment in a destructive manner (Jahn, Klesel and Kordyaka, 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.** The gaming industry is a growing area that includes gender diversity. Nowadays, it is identified as a men-dominated community due to gender stereotyping, but this identification changes by increasing numbers of women game players day-to-day. 41% of computer gamers and video gamers in the USA are women gamers. The gaming industry is men-dominated. Hence the women game players' experiences play an essential role in understanding and analysing the community thoroughly (Statista.com 2020). While 79.4% of women agree that video games can be inspiring, 75.9% say they are abused or discriminated against during online games (McDaniel, 2016). Girl gamers experience additional toxicity aimed at them because of gender and the general toxicity that video game players experience (Khandaker, 2019). Toxicity includes sexual harassment, discrimination, and verbal abuse, are the main salient negative experiences that women game players face. The classes 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 (Ballard and Welch, 2017). That is to say, and most women game players have negative experiences in the gaming community.

**Toxic Keywords in Sexist Discriminative Language:** Below literature review gathers 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 since the word fuck 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 discussion on “the directness and specificity of its reference to bodily penetration“, we classified this word as “toxic” by its grounds and underlying masculinity patterns as an expression of gender-based discrimination.

Despite being used in a way to express “strong” or “fancy” females in daily language, “bitch“ is also classified as a sexist hence a toxic word, as Kleinman et al. (2009) defined it as a term that reproduces sexism, and using it (even by women) reinforces the idea that women are essentially different from men and in a negative way.

### 2.3. Past and Current Solutions

As the concept of online gaming becomes more and more social and interaction between the players' increases, one core issue occurs: toxic behaviour among players. It was inevitable for gaming companies to recognise this as a concrete problem, 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 (Will Tribunal Return Back, 2018). As cited in Lapolla, Kou and Nardi (2014) state that as of May 2011, the Tribunal was formed, and over 47 million votes were cast in its first year evaluating toxic behaviour; 74% of players facing punitive toxicity interventions subsequently changed their in-game behavior. Blizzard, the owner of Overwatch, tried a different approach to stop toxic behaviour and 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% in 2019 (Ziegelman, 2020). CS:GO’s producer Valve has also taken a few steps to fight against 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). The other 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, 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% 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).

### 2.4. Text Mining Models for Online Game Content: Literature Review

Though it seems like a relative niche area, many studies searched for automating and speeding up detecting online games' toxicity. Some of the researches which proposed a text mining modelling for detecting online game toxicity are as follows;

**Table 2.** Model Literature Review

| **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 |

## 3. Methodology

In our application's 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. Data preparation was initialized with random sampling from chat.cvs dataset on Excel to get a randomly sampled dataset. Using a clustering technique to assign a label to each dataset sample to divide the dataset into disjoint clusters is a naive approach to the labelling problem (Plessis, Niu, and Sugiyama, 2013). So, this sample dataset was binary-labelled manually by co-authors as "sexist" or "non-sexist" regarding sexism based on toxic language literature provided before. The number of rows in the sample dataset was decided heuristically. After manipulating this sample dataset with pre-processing, 75% 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.

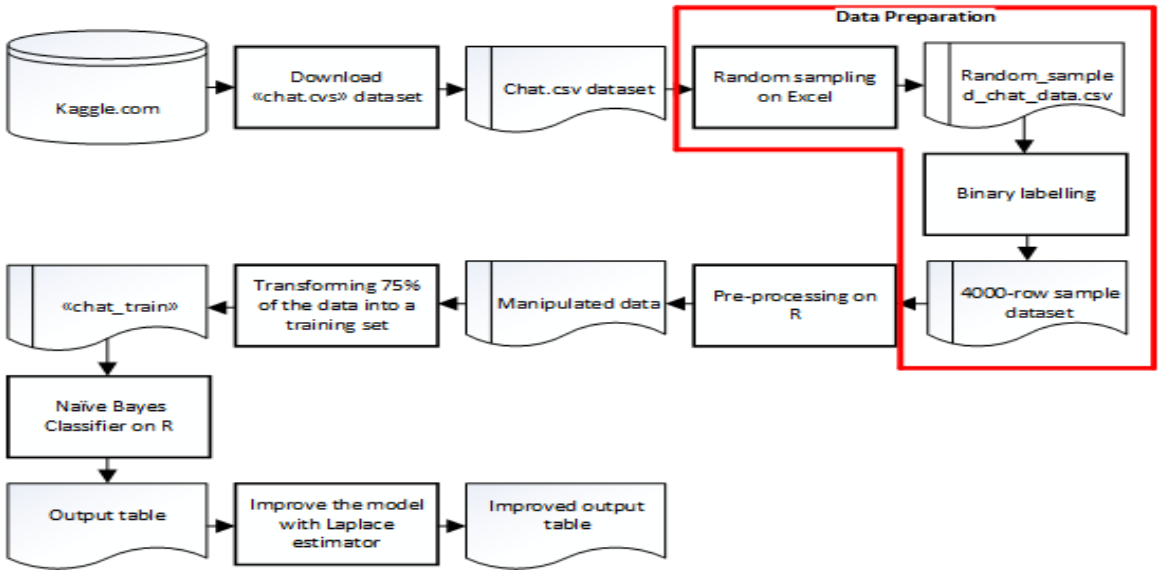


**Figure 1.** Process Flow of The Text Mining Methodology

### 3.1. 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 |

### 3.2. 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 36655 matches in total. This dataset is downloaded from Kaggle.com Anzelmo, Dota 2 Matches).

| | 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 |

**Figure 2.** Dataset downloaded from Kaggle.com

### 3.3. Random Sampling

We first needed to create a sub-sample of our dataset to be labelled manually in each row's sexist, toxic language, including content. Therefore, each chat entry contains. By this aim, each row of the dataset was assigned a random number with Excel's "RAND()" formula. 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".

### 3.4. Labelling the Sample Dataset

After creating a 4000-row sample dataset, each row of this dataset was labelled binary as if the row's content; therefore, the chat entry uses toxic language according to keywords provided in the Literature Review section. We referred to the toxicity definition of Jigsaw (2021) in this study If toxic, the statements are labelled "1" and "0" if they are not. This evaluation is done by the authors separately, dividing the random sampled set into two.

| | 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 |

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

### 3.5. 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

### 3.6. Data Transformation

75% 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.

### 3.7. Text Mining Application

#### 3.7.1. Naïve Bayes Classification Method

Classification can be simply defined as the process in which the output variables of interested 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 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).

## 4. Findings from Qualitative Data Analysis

As a result of our qualitative content analysis research and work on the data, we had achieved the Keyword Map in Figure 4. This illustration was created with the help of MAXQDA Tool. Online game chat messages of 1000 online game matches were scrutinized, and sexist words were selected as keywords. The most used keywords are associated according to their use 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.

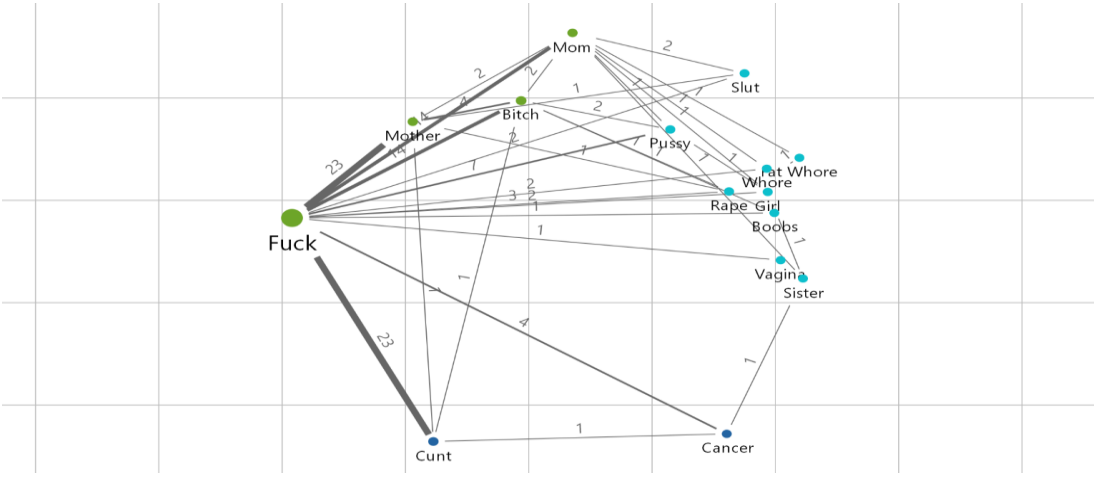


**Fig. 4.** Keyword Map

Their proximity indicates that those words are built on top of each other, mainly used together. The lines between the words are the visualizations of the relationships of words with each other. The dots, the dots' size, 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 usage is the highest number of uses compared to using other words with each other.

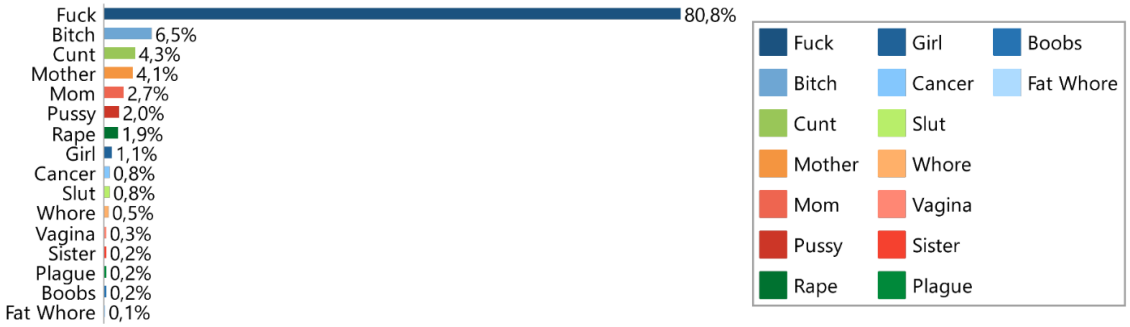


**Fig. 5.** Keyword Percentages

Figure 5 was created with the help of the MAXQDA Tool. Online game chat messages of 1000 online game matches were sifted through, and the mostly used sexist words were selected as keywords. This figure shows the percentage of each specified sexist keyword in total chosen sexist keywords. Unique colours are assigned to the keywords to explain the keywords more clearly. Keyword Percentages in Figure 5 shows that 80,8% of all sexist keywords used is the word “Fuck”.

## 5. Findings from Naïve Bayes Classifier Text Mining Results:

By using Naïve Bayes Classifier for the text mining process, accuracy percentages in Laplace 0 columns of Table 6 were achieved. 1 means “includes toxic content,” and 0 means “does not include toxic content”. 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 etc. 75% 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% and chat logs including toxic, sexist language with a rate of 62,5% out of 657 meaningful rows. A model with a Laplace estimator achieved these rates equals to 0. To improve the results, we run 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%. However, it has downgraded in terms of accurate prediction of non-toxic content with a rate of 94,5%. Running the model one last time with a Laplace estimator equals 2 yielded the results in Figure 10. This model has predicted non-toxic content with an accuracy rate of 73,6% and toxic content with an accuracy rate of 77,5%. As we inserted a higher Laplace estimator's degree, non-toxic prediction accuracy has declined as the toxic language prediction accuracy has inclined.

**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 | |

Finally, the toxic chat log classification model results in summary is given in Figure 11. 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% |

## 6. Conclusion

This article aimed to propose an analytical model that explains the sexist toxic language in the online gaming community by using a hybrid method composed of text mining and qualitative data analysis methods. Results from both methods proved that the gender discriminative toxic language is evident in online gaming chat logs. However, there are no existing preventive measures or moderation on this toxicity. As revealed in previous researches, 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 stopping the bullying games if empowered 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 our accuracy to the model, which can be a methodological contribution to similar practices. Our model predicted

non-toxic and toxic content at 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. However, results revealed that the Laplace estimator did not correspond to the accuracy increase expectation. Laplace didn't correct the model as expected. While the true negative ratio constantly decreased, the true positive has increased with every step of Laplace. 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. 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.

Though we had pleasing results from our work, we firmly believe that it would have more accurate and more grounded outcomes if our database was newer and had extensive features. By cooperating with the 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 to racist and similar discriminative contexts for 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

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.

Bryter (2019). Female Gamer Survey. Available: https://blog.bryter-research.co.uk/online-abuse-still-hindering-female-gamers-market

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.

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

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.

Gaikwad, S. V., Chaugule, A., & Patil, P. (2014). Text mining methods and techniques. *International Journal of Computer Applications*, Vol. 85 (17), pp: 17-27.

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.

GTZFUoS. (2011). Fat, Ugly or Slutty?, June 20, 2011. Retrieved December 29, 2020, from https://bit-tech.net/reviews/gaming/fat-ugly-or-slutty/1/

Gupta, V., & Lehal, G. S. (2009). A survey of text mining techniques and applications. *Journal of emerging technologies in Web Intelligence*, Vol. 1, Issue 1, pp: 60-76.

Hearst, M. (2003). What is text mining?, SIMS, UC Berkeley.

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

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

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.

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

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.

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/

Gomez, R. , Gibert, J., Gomez, L. and D. Karatzas (2020). "Exploring hate speech detection in multimodal publications," in *The IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 1470–1478.

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

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. (2019). 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

Whittaker, E.; Kowalski, R.M. (2015). Cyberbullying via social media. *J. Sch. 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.

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

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