A FRAMEWORK FOR UNDERSTANDING AND DETECTING HARASSMENT IN SOCIALVR

LANCE POWELL

BOĞAZİÇİ UNIVERSITY

2018

A FRAMEWORK FOR UNDERSTANDING AND DETECTING HARASSMENT IN SOCIALVR

Thesis submitted to the

Institute for Graduate Studies in Social Sciences

in partial fulfillment of the requirements for the degree of

Master of Arts

in

Cognitive Science

by

Lance Powell

Boğaziçi University

2018

DECLARATION OF ORIGINALITY

I, Lance Powell, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature .....

Date 11 - 12 - 2017

ABSTRACT

A Framework for Understanding

and Detecting Harassment in SocialVR

SocialVR, as experienced in immersive audiovisual environments, is a symmetrical communication medium that allows for both verbal interaction, and limited physical interaction through first-person avatars. Through a qualitative analysis of discourse among SocialVR users, this research finds examples of harassment and evidence for patterns in that harassment, advancing how we understand the current problem. In response, methods of recognizing user and environmental trends towards harassment are discussed. Informed by the qualitative data and literature on harassment in social media, natural language processing is used to classify speech as being harassment according to lexical and structural elements. When implemented by SocialVR platforms, this initial step can be added to and altered, making it an effective tool for preventing abuse among users. This research also provides a method for using convolutional neural networks to classify three-dimensional, vulgar imagery that is produced in SocialVR, narrowly targeting the most common forms of vulgarity. Using a CNN, a classification model is made, which can be used to remove unwanted images with 78% accuracy at testing. This framework includes recommendations on how data should be collected going forward, how data should be used, and the design considerations that should be made for both harassing and non-harassing alike.

Özet

Sosyal Sanal Gerçeklikte Tacizi Anlamaya ve Saptamaya Dair Bir Çerçeve

Sosyal Sanal Gerçeklik (SG), kapsayıcı görsel-işitsel çevreyle deneyimlenen, simetrik iletişim aracıdır. Sosyal SG, sözel etkileşime fırsat tanıyan ve kısıtlı da olsa, avatarlar aracılığıyla fiziksel etkileşimin kurulabildiği ortamlardır. Bu araştırmada, Sosyal SG kullanıcıları arasındaki söylem, nitel analiz yöntemiyle incelenerek, taciz örnekleri ve bunların gerçekleşme örüntüleri sunulmuştur. Bulunan bulgular ve örüntüler, güncel taciz problemini nasıl anlamamız gerektiğine de yardımcı olmaktadır. Bu doğrultuda, tacize yönelik kullanıcı ve çevre alışkanlıklarını tanıma metodlarına dair tartışmalar sunulmuştur. Nitel veri ve sosyal medyada taciz literatüründeki bilgiler ışığında, Doğal Dil İşleme kullanılarak konuşmalar, içerdikleri sözcüksel ya da yapısal taciz unsurlarına göre sınıflandırılmıştır. Bu ilk adım, Sosyal SG platformları tarafından uygulandığında, kullanıcılar arasındaki tacizi engelleme konusunda etkili bir araç olabilir. Bu araştırma, aynı zamanda Kıvrımlı Nöral Ağlar (KNA) ile Sosyal SG'de üretilen üç boyutlu bayağı görsel unsurları sınıflandırmak için de bir metod sunar. KNA kullanılarak, istenmeyen imajları test aşamasında %78 tutarlılık oranıyla ayıklamak için kullanılabilecek bir sınıflandırma modeli üretilmiştir. Bu çerçeve, daha ileri çalışmaları yürütebilmek için verilerin nasıl toplanması gerektiğine ve taciz içeren içerikleri diğerlerinden ayırmayı kolaylaştıran tasarım yöntemlerine dair öneriler de içerir.

ACKNOWLEDGEMENTS

My foremost thanks go to my advisers, Arzucan Özgür and Didar Akar, who believed in this strange union between discourse analysis and SocialVR under the umbrella of cognitive science. They have been the two pillars that kept this project standing. I'd like to show my sincerest appreciation to VRFirst at BAU and Crytek for the use of their VR laboratory throughout my research. Without their support, this research would not have been possible. I'd also like to thank the We Make Realities and Virtual World Society, founded and co-founded by Eva Hoerth, for making me aware of this issue and inspiring the thesis topic. Thanks to Shawn Whiting of Against Gravity for sharing his insights on the SocialVR industry and talking over several of my notions regarding SocialVR. Thank you to Lisa Kotecki for being a cheerleader for this anti-harassment research. Thanks also to Sami Hamid of Glitch Studios, who championed much of my writing on SocialVR, forcing me into a deeper understanding of interactive design and industry concerns. Thank you Arda Celebi for his hand in debugging a program and calling my thesis *cool*. Thanks also to Dr. Suzan Üsküdarlı who sparked an interest in virtual worlds even before SocialVR entered the picture and guided my thesis work throughout the process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
EEOC	Equal Employment Opportunity Commission
FN	False Negatives
FP	False Positives
HMD	Head Mounted Display
MCDA	Multimodal Critical Discourse Analysis
NLP	Natural Language Processing
RGB	Red Green Blue
SocialVR	Social Spaces in Virtual Reality
TN	True Negatives
TP	True Positives
VR	Virtual Reality

CHAPTER 1

INTRODUCTION

Quickly and efficiently, computer mediated communication allows user engagement across an increasing number of modalities and environments. Once computer networks were built, and the needs for personal and professional communications were met, these networked users devised recreational uses for the computerized medium. Geographically separate people suddenly inhabited virtual locations built on raw text and their imaginative powers. As this early form of virtual reality became popular in the late 1980's and 1990's, the experience became tarnished when the abusive tendencies of some users surfaced. They were people who treated these spaces as platforms for slander, racial intolerance, and misogyny, forcing common users to create tactics to deal with them or to abandon the hostile platform. These early tendencies towards abusive behavior evolved with the interconnected technologies and harassment became commonplace alongside mainstream access to the internet and social media platforms (Duggan, 2017). By necessity, social media companies devote many of their resources to harassment prevention, but their success is limited to say the least. Discussions over technology and harassment stemming from this past are reignited due to the release of communication in virtual reality (VR) as mediated by head-mounted displays (HMDs) in 2015 and 2016, allowing users audiovisual interactions with one another in a shared, image-driven virtual space. Interactions between users in VR cannot be completely understood in the context of social media or real life, creating a need for more research

in the new medium. Harassment taking place in Social Spaces in VR (SocialVR) has also required new methods for their classification and detection.

This thesis will give context to harassment in SocialVR by describing the history of harassment, interactive technologies, and the effect they have had on one another. The current and former states of SocialVR as an immersive tool and communication medium will be explained, followed by several anti-harassment measures in place at the time of writing. These descriptions will act as a snapshot of SocialVR as it is developing rapidly. Currently, little data is available that would help optimize anti-harassment features since usership of SocialVR is very low relative to social media, data about individual users is generally not shared within a platform, some SocialVR platforms lack the resources for the proposed types of data collection, and the platforms large enough to collect actionable data may be unable to share it. Therefore, this thesis proposes methods of harassment detection based on current industry needs for simplicity and quick implementation. The methods are designed to be a starting point and proof of concept for gathering data and classifying harassment among users, which may be built upon for utilization in SocialVR platforms and further research of this type.

CHAPTER 2

A HISTORY OF HARASSMENT AND ITS CATEGORIES

This section will explore the forms harassment has taken during its history, both in terms of where and how it happens. A full explanation will give context to the types of harassment I will be dealing with in this thesis and make explicit the methods of harassment I am seeking to prevent.

2.1 Workplace harassment

Discussions and legal actions regarding harassment may have originated with workplace harassment since it predates the internet and it is an environment which brings together a diverse group of adults wherein there is generally a need to communicate and form relationships with others. Where a group of people who lack mutual understanding come together for a shared purpose, there lies the potential to create hostile environments either out of ignorance, apathy, or malicious intent. A hostile work environment is one in which employee behaviors lead to changes of emotional discomfort to the working environment or abuse (Rotundo, 2001). In this case, harassment would be classified as the behavior that leads to this environment and it can be severe enough to prompt emotional distress and the victim's resignation from their position resulting in lost wages.

2.1.1 Landmark harassment cases

As an example, one legal case finalized in 1993 comes from the US Supreme Court that tied harassment to the psychological injury of the victim and discriminatory workplace practices based on the employee's sex, race, national origin, or religion (Case no. 92-1168, Teresa Harris v. Forklift Systems, Inc.). In that case representing discriminatory practices against an employee's sex, the defendant had referred to the plaintiff as a 'dumb ass woman' and asked her to pick up items from the ground so he might look inappropriately at her body alongside other male colleagues (Epstein, 1995). A 1998 US Supreme Court case (Case no. 523 U.S. 75, Onacle v. Sundowner Offshore Services, Inc.) clarified sex-based workplace harassment as being possible against transgendered persons, between litigants of both the same and different genders, and other instances of gender non-conformity such as homosexuality. Cases of racially motivated harassment in the workplace have also been taken up and prosecuted by the US Supreme Court, including Vance v. Ball State (Case no. 570 U.S. (2013)). The behavior that led to this law suit involved a fellow employee intentionally blocking Vance's path, weirdly smiling at her as if to ridicule her, intentionally banging cookware around her, and making Ku Klux Klan references to her, but the judgement of this case clarified the employer responsibility for harassment cases where there is a power differential, such as the instigator is a supervisor of or has a supervisory role over the complainant (Woska, 2014).

2.1.2 Harassment definitions

The Equal Employment Opportunity Commission (EEOC) goes into greater detail on what constitutes sexual harassment at work by giving a list of potential scenarios along with the classification: verbal, non-verbal, or physical (EEOC, n.d.). Examples of verbal harassment might be sexually explicit jokes, questions, or even sounds while non-verbal harassment may include the exposure to pornographic material or inappropriate gestures. Physical harassment involves the touching of another person or oneself in a threatening and potentially sexual manner. According to the EEOC, what links these behaviors as examples of harassment may not be their sexual nature, but the fact that they are 'unwelcome' to the recipient. The victim of harassment need not outwardly protest the unwanted treatment. Whether the conduct is unwelcome, and thereby defined as harassment, depends on how the individual considers it.

2.2 Harassment in schools

Harassment concerns extend also to educational institutions, which are similar to the employment environment insofar as they include diverse sets of people who will be required to interact with one another. Students, with a varying knowledge of standards for behavior and typically lessened consequences for violating those standards, may find a greater likelihood of experiencing harassment as either the culprit or the victim. In university, where the parents may be uninvolved in such cases, there is typically a code of conduct that lays out a standard process for reporting breeches to this code. As an example, the University of California Berkley in their Code of Conduct defines

harassment as actions that prevent a student's participation in programs or activities in the university and extends the previously listed bases for harassment to include age, marital status, veteran status, and disabilities (Section V, Article 102.09). The Administrative Guide to Stanford University repeats this sentiment, describing harassment as '*unreasonable interference*' or the creation of a '*hostile environment'*, a term also used in the context of the workplace. Furthermore, harassment may occur on repeated occasions or on a single occasion if the infraction is particularly extreme (Stanford University, 2016).

2.3 Street harassment

It is also possible to find instances of harassment beyond the confines of formal institutions. Depending the location, gender-based harassment can be a common occurrence on the streets or in public settings and it has been given the name *street harassment*. Unlike the previously mentioned forms of institutional harassment, street harassment is a sexually motivated type of harassment done to women by people who are strangers to them (Bowman, 1993). As others have defined it before, this may include verbal, gestural, or physical assault intended to objectify or humiliate women (Peoples, 2008). It traditionally may receive less attention from lawmakers and academics than other forms of harassment since the perpetrators will typically be unknown and its potential harms, such as the feeling of being threatened or negative body image, are less quantifiable than the loss of career or academic opportunities. Likewise, countries such as the United States grant freedom of speech as a constitutional right and legislating street harassment could be interpreted as a violation of those rights

(Nielsen, 2000). However, the freedom of movement has long been argued as a civil right and the inducement of fear through sexually aggressive speech or actions would present a limitation to women's access to their basic human rights (Blackstone, 1915). Based on the US Supreme Court rulings in the aforementioned workplace harassment cases, I will include all sex-based harassment, not exclusively heterogeneous, as a candidate for street harassment.

2.4 Hate speech

The provided definition of street harassment does not adequately cover all forms of harassment occurring in a public place since it fails to include instances which are motivated by violence or the spread of hostility towards a specific group. For these instances, we will use the term *hate speech* which, according to Anne Weber of the Council of Europe's Committee of Ministers is '...understood as covering all forms of expression which spread, incite, promote or justify racial hatred, xenophobia, anti-Semitism or other forms of hatred based on intolerance, including: intolerance expressed by aggressive nationalism and ethnocentrism, discrimination and hostility against minorities, migrants and people of immigrant origin' (Weber, 2009).

Unlike street harassment, hate speech does not require that a member of the group under discussion be present. One person expressing ill feelings toward a group may be committing harassment by spreading their hostility among those people listening. The laws pertaining to hate speech vary from nation to nation, but United States law has been explicit in its consideration of hate speech as being protected

according to the First Amendment of the Constitution (Bleich, 2014). If a crime committed against another person is motivated by hatred towards a group, it may be considered a hate crime and the perpetrator will receive an unfavorably modified sentence base on those overt declarations, but having and expressing those sentiments are not punishable by law.

2.5 Harassment summary

Having covered different types of harassment and the importance of the location in which they occur, it might be helpful to illustrate the differences here. Sex-based harassment between people of different races is still street harassment so long as race is not the focal point for that harassment and it happens outside any formal institution, such as a workplace or school. A person who hypothetically discusses raping another person in a public setting is still performing verbally based street harassment since, even though rape is a form of physical assault, it is still spoken, sexually themed, and creates a threatening environment. On the other hand, if a White person in the presence of a Black person raises a fist to the sky, cocks their head, and sticks out their tongue, it is likely that this is mimicry of someone being hanged by a noose. Using United States history of lynching Black populations as the context, this is a strong candidate for both non-verbal harassment and hate speech. If one person corners another or a group of people crowds around an individual, this could be an instance of physical harassment, which could be sex-related or hate-related depending on the motivations of the crowd. These distinctions between different types of harassment will become important when discussing it in the context of virtual worlds where examples not unlike these were found.

CHAPTER 3

A DESCRIPTION OF VIRTUAL REALITIES FROM THE PAST AND PRESENT

In this section, virtual reality is under discussion with a brief description of its earliest electronic forms and an extensive discussion of its most modern iteration. The descriptions encompass not only the technology, but functionality and user identity as well.

3.1 Multi-user dungeons

Multi-user dungeons (MUDs) are online recreation centers and an early version of virtual reality in which visitors may act out a fantasy with multiple other users, where the imagined environment is built upon its text-only representation and the players assume an identity of their choosing. Though this venue of interaction has been long absent from the public eye, the number of multi-user games numbered in the hundreds back in 1993 and the type of verbal engagement among users ran the spectrum from mannerly to graphically sexual (Rheingold, 2000). Design decisions regarding the appearance of the collectively imagined environment might be based on an individual host of a dungeon or the environment could be designed by the whole group and in real time. Likewise, pseudonyms were generally used and, should they so choose, a user's physical appearance could be determined through description by the users themselves. The dungeons could be centered on a theme, such as the Medieval or Science Fiction, and it could be directed towards a common goal or game that users play together.

3.2 Virtual realities

As consumer versions of virtual reality headsets that immersed people in digital 360-degree environments were released throughout 2016, including: Vive, Oculus Rift, Samsung Gear, and PSVR, the association between MUDs and the term *virtual reality* has lessened considerably. VR has come to be an immersive, and perhaps interactive, visual environment mediated by HMDs and controllers. Users might be able to move within that space and there may also be audio components suited to the environment. There is 360-video content in which the user is a passive observer, but VR may also contain objects with which users interact. These objects cannot be spoken into existence as with MUDs. They must be designed and scripted, either by the developer or individual content creators, so they may be handled by the user and the objects may physically respond to one another. In short, a photo-realistically rendered, fully immersive, and fully interactive VR experience would be indistinguishable from reality as far as the eyes and ears are concerned (Steinicke, 2016).

3.3 Social Spaces in VR (SocialVR)

An experience in VR is defined here as SocialVR when multiple users simultaneously inhabit a virtual space where they are capable of interacting verbally and through the movement of their avatars. Variations of these virtual spaces in SocialVR include the simultaneous visitation of users to a single-user VR experience by means of an external program, but these are not in discussion here.

Table 1 Full List of SocialVR Platforms Mentioned in the Text

SocialVR Platforms		
Rec Room	Facebook Spaces	Anyland
AltspaceVR	VR Chat	Bigscreen Beta
High Fidelity	vTime	QuiVR
TheWaveVR	OrbusVR	Sansar
Pararea	EmbodyMe	Basement VR

3.3.1 Communication in SocialVR

Communicative interactions in SocialVR are generally spoken. Communication through text is sometimes available, but it is not often used because the act of typing with VR controllers is slow and attempting to type with a real keyboard while in VR may, at the moment, be problematic. Most SocialVR platforms attempt a faithful simulation of in-person speech, where its volume is loudest directly beside the speaking avatar and decreases the further from them one travels. If spacious enough or there is an obstruction, speech in the environment may be inaudible to other users who are an adequate distance from the source. There are cases in which speakers can use available objects to project their voice at an equal volume throughout the environment. These objects may be microphones or megaphones (*Rec Room*), which correlate to objects in material reality, or more fanciful objects like magical cookies (*Anyland*), which do not have a widely known correlation. Therefore, users of environments with a large enough capacity can reasonably expect to encounter multiple conversations happening simultaneously throughout the environment. Users will often join, leave, and rejoin

conversations, so the social norms of discourse, such as leave-taking, may not apply (Schourup, 2016). The breaking down of social norms is exacerbated when the platform or its users suffer technical failures which force them out of the conversation.

3.3.2 Movement in SocialVR

The movement of avatars can vary from user-to-user and platform-to-platform depending on the technology being utilized by users and the availability of its support within the SocialVR platform. Minimally, the movement of avatars will include their travel on the X-axis, Z-axis, and sometimes Y-axis. SocialVR platforms will often indicate when a user is speaking, either through manipulation of the avatar's mouth (*Rec Room*, *Facebook Spaces*) or a flashing light emitted by the avatar syllabically corresponding to the users' speech (*AltspaceVR*). Most SocialVR platforms use hand-held controllers whose movements are captured by sensors and represented as the users' hands within the virtual environment. Some SocialVR platforms have support for motion capture, which will interpret the body, head, and limb movements of users (*High Fidelity*, *AltspaceVR*). In virtual environments, users may sometimes move from point A to point B through continuous, linear travel, but teleportation between points is more common because it is faster and generally more comfortable, which is important for the avoidance of motion sickness (Bozgeyikli, 2016). In teleportation, users direct a cursor to a distant spot on the terrain and they are immediately transported to that point when the necessary input has been received.

3.3.3 Privacy in SocialVR

SocialVR experiences can be divided into two types, private and public. Private experiences, or events, within SocialVR are between the hosts and their invited guests. These private experiences are intended to bring together people who are already somehow connected through the platform, people who have begun a conversation and want to continue it uninterrupted, or people who have met elsewhere and connected in the SocialVR platform. People meeting privately may consider SocialVR a venue for being with preexisting friends and they exclude unknown users out of convenience, but it is also a sure way of preventing contact with harassing or otherwise unwanted users.

In most SocialVR platforms, experiences in public environments are the default because they do not require an additional setup process. These experiences take place in common areas which all users of the SocialVR platform may visit. In some cases, these users need not even be registered and will be designated as *guests*. Some SocialVR platforms put upper limits on the number of visitors allowed within a given environment, also referred to as a *room* or, more literally, *server*. For example, *Bigscreen Beta* and *vTime* have a capacity of four users per room. Meanwhile, other platforms will put as many users in an environment as the server will allow. For example, *AltspaceVR* and *VR Chat* might place twenty or more users in the same room. A user may or may not know the other users in their room and, while within capacity, users may freely travel between rooms.

3.3.4 Avatars in SocialVR

Users will inhabit avatars during their time in SocialVR and each platform differs in its customizability options for avatars (see Figures 1 and 2). There are SocialVR platforms offering only generic avatars, which may be customized with regard only to clothes, gender, skin color, hair color, and eye color (*Rec Room*), while other platforms have a collection of avatars to choose from, which may be humanoid or robotic (*AltspaceVR*). A few SocialVR platforms offer greater degrees of customizability or even the ability for users to design and upload avatars to their SocialVR account directly (*VR Chat*). To be permitted, avatars may need to conform to the platform's aesthetic, but other platforms will put no restrictions on their avatars' appearance beyond the prevention of nudity. This means that the avatars of some platforms will include human-like physical curvature, which may even be highly sexualized or fetishized, and the avatar forms of other platforms will consist mainly of straight lines. There have been a few attempts at avatars that are photo-realistically representative of the user inhabiting them (*EmbodyMe*), but otherwise there would be little reason to assume a user looks at all like their avatar. This means that a user and their avatar might not share the same physical properties, including gender and skin color. Changing the appearance of one's avatar is done either by accessing a menu or traveling to a specific location, like a dressing room, in the SocialVR platform. This means that the avatar of an individual user may change multiple times within a session, one moment inhabiting an icon of popular culture and the next moment an anime-style character in a school uniform. Despite these changes, the users will still be identifiable by a username, even if it is a generic name of a guest

account, and that name will either be continually visible or accessed by clicking on their avatars.

Figure 1 Avatar customization in Rec Room

Figure 2 Avatar selection screen in AltspaceVR

3.3.5 Environment in SocialVR

The environments that users inhabit may also change throughout a session. Some SocialVR platforms allow for party membership, meaning when one party member leaves one environment for another, the other party members are invited to join (*Rec Room*). Users may be motivated to move to different environments for the sake of exploration or to pursue a desired activity. The environments may be developed either by the SocialVR platform or the users themselves. They may be constructed within the platform or imported from a game engine. The environments may also have interactive components, such as: a bed to lie down in, a torch to give light, or a writing instrument for use on a flat surface. Some environments also allow for the inclusion of external media, such as photos, web content, and streaming videos, which may be exhibited by the SocialVR platform or the user directly. This shared content is the main feature of some environments and, arguably, it is the main part of some SocialVR platforms (Basement VR).

3.4 Summary to virtual realities

As we have seen, computing technologies have advanced considerably over the past decades in their level of sophistication and their adoption. We will now see what happens when the trends in technological advancement and expansion meet with longstanding faults found in human behavior.

CHAPTER 4

HARASSMENT IN SOCIALVR:

ITS HISTORY AND CURRENT METHODS OF COUNTERACTION

This will be a discussion of challenges related to harassment given the increased availability of SocialVR and the greater number of modalities that it uses. SocialVR platforms have introduced measures for coping with harassment cases and their proven insufficiency in the face of trolling behavior.

4.1 Harassment in MUDs

MUDs represent an early form of computer-mediated virtual reality because they enabled symmetric dialogue between users who could interact within environments. To create elements within an environment or make objects, users needed only to write about them within the dungeon, so others may acknowledge their existence and interact with them. As seen below, this can be a powerful tool in the hands of a harassing user even though the abuse happens through text alone and people do not share a literal space. As it is in-person, discriminatory or threatening behavior is emotionally detrimental to the victims and it is harmful to the virtual world itself as wronged users abandon the platforms or assume male identities to avoid being harassed (Fox, 2016).

Even if dialogue within a dungeon is directed towards specific users, the chat may be visible to everyone present in the space and, therefore, each of them may be subjected to harassment through humiliation or expressions of hatred. Take the

following case, which is a real example of how two MUD users (ViCe and Aatank) performed sexual harassment against other users (sm, st, and rani) with only textual cues (Herring, 1999):

> <ViCe>Aatank man i got women here u'll fall in love with!!
>
> <Aatank>vice like who
>
> <ViCe>Aatank a quick babe inventory for u: st / sm and rani :)
>
> <Aatank>sm hi u can call me studboy. what color are your undies
>
> <ViCe>haha ¤ ¤ ¤ Action: Aatank rushes up to st and yanks her panties off. BOO!

Also in 1993, a case of simulated rape was reported in a MUD named LamdaMOO, where an abusive user assumed the identity of two female users, thereby forcing demeaning acts upon them which prominently featured sodomy and pubic hair (Huff, 2003). The incident initiated a broader conversation on virtual identities, censorship, and means of preventing harassment in virtual spaces. These cases do not represent forms of institutional harassment since any registered user has access to the MUD. Rather, this type of harassment is nearer to street harassment because it may happen between strangers and its resulting humiliation impedes the victims' ability to move and act freely within the environment. One can easily imagine a case of hate speech also taking place in the same venue. Harassed users may feel frustrated and anxious due to their inability to participate meaningfully in the MUD. They may feel greatly embarrassed by the experience of being publicly targeted for the violation of social norms and the text-based sexual aggression could easily be triggering for users. This disruption to their

experience could cause them to leave the platform, or perform *gender masking* in which users choose male or gender-neutral names to avoid unwanted, sexual attention (Fox, 2016)

Some MUDs had regulations that users were expected to adhere to while interacting in dungeons, but others lacked such precautions, especially in the pre-commercial days of the Internet. Creators of the MUDs seemed not to know what behavior to expect from their visitors. Users and administrators of MUDs may have had the ability to kick out offending players, but they had little, if any, power to prevent them from returning to the dungeon soon after. It is within this set of circumstances that a new type of harassment was conceived of and it has found renewed relevance in modern iterations of SocialVR.

4.2 Harassment in SocialVR and counteractive measures

At a minimum, harassment in SocialVR may be spoken to a user or acted out physically by standing too closely, participating in unwelcome touching, or acting out sexually suggestive pantomime. In this medium, gender masking generally ceases to be an option since users may suppose one another's gender by the sound of their voice. SocialVR has already experienced a high-profile case of sex-based harassment when a woman was groped while playing the game QuiVR (Belamire, 2016). As of writing, how the body is captured in virtual reality is limited and, therefore, the sophistication of a user's gestures is low, but this aspect of virtual reality is seeing continual advancement (Han, 2017). Some solutions to harassment have been implemented in response to verbal and physical

harassment, and more may be done still, but SocialVR platforms are designed as a source of entertainment and a means of connecting to others. Building an environment in which users are continuously on guard, carrying a full arsenal of defensive measures, could easily undermine the purpose of the experience (Shriram, 2017). Therefore, the ideal measures for stopping harassment should be automated or easily accessible to the user, but they would not be obtrusive to the positive experiences of SocialVR. To elaborate further, there may be language and behavior that is appropriate between friends, or even users with mutual romantic interest, that would not be appropriate among other users, so we would ideally not want to stymie a good experience in SocialVR for the sake of defending against something negative. Furthermore, if users are given a reactive feature, making the victim responsible for initialization, then the means of accessing the anti-harassment tool should be clear. It should not require too many steps and it should definitively end the harassment from the offending user for at least the length of the session. Finally, the solution to harassment should not be subject to abuse, giving the offenders another way to disrupt the experiences of others as can happen with online tools (Ehrenkranz, 2017). The solutions discussed in the following sections are user-initiated features as automated tools seem to not be active.

4.2.1 Muting

The mute feature allows one user to silence another within the virtual space. The effect may or may not be reversed by the user who is muted. Additionally, the muted person may be silent only to the muting party or to everyone using the SocialVR application. Some platforms allow a user to mute any other user, or themselves, by simply clicking

on a menu button beside the user's nametag and the reason may be for the sake of harassment (*AltspaceVR*), but it's more often to silence someone experiencing disruptive feedback from their microphone or someone speaking too loudly at a public event and preventing others from hearing the main presenter. In harassment cases, it could be effective in signaling one's annoyance to the offending user, but the muting is quickly reversed, doing nothing to prevent the harassment and perhaps goading the offender into further attacks. At the same time, not allowing the muted person to unmute themselves would unfairly penalize the innocent. It would also fully remove real offenders from participation, but still enable them to enact forms of physical and non-verbal harassment against other users.

4.2.2 Blocking

Blocking, or Ghost Mode, is another option in some platforms, where the harassed user may click a menu button near their harasser or perform a specific gesture in the harasser's direction (*Rec Room*). The only added action may be a request for confirmation. When given, the offending user will neither be visible nor audible to the blocker and vice versa. The two users will be unaware of the other's presence within that space even though everyone else will be. This effectively ends instances of non-persistent harassment or annoyance by other users, but it does not prevent blocked user from returning under an alternate account, perhaps using another email address, to do further harm. Also, if the user proves to be a general nuisance, they must still be blocked by every individual user, which requires a lot of menu access cumulatively.

4.2.3 Kicking

Kicking is a feature of a SocialVR platform in which a group of users, or the host of a private room, may vote to remove a player from a room. When one player initiates the "kicking" of another, bystanders receive a notification in which they may also vote to remove that player. To illustrate, Rec Room is a SocialVR platform consisting of a common area that connects to multiple games and each game area may contain multiple rooms, the number depending on the games' capacity. Once kicked, the player is ejected from the game and there is a short delay before they may re-enter, but the kicked player will be unable to join the same specific room again. This puts the decision to censor behavior into the hands of groups who may be frustrated by harassing behavior or the kicked players attempts to sabotage a game. However, the same tool can be abused by players who might wrongfully remove good players from a game in order to increase their chances of victory. In a justified incident of kicking, harassing players still have access to all other areas and it will not protect the victims of harassment if they leave the game where the harassment occurred.

4.2.4 Bubbles

Rather than filter out other users through muting or blocking, some platforms have a version of the protective bubble feature through which, from the users' perspective, the physical form of another user's avatar may not pass (*High Fidelity*, *Rec Room*, *QuiVR*) (see Figure 3).

Figure 3 AltspaceVR buttons (Left) for self, bubble on bottom (Right) for other users, block in the center, mute on the right

If another player nears or penetrates the invisible barrier, the body of the intruder will initially fade and then disappear. This is a solution working against sexually themed attack or intentionally intrusive users, but not verbal harassment or offensive mimicry from a distance. The bubble may be turned on at all times, but the user may choose to disable it should they want to come closer to other avatars. The radius of the bubble may be customized in some platforms, giving each user as much personal space as they require. Here, the user experience may suffer since they must turn off the bubble when making voluntary physical contact, but it might be a worthwhile tradeoff where the other users are unknown, or harassment seems likely. Bubbles may be called upon by accessing the menu, but there are some platforms that utilize gestures, such as raising one's arms outwards, to access the protective bubble (D'Anastasio, 2016). Since it is a simple gesture, this solution presents a fast way to escape an uncomfortable or threatening situation, but the harasser is still present within the space at a short distance.

4.2.5 Reporting

SocialVR platforms generally have a method of reporting, or flagging, harassers, which can either be done from menus within VR or through a standard form available on the website. The terms of service for these platforms vary in the amount of detail in their descriptions of harassment and the penalties enacted for each type of harassment may not be stated explicitly. Potential outcomes for harassment claims are sometimes given and they may include the suspension of an account, the closure of an account and further blocking of a user's Steam account, or complete blockage of access by someone using a specific IP address. The enforcement and penalties for harassment are at the discretion of the administrators of the SocialVR platform, whose interests in preventing access to their platform may conflict. First of all, barring someone from accessing the platform directly lowers the number of users on the platform and removing an individual user could potentially lead to the loss of the social network connected to that user. Therefore, the risk of losing the harassing users might be weighed against the likelihood of retaining the harassed users and the potential for further harassment from the offending users in the future. In addition, heavy handed enforcement of harassment policies carries the potential for a backfire effect in which a network of users engages in systematic trolling behavior for the explicit sake of disrupting the SocialVR platform entirely (Binns, 2012).

4.2.6 Admins

SocialVR platforms with a large enough usership are known to keep admins stationed in continually populated common areas. The admins are humans employed by the platform to monitor users' behavior by remaining in the environment with them, engaging in conversation with them and warning them away from excessively harassing behavior. This solution is the surest method of classifying harassment, but it is likely to be untenable when SocialVR usership grows and it will be a superfluous position when unsupervised and automated methods of harassment detection become available.

4.3 Trolling behavior

Since SocialVR integrates aspects of online and in-person communication, an understanding of behaviors relevant to both arenas will give a broader picture of the players involved in an instance of harassment. Online trolling is defined as malignant actions intending to compromise a social environment and, as studies have shown, this behavior is often correlated with the sadistic tendencies of trolls generally (Buckels, 2014). Since they share the same potential for harm and havoc alongside the cloak of virtual anonymity, one can assume that the troll's motive of deriving pleasure from another's pain crosses over from the old domain of online social networks to the new domain of SocialVR. It is also consistent with trolling behavior to abuse or skirt systems of preventing their harassment, so their unwelcome behavior may continue unabated. Trolls have also been known to coordinate their attacks against entire platforms if, for

example, they disagree with the introduction of a new policy, and this may be disruptive to every other user in a highly publicized manner (Higgin, 2013) (see Figure 4).

Figure 4 Dark-skinned avatars in swastika formation in Habbo Hotel Raid (The Awesome Patman, 2013)

Concurrently, SocialVR may simulate the experience of being physically present with a person insofar as users may see one another by proxy and speak to each other in real time. For this sense of presence, SocialVR lends itself to sexual advances by users who may feel a heightened sense of gratification from their behavior. That is not to say every case of sexually themed conversation or physical movement is unwanted or improper as some users log in specifically to meet with a romantic partner or flirt congenially with other users. However, sexually motivated users with harmful intentions may easily

address other users or initiate physical contact inappropriately while there is a lessened chance of repercussions for this behavior. It can be argued that the unique sense of presence and altered mobility that comes with modern SocialVR, to a degree, increase both the likelihood of harassment taking place and the stakes for the SocialVR platforms to keep that harassment from happening.

4.4 Proposed methods for detecting harassment in SocialVR

The current methods listed above for preventing or responding to harassment all share the property of being user initiated. Making users responsible for responding to harassment against them requires educating them on the available anti-harassment tools and encouraging them to use it. However, if tutorials on preventing harassment become mandatory for registration on the platform, this has the potential side effect of discouraging new users from a lengthy registration process while making them wary of the SocialVR experience since they may now expect to be harassed. There may be an additional reluctance to use these tools on the users' part because they foresee it as a source of conflict if the harasser learns of their action. Finally, the harassed users, on principle, may not wish to disrupt the harassing user's experience.

One answer to identifying and preventing harassment may come from an automated response initiated by tools in the SocialVR platform rather than the user. The tools needed to detect harassment would begin with user profiling that considers data on the users, such as length of membership and history of abuse, and users would receive a score based on this profile. Features of the virtual environment may also count towards

the score, such as the time of day and current number of users. Transcriptions from the users' sessions may be taken and NLP may be used to find patterns in harassing speech, including the repetition and syntactic context of taboo words and hate speech. Analysis of the transcribed discourse would also result in a score to be added to the total and, if the score rises above a preestablished threshold, an action is triggered to deal with the potential harassment.

Since two-dimensional and three-dimensional image creation is also available to users of most SocialVR platforms, there should also be automated tools for preventing its abuse. The image could either be sexually explicit, related to hate symbols, or written words. There should be a method of capturing user-generated images and classifying them as harassing. Resulting actions taken by the platform should alleviate the potential hostility sparked by the drawing and prevent it from reoccurring in the future. These proposed methods for detecting harassment will be discussed throughout this thesis, including their implementation in software.

CHAPTER 5

METHODS OF QUALITATIVE ANALYSIS OF USER DISCOURSE IN THE DEVELOPMENT OF NLP TOOLS

Building a dictionary of lexical items and syntactic structures for use in the analysis of text can be partially done through literature review, but most come from online resources, such as social networking platforms. Data that specifically regards harassing features of speech while embodied is much less common, which resulted in the need to collect original data for the NLP implementation. Information on methodologies for autoethnographic studies in SocialVR was also uncertain, but here they attempt to follow the tenets established by discourse analysis.

5.1 Qualitative analysis of discourse in SocialVR

Since the virtual environment and its interactive features play an ancillary role in most SocialVR discourse, the methods of Multimodal Critical Discourse Analysis (MCDA), which would allow me to integrate discourse and the visual environment in the interest of discovering the factors contributing to harassment (O'Halloran, 2004). As revealed in the data, MCDA became integral to contextualizing much of the analysis since items such as avatar appearance, phallic-looking props, three-dimensional drawing tools, and the VR equipment itself needed consideration to allow for the recognition of harassment. This approach also considers language and other sensory data as a symbolic system which may be used to unlock an underlying intent or thought process (Wodak, 2011).

Multiple environments and platforms were used in the data collection. The discourse was transcribed from video taken from my first-person perspective when visiting SocialVR. I did not initiate conversation while recording and did not intervene when witnessing a likely instance of harassment. Due to technological aspects of VR, voices may become unclear, there may be voices of people outside VR watching another person in VR, people may speak in different languages, and there may be unexpected sounds. People appear and disappear suddenly as they log off or move to a different area. Likewise, people join and leave conversations suddenly, which indicates that the social norms of in-person discourse may not apply.

5.2 Summary of harassment activity from qualitative analysis

In the same manner a medical practitioner would diagnose a disease, a person creating these preventative tools first needs to recognize what defines a case of harassment. Collecting first-hand data on harassment required spending indeterminate lengths of time in virtual environments, essentially waiting for a situation to arise. During this data collection period, which was to be two or three weekly visits over a period of approximately three months, the majority of SocialVR sessions and the majority of interactions between other users did not produce instances of harassment. However, the minority cases where harassment did occur were severe and frequent enough to produce patterns and inspire action to halt them. Actionable data from this analysis was used in the NLP program. Categories of harassment from the analysis are shown in the table below and summaries of harassment incidences are found in Appendix A with aliases used for each of the usernames.

Table 2 Categories for Instances of Harassment in Qualitative Analysis

	Verbal					Images	Physical / Gestural		
	Group-Based			Sexual	Profanity		Sexual	Confrontational	Invasive
	Race	Gender	Other						
Session 1	x	x	x	x	x		x	x	x
Session 2	x	x	x	x	x		x		
Session 3				x					
Session 4	x						x		
Session 5					x				
Session 6					x				
Session 7			x			x			
Session 8		x		x			x		
Session 9				x		x			
Session 10		x		x			x		x
Session 11		x		x					x

5.3 Results from the qualitative analysis

Given these cases from the qualitative analysis and their identifying features, it becomes possible to identify the types of harassment one would like to defend against. Listed below are harassment-related patterns and identifying features that might be used to help classify them as harassing behavior:

- Sounds Related to Sexual Activity: The sound of orgasm, whether fake, pre-recorded, or genuine, is to be considered harassment in a public setting. Analysis of audio signals can be used to detect this class of sound and, when performed in a public space among non-friends, the behavior of the offending user will be classified as harassment. There are also lexical patterns found in the sound of orgasm the may be detected by NLP tools.

- Presence of Sexual or Hate-Related Imagery: The ability to draw images within a three-dimensional space or on a flat surface, such as a whiteboard, is a feature of many popular SocialVR platforms. This also allows users to produce images which can either be vulgar or associated with hate speech. Upon their creation, this type of imagery can be captured and analyzed, classifying the person who made it as a harasser.
- Incitation to Violence: Users using violent language in conjunction with a specific race, gender, or sexuality are performing hate speech. Therefore, they would be flagged as engaging in harassing behavior. In the transcript, there were examples which would have been classified as harassment, such as: 'Kill Black People', 'Rape Women', 'Kill the Midget', 'Punch this Bitch'.
- Large Quantity and Repetition of Vulgar Language: Users directing vulgar language towards another user may be considered harassing or playful. The repeated use of vulgarity among strangers often creates hostility, making it a form of harassment. The user response may be determinant in these cases. For example, not responding with the same vulgar language, requests to stop, victims leaving the area, or victims not giving any response could all be signs of harassment taking place.
- Large Variety of Taboo or Controversial Topics: Users engaged in trolling behavior will introduce multiple topics to provoke an angry reaction. If enough of them are used in close proximity and in combination with vulgar language, it could be an indicator of harassing behavior. In Session One, a user gives topics of discussion, including: Donald Trump, abortion, spousal abuse, immigration,

multiple types of sex, and racial minorities. He does this all within the span of 12 minutes. Speaking about most of these topics may not be grounds for a harassment classification, but speaking about all of them in conjunction with abusive or vulgar language may be. For example, ‘You’re an immigrant’ may not be a harassing statement, but ‘You’re an immigrant and you don’t pay taxes. Fuck you.’, as found in the transcript, is assuredly a case of harassment.

- Proximity of Vulgar Language and Behavior to Login Time: A user’s trolling behavior often began immediately after logging into the SocialVR platform. In the early parts of a visit, most users have not had time to discover whether or not vulgar speech is appropriate or wanted by other users. This may be a sign of intentional rudeness from the suspected troll, so the consideration of behavior in the first minutes of a visit may help classify harassing behavior later in the visit.
- Physically Mimicking Sexual Activity: Users may simulate sexual contact with other users either to disrupt their VR experience or for the harassers’ own sexual gratification. Currently, harassers may perform this mimicry with a wide variety of available hand gestures and objects. They may also simulate sexual acts through repetitive back and forth motions in which they repeatedly come into contact with the harassed user’s avatar. Positional, gestural, and movement logs could track user movements to determine if harassment is taking place, but relative physical coordination and skill with the users’ VR controllers could lead to unreliable data or false positives. In addition, the types of logs needed for such an analysis are not yet publicly known to exist on any large scale.

- Suprasegmental Features of Speech: Harassing speech may include elements of speech beyond the individually spoken words, such as: raised volume, changes in pitch, changes in tempo, and falling or rising tone, among others. Understanding which patterns indicate disapproval on the side of the harassed or an attempt to harm on the side of the harasser would require a spectrogram analysis of data. This approach, however, is met with significant complications since research into human behavior has shown us that expression of emotion is heavily variable depending on people, cultures, and elements within the situation (Barret, 2006). Gathering and applying suprasegmental data would require many data samples, which are currently insufficient for this research.

With the exception of the last two items, which will both be discussed in the Future Research section below, the current project seeks to utilize these patterns of harassment to classify harassers for their behavior so their influence within SocialVR platforms might be reduced. All of the considered approaches may be used in the context of Harassment Scores and Harassment Thresholds, where users and their behaviors add to a score which will lead to a harassment classification if the threshold is exceeded. This scoring will be discussed in further detail in the following sections.

CHAPTER 6

COMPONENTS OF HARASSMENT SCORING

The methods currently employed in harassment detection rely on the actions of harassed users or witnesses to the incident. Until speech processors can detect when one user is harassing another with virtual certainty, multiple probabilistic approaches may be used in the detection, triggering actions that will help potentially harassed users, or prevent it from happening in the future. The components of harassment scores are described below.

6.1 User profile scoring

As a SocialVR platform ages, they have the opportunity to learn more about their users and behavioral trends within their demographic. This allows the platform to target segments of the population that may be most interested in their service and find ways to hold the interest of the already existent user base. The same principle may be applied to incidences of harassment where demographic data may be recorded alongside user histories to determine the nature of the behavior they are likely to participate in. Demographic data may include:

- Age: Platforms may choose to give an initial score based on the users' age, which can be applied either on a trajectory or within a range. Applying a score

on a trajectory would mean that users would either get a higher or lower score depending on how young or old they are. A ranged system would mean users are given an initial score based on a grouping of ages. For example, a 13-year-old might get a higher score than a 14-year-old on a trajectory, but both of them could be given the same score if they are scored equally within the range of 13 to 17 years old. SocialVR platforms generally have an age restriction for the users' own protection, so many younger users would have an incentive to lie about their age when registering, subsequently reducing the platforms ability to profile all users. However, this can be partially counteracted by disregarding users whose birthday is on January 1st since that is commonly the default date when entering one's birthdate.

- Gender: Where gender data is available, and their relative likelihood of harassing other users is known, platforms may also choose to add a pre-set gender score to the harassment score. However, this practice seems to be falling out of favor as SocialVR platforms and online entities become more sensitive to non-binary gender classifications (Cole, 2000).
- Geographic Location: At any time, the geographical location of a user may be recorded by the SocialVR platform. It is indeed a feature of some platforms to include a map within a virtual space that reveals the real-world, geographic location of every user within the room. If a platform finds there are users from a specific region who have been disproportionately flagged as harassers, the platform may use it as criteria for adding to the harassment scores of all users

from that region. The reason for such an occurrence may include coordinated trolling attacks from people known to each other living within a region.

- Means of Access: Users can access most SocialVR platform through the leading HMDs on the market, but some platforms are also accessible through mobile VR or even a flat-screen computer, not requiring a VR device. Likewise, the SocialVR platform may be accessed through different online stores or bypass the digital marketplace by offering a downloadable app. These means of access vary in price, functionality, and quality of experience, which allows a platform to make inferences about the users' economic standing, how they use VR, and their sophistication with the technology. SocialVR platforms may add a score according to this category if the data support it.

The relationship between the platform and the user is defined as user history. This may include:

- Registration Status: SocialVR platforms do not uniformly require registration, allowing users to enter the virtual environment as a guest. Non-registered users may be assigned a standard avatar and given a completely numerical ID, or they may have limited ability to advance within the platform and the title guest will be appended to their username. These steps are taken by the platform to encourage registration while also allowing newcomers to preview the virtual experience. However, these users have even greater anonymity and less persistent identity

than the registered users, meaning the potential penalties for harassing behavior from the platform and the social costs from non-harassing users are lessened (Suler, 2004). SocialVR platforms may also want to hold new users to a higher behavioral standard since they have yet to understand the social norms of the platform and it may model user behavior in future visits to the virtual environment.

- Duration of Registration/Login Time: It is assumed that the length of a user's registration, or their time spent in the app, correlates positively to their knowledge of social norms within the platform. This could mean that newer users pose a greater harassment risk due to either ignorance or low social cost. There are even platforms that apply a literal level to their users based on the amount of gaming activities performed within that platform and it can be imagined that users have a more personal stake in their good standing with the platform as their levels accrue. On the other hand, there is also a potential for long-term predatory behavior from a user, so the relevance of the registration length should be supported by the data before adding it to the user score.
- Past Harassment: Previous incidences of harassment are a strong predictor of their future behavior. Assuming an offending user's account is not banned or suspended from the SocialVR platform, the user may be given a harassment score to catch them more quickly if their harassing behavior is repeated in a future session.
- Number of Friends: The number of friends a user has might be an indicator of how they use the platform. It is expected that users who behave well,

successfully socializing with others, will become friends with them. However, the higher or lower relevance placed on these social connections depends on the individual users. For some, the decision to add another user as a friend is haphazard as they may send requests to whomever is in the room with them at that time. The relevance of the number of friends with regard to harassment should also be determined when the data is available.

- Friends with a Harassment History: In social network analysis, the behavior of connected persons is often a reliable predictor of their own behavior (Mouttapa, 2004). This principle might also apply to friendship networks within SocialVR platforms. A higher number, or percentage, of friends with a record of harassment could indicate that the user's personal standards of behavior does not comply with the platform's.

SocialVR platforms who implement profile scores based the criteria listed above would need to continually update their scores as the user gets older, gains friends, and spends more time on the platform. Also, the reasoning behind scores might not be borne out by the data over time and they may decide either to reevaluate scores attached to the criteria periodically or automate the rescoring based on any harassment events. However, platforms should be wary of automatic scoring because systems like these could be self-perpetuating much in the way that encoding biases effect human judgment (Lewicki, 1989), considering a raised profiling score reliably predicts the classification of future behavior as harassment. Creating score limits for the specific criterion or user profiles generally will help mitigate this risk of false positives.

User profiling could be generally controversial, especially demographic data, and its mismanagement could be harmful to the reputation of a SocialVR platform despite its good intentions. SocialVR platforms may wish to be seen as socially progressive and they would want to weigh the supposed benefit of user profiling against the likely reaction of users who discover they are being profiled. On the other hand, giving profile scores based on user history can be helpful for weeding out the perpetrators of unwanted behavior. For these reasons, I recommend:

- Do not use age-related data, which cannot be easily verified in any case.
- Scoring based on gender is also discouraged since people could easily accuse the platform of gender discrimination in addition to being inconsiderate towards users who have a non-binary gender identity.
- Giving scores based on geographical data should be done with extreme caution and not inadvertently target cultural or racial groups. Scores given to geographic regions might be given to entire cities or metropolitan areas rather than neighborhoods, which might contain a disproportionate number of people belonging to a group.
- Profiling users based on the HMD manufacturer or online store they use to access the SocialVR platform carries a slight risk of damage to the business relationship between them. Though retaliation would be unlikely since the mere fact of having a greater number of harassers using their product or service as opposed to their competitor would harm their reputation and they may prefer to overlook the issue.

- All profile scores based on user histories may be considered objective in that they are based entirely on user behavior and not individual assumptions. Therefore, they can be used with less fear of outside criticism. These criteria will also change over time, giving the users more agency over their scoring as opposed to scores given for demographic data.

6.2 Environment scoring

While user profile scores are applied whether or not a user logs into the platform, environment scoring is applied from the moment they enter the VR platform and it changes depending on what is happening there. Shifts to the environment score may be based on individual users, groups of users, or when people arrive in that virtual space. Environment scoring may include:

- Number of Users: SocialVR platforms may want to use this score to reduce the number of people who witness a harassment event by giving a harassment score proportional to the number of people sharing the virtual space. The platform may decide that having few users together in an environment puts unwitting users at greater risk of being placed with perpetrators of harassment or trolling behavior and the environment score should be raised. Likewise, they could find that trolls are more prolific in their attacks among large group of people because there is an increased number of targets and the potential to upset a larger volume of users.

- Number of Harassing Users: The qualitative analysis revealed that harassers will support one another in their harassing behavior through declarations of approval, simultaneous laughter, and reiterating one another's threats against non-harassing users. There have also been instances of potentially harassing users seeming to test a vulgar or controversial line of discussion, but dropping it when no one else reacts or joins in. Given this evidence, platforms may decide to increase the harassing scores further when found in the presence of other harassers.
- Number of New Users: When users first experience a SocialVR platform, they are building an impression which will be based largely on their first encounters with other users. By definition, a harassed user does not wish to be harassed, and users who are harassed on their early visits to a platform are essentially getting an experience they do not wish to have and may choose not to continue visiting the platform. Furthermore, word-of-mouth about their experience will sometimes spread, risking a reduced number of new users. An increased environment score may be applied when there is a higher number of new users to help prevent the negative experience, to reinforce cultural norms among new users, and increase the likelihood for positive word-of-mouth.
- Current Time: Depending on the time in a user's geographic location, such as late evening on the weekend, a platform could find that users engage in behavior that is considered vulgar, which might result in a harassment classification. If this is problematic for the platform and they wish to curb the behavior, they could raise the environment score for these particular hours. The platform may also decide to do the opposite by raising the threshold if they find that people

visiting the platform at that hour are generally likeminded and not offended by the manner of conversation.

- Session Duration: The qualitative analysis has shown, and further data may reveal, that users visiting a platform with the intention of committing harassment may begin doing it soon after logging in. Users who introduce vulgar or controversial topics of conversation with new users immediately upon signing in might more likely be creating a hostile environment among the users around them. Therefore, setting a higher score for users in the first minutes of their visit might be deemed appropriate.
- Avatar Proximity: It has been found in the qualitative study that verbal harassment often occurs at close proximity and physical harassment would, by definition, happen near, or inside, the victim. The protective bubble was implemented for this reason, specifically as a response to the incident in QuiVR (Wong, 2016), but the nearness of a harasser could still be used to increase the environment score. This may be combined with the bubble feature, meaning the score would increase when someone is close enough to trigger the bubble.

If implemented, Environment Scores would be added to User Profiling Scores to increase the accuracy of harassment classification, but there should be an upper limit to their combined score so that users will not automatically be classified as harassers and then blocked or kicked out of the platform. SocialVR platforms can use Environment Scoring to curate the type of experience they wish users to have without users being explicitly aware of it. For this reason, I recommend gathering data on each of the listed

points and using them within all public areas of the platforms. If there are private areas or events in the VR platform, different classification thresholds may apply, or harassment classifications could be disabled altogether since attendees are more likely to be friends of the host and one another.

6.3 Lexically based harassment classification

The primary mode of interaction between users in SocialVR is verbal conversation, so it is anticipated that most of the harassment in VR will also be verbal or a combination of the verbal and physical. This makes the detection of harassment in SocialVR a different task than classifying harassment in online social networks where, until a message is deleted, the text is accessible to both the sender and the recipient. Ambiguity as to the intent of the harasser is less common in these social networks and evidence for the harassing statement is found in the message itself. Social media platforms have the ability to use keyword searches and sentiment analysis of online exchanges to target those who may potentially be a nuisance or even threatening to other users (Yin, 2009).

A complete lexical analysis of verbal interactions in SocialVR is different because it requires an audio log of every user's visit to the social platform which, while far from impossible, may be resource intensive. There is a computational cost, storage costs, and cost of labor in managing the information. In place of audio files, platforms may use speech-to-text programs, allowing them to keep a file on each user in their records. This way, the platform can confirm or disconfirm harassment in the event that another user reports it. The end of the users' statement may be timestamped, so the truth

and timing of a claim is known. Moreover, an immediate or periodic processing of their natural language can be done to provide a classification which the platform can act upon while it is still taking place.

6.3.1 Method

The program written for this project was done in Python and the Google API (https://pypi.python.org/pypi/SpeechRecognition/) was used for the speech-to-text processing; online natural language processors like IBM Watson (https://www.ibm.com/watson/) are also available online while others like PocketSphinx (https://github.com/cmusphinx/pocketsphinx) are available offline. Users should be aware that their quality varies, and the speech processors may have difficulties understanding some users, so they should be tested for accuracy or common transcription errors should be considered in the analysis. The Google API in particular censors its results by using the initial letter of a curse word and replacing the remaining letters with asterisks, but their meaning can be assumed, or the censorship tool can be circumvented in the code, returning the originally spoken curse word unchanged.

This section includes discussions of the vocabulary and sentence constructions that are targeted in the code (see Appendix B), how the code may be implemented in a SocialVR platform, how the program may be tested and improved, the challenges present in lexically based classification, and how the program may be used in the future. Each of the sections will add to the total score (macro-score), but some of them will also include a score for that section or shared between two or three sections (micro-scores).

The macro-score is meant to include every known type of harassment, behavior and non-behavior-based, each of which will add to their score. Harassing users may only be performing one type of harassment repeatedly and adding to the same macro-score continually may lead to a slow harassment classification or a false negative. The micro-score considers the number of times a form of harassment has been committed and either adds significantly to the total (macro-) score or leads to a harassment classification directly. The targeted lexical data, sentence structure, and harassing patterns came from the qualitative analysis, synonym searches, and literature review on harassment classification (Gitari, 2015; Silva, 2016; Davidson, 2017; Geen, 1975). The given scores are meant to represent the priorities of the website with regard to the behavior they would wish to detect and behavioral patterns that may emerge in types of harassment, but they have not been fully optimized since they are awaiting more data to be validated. In essence, they are representative placeholders free to be adjusted by the platforms that implement them.

- Singular Lexical Items: The first level of analysis is the quality of individual words in the discourse and the four categories that are considered include: swear words, controversial topics, abusive terms, and taboo words. The discussion of controversial subjects and swearing are not explicitly discouraged, but their overuse may be a sign of harassing behavior, especially in combination with other categories. Abusive terms are always problematic when used sincerely against another user and taboo words are always considered unacceptable in a public setting, like a common area in SocialVR. The use of any categories add to

the total harassment score and the use of three or more in combination can lead to an automatic harassment classification (see figure 5). The only reason taboo words do not result in an immediate harassment classification is the chance for incorrect transcription by the speech recognizer. Specifically, some pronunciations of the word *can't* can be misunderstood as the taboo word *c****.

- Harassing Imperatives and Abusive N-Grams: There are common bi-grams and tri-grams that may be used to insult someone or demand a sexual act be performed on them. Some of these n-grams contain swear words, but other times the components of an n-gram, such as *blow* or *jerk*, may be completely inoffensive when used individually. When a higher number of lexical items is used, the evidence for a propensity towards harassment becomes stronger, which justifies higher scores being added to the total score. There are also high-frequency, abusive n-grams which, as dictated by the micro-score, may lead to a positive harassment classification if repeated too many times in too short a time span.
- Name Calling: In this form of harassment, the user calls another user by an abusive name. Optionally, it may also include an abusive adjective and an intensifier. Because there is no verb, these expressions do not constitute complete sentences. Their sincere use is intended only to belittle the target of the abuse. Each lexical item in this section would lead to a higher total harassment score since the abuse becomes stronger and the intent to harm is made clearer. For example, 'you' + ABUSIVE TERM would receive the lowest available score while 'you' + OFFENSIVE ADJECTIVE + INTENSIFIER + ABUSIVE TERM

would receive the highest available score. A substantially high score in this section alone may lead to a positive classification for harassment.

- Name Calling (Complete Sentences): This section is like the previous one, the only difference being the presence of a verb. Since the harasser's intent is similar, the micro-score for this section may be added to the previous section's and a substantially high score will lead to a harassment classification.
- Explicit Threats of Violence: Despite the inability to carry out a threat of violence in the medium of SocialVR, making threats of violence leads to a hostile environment at minimum. With the possible exception of Facebook Spaces, the identity of users is more difficult to discover in SocialVR than it is on popular social media platforms since only the username is shown amongst other users. Still, users may learn one another's identities if they make the mistake of divulging them or their username appears also on their social media profiles and, in this case, threats must be treated with the utmost seriousness. In this section, there is a stated intention or desire to harm someone, specifically when using the second person pronoun 'you'. Intensifiers also add to the score. A micro-score is included in this section, resulting in a harassment classification if the threats are repeated.
- Hate Speech in Participle Constructions: This is also a violently themed section, but the target shifts from the individual to people grouped by race, gender, nationalities, political affiliations, sexualities, and religions. The harasser does not explicitly say they will perform the violent act, but only that the group, for example, 'Should Be' or 'Must Be' harmed in a particular way. Since this is

considered hate speech and, furthermore, extremely detrimental to the life of a SocialVR platform, repeated infractions of this type would quickly be classified as harassment by using the section's micro-score.

- Unfavorable Descriptions or Comparisons of Groups: People can also express their hatred of a group by unfavorably describing them or directly comparing them to things of lessened repute. This section looks at offending adjectives alongside unfavorable metaphors and similes used to degrade the same groups of people listed in the previous section. Sentence constructions in this section include: GROUP + 'are' (+ 'like') (+ OFFENSIVE ADJECTIVE) + ANIMAL. The score from this section is added to a GROUP AGGRESSION micro-score used also in the following few sections, which can lead to faster harassment classifications.
- Invoking Violence and Hatred against Groups: Harassers who openly express their negativity towards the groups listed above or promote violence against them using a sentence initial 'Let's' will receive a harassment score from this section. The additional GROUP AGGRESSION micro-score is used here since these users have displayed an openly hostile attitude towards a group of people, which is an unacceptable form of expression for SocialVR platforms.
- Promoting Self-Directed Harm: This section is an extension of the previous few sections in its attempt to detect harassment against a group. The lexical cues include a reflexive pronoun preceded proximally by one of the aforementioned groups. The GROUP AGGRESSION score is also included here. This section also serves to detect directives towards self-harm and suicide where the target is

an individual user, which is a grave problem in social media that SocialVR platforms would undoubtedly not wish to see replicated in their domain (Mukhra, 2017). Having a user harm themselves after an unfortunate encounter with another user would be greatly detrimental for both the targeted user and the reputation of the platform.

- Discussing Death of a Group: Speech including mention of these groups proximally to words associated with death could signal that an instance of harassment is occurring. The phrasing could include the expression ‘die’ + GROUP, but longer sentence structures with the same sentiment are also included in this section. This section also includes a GROUP AGGRESSION micro-score.
- Demands or Expressed Desire for Sexual Activity: SocialVR platforms do not, in principle, disapprove of sexual contact between consenting users, but they strongly disapprove of sexually propositioning unwilling users or open expressions of or about sex. There can be legal ramifications if minors are exposed to sexual content and there is a high likelihood of making the general usership uncomfortable or offended. Many SocialVR platforms include private or custom areas where host users may choose who is allowed to visit and may also choose to express themselves sexually with other consenting users. With this in mind, it is important to flag overt or excessive sexual themes in open conversations and divert users to private areas. This section of the program targets potentially sexual verbs used proximally to parts of the body associated with sex acts or the person themselves. The micro-score for this section includes

repeated demands for sexual activity, which results in a positive harassment classification.

- Ejaculation and Prepositional Phrases: There are some sex acts considered especially vulgar and lexically particular, which require added detail to detect in conversation. Synonyms for the verb *ejaculate* are covered in this section of the program and they are used with prepositional phrases. The challenge of this section is highly sexual verb *cum* and its frequently used homonym *come,* so the program seeks to limit the number of false positives by targeting the prepositional phrases including parts of the body that are used with the verb. Using the non-sexual verb *come* with the prepositional phrases is semantically incongruous, allowing the program to give a harassment score with some confidence. Scores may be higher or lower depending on how sexually suggestive the part of the body in the prepositional phrase is. A micro-score is included for this section and exceeding its threshold will result in a positive harassment classification.
- Excessive Repetition of Vulgar Phrases: The qualitative analysis revealed some users who repeat the same vulgar expression, to one or perhaps several users, and sometimes to the exclusion of all other words. The continued repetition of vulgar phrases may be considered an attempt to disrupt the experience of other players and it is considered harassment, whether its directed at individual users or the room generally. A micro-score is included to ensure that users who repeat these phrases excessively are classified as being harassers.

- Protestations: This section is sourced in users who are potentially being harassed and the resulting score from their protestation may be added to either the nearest players or all users presently in the room. Demanding someone stop what they are doing could be a sign that they are being harassed, but it could also be a normal part of their conversation. For this reason, the score limit is kept low, which will prevent a high number of false positives, and also force true positives, whose score is near the threshold, over it. The current program takes a second, subsequent recording for the sake of including protestations, but this would be done differently in a SocialVR platform.

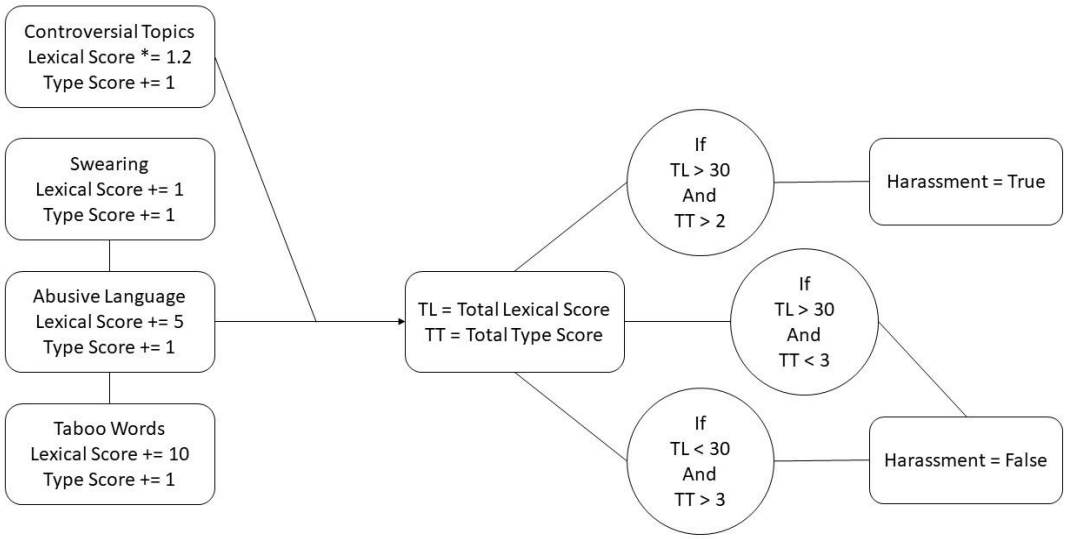

Figure 5 NLP scoring for single lexical item category

6.3.2 Testing

The lexically based harassment program was written with the expectation that lists of applicable lexical items would be added as more data is analyzed. Since user data was still not widely available from the platform, the testing data had to be collected according to methods similar to those of the qualitative analysis. I again entered SocialVR platforms and kept a record of the data that was overheard. The main difference between the data is the absence of context in the testing data as opposed to the qualitative data. Most, but not all, of the expressions taken from the conversation during the collection of testing data were not harassment, but they are still included because they might have been considered harassment if they had appeared in a less familiar and consenting social environment.

Each expression that could be considered harassing in the wrong context was transcribed and given a score by the NLP program written for this project. Additionally, each of the users were given a harassment score for their behavior during a session. Since some scores grow exponentially with each new harassing statement, the individual statements could not just be summed, but had to be read together. Of the 50 harassing statements, only 13 of them were given a harassment score for a total of 27 points. Using the testing data, new lexical items were added to lists in the program, new noun phrases were added, and causative structures were included for sexually themed statements. The phrases and users were tested again, and the adjusted sum was 45 points from the 20 harassing phrases that were given a score, increasing the true positive rate from 26% to 40%. This means that more than half of the tested phrases were left unscored, but they were too ambiguous, context specific, or euphemistic to be captured. Any attempt to

include them might easily lead to false positives in further testing. Table 3 (see below) gives the most relevant examples from the testing data, revealing the original NLP score, resulting additions to the lexicon and sentence structures, test scores for the phrases after changes to the program, and the classification status before and after the changes.

6.4 Summary of harassment scoring

This chapter described three proposed data sources that may be used to determine harassment classifications for users. The first two, user profiling and environment scoring, are probabilistic while the analysis of transcribed speech is more determinative but error prone and subject to mitigating circumstances throughout user sessions. As this solution is implemented to gather more data on users, environments, and discourse, scores will have to be continually modified and data points will be added or subtracted based on relevance. Transcribed discourse will also reveal more candidates for inclusion among lists of lexical items and sentence structures likely to be harassing.

Table 3 Harassing Statements, Scoring, and Classification Results

	Expression	Score	Resulting Action	Adjsuted Score	Classification (Before)	Classification (After)
1	Kiss me	3	None	3	TP	TP
2	Let's have VR sex.	0	Add Construction	2	FN	TP
3	I didn't mean to interupt your circle jerk.	0	Add N-Gram	1	FN	TP
4	I came out to my dad yesterday.	0	None	0	TN	TN
5	Do you have some condoms?	0	Add Term	0	FN	FN
6	Are you wacking off?	0	Add N-Gram	1	FN	TP
7	This is fucking bullshit.	2	None	2	TP	TP
8	It's a fucking pain in the ass.	0	Add Term	2	TP	TP
9	You're a Bronie?	0	Add Term	0	TN	TN
10	Do you know what a Bronie is?	0	Add Term	0	TN	TN
11	I'm a furry.	0	Add Term	0	TN	TN
12	I've got hips coming out of the ass.	1	None	1	TP	TP
13	Who lost their virginity?	0	Add Term	0	FN	FN
14	Could you get your ass out of my face?	1	Add Construction	3	TP	TP
15	Where the fuck did she go?	1	None	1	TP	TP
16	My cat has a shoe fetish.	0	Add Term	0	TN	TN
17	Shut your fucking eyes.	1	None	1	TP	TP
19	Fuck it.	1	None	1	TP	TP
20	Oh, I got a little excited.	0	None	0	FN	FN
21	I'm not old enough to be a cougar.	0	None	0	TN	TN
22	So you lost your virginity.	0	Add Term	0	TN	TN
23	He popped my cherry.	0	Add Term	5	FN	TP
24	I had my cherry popped.	0	Add Construction	3	FN	TP
25	I want to tickle his pickle.	0	None	0	FN	FN
26	She just went inside me.	0	Add Construction	0	TN	TN
28	I'm killing myself.	5	None	5	TP	TP
30	I'll take you in the woods. Deep in the woods.	0	None	0	FN	FN
31	She's got a penis.	5	None	5	TP	TP
33	Mister fister. (2x)	0	Add "FIST"	0	FN	FN
35	I just got raped in the woods.	2	None	2	TP	TP
36	Prositute	2	None	2	TP	TP
37	Prostitution	0	Add Term	2	TP	TP

CHAPTER 7

CLASSIFICATION OF VULGAR IMAGES USING CNNs

While thus far the focus of the project has been on lexical properties of user interactions with brief consideration of users' physical proximity, an added threat of non-verbal harassment comes from users who expose others to drawings of vulgar or hateful imagery. This phenomenon damages the reputation of SocialVR platforms and drives away users believing the platform's users to be overly immature or intolerant. One method of handling this image problem and its implementation is described in this section.

7.1 Introduction to image creation in SocialVR

The qualitative analysis included multiple instance of users drawing male genitalia, female genitalia, and breasts. The most commonly found image was male genitalia and this finding has played out also in my recreational use of the SocialVR platforms. Depending on the features of the platform, the images may be drawn on a flat surface, such as a notepad or whiteboard, or drawn in a three-dimensional space with a drawing instrument or the users' own finger. Some of the SocialVR platforms with this feature or related features at the time of writing include: *AltspaceVR*, *Rec Room*, *Sansar*, *VR Chat*, *Pararea*, *Bigscreen Beta*, *High Fidelity*, *Anyland*, *TheWaveVR*, *OrbusVR*, and *Facebook Spaces*. While the intent of these writing tools is to invite users to express themselves creatively, write messages, or play guessing games (see Figure 6), abusers may use it as

a tool of non-verbal harassment. The SocialVR platforms benefit from the abstract (i.e. non-photo-realistic) rendering of these drawn objects, but this does not guard against the hostility that derives from the drawers' vulgar intent. Some SocialVR platforms allow for photo and video sharing in some environments, which results in the spread of more explicit material, but this is a matter not covered in this research.

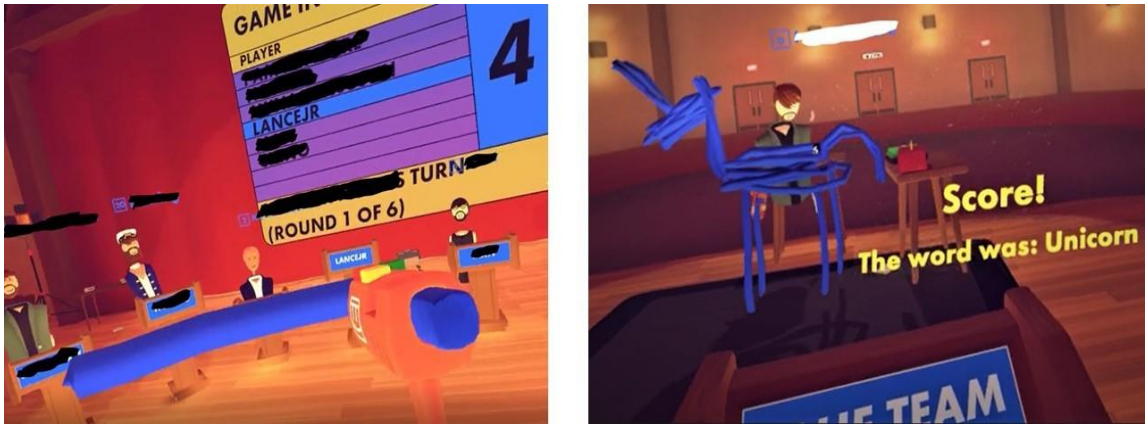

Figure 6 3D drawing game in Rec Room (Left) first-person perspective (Right) third-person perspective

There are a few justifications for this project seeking to classify drawn images of male genitalia in particular. The display of phallic imagery has a symbolic tradition that predates its appearance in shared digital environments and it often denotes aggression; thus, it has become a cultural icon for some groups and informs their perception of local cultural norms (Revi, 2015). Seeing male genitalia drawn into the three-dimensional space suggests a lower standard of behavior which neither the SocialVR platforms nor the non-consenting users have agreed to. Detecting phallic imagery has also been a long-standing problem for virtual worlds and attempts at getting rid of it are known to be costly and never completely successful (Phillips, 2015).

The drawings of male genitalia in the qualitative analysis were also numerous enough to give some patterns as to their form, which would be utilized in producing training data. Universally, the drawings included a long, cylindrical form pointed vertically, meant to represent the shaft of the genitalia, with a rounded edge at the top, and two round forms near the base of the object meant to represent the testicles. In addition to these components, all of these drawings included one or more of the following features: a series of short lines on the testicles meant to represent hair, a horizontal line near the rounded top of the shaft meant to represent the corona, and lines protruding from the top of the drawing meant to represent the trajectory of ejaculate.

7.2 Training data

The features included in drawings of male genitalia may be similar across SocialVR platforms, but the environments in which they are drawn may be very different. There may also be differences in the color or texture of the ink. SocialVR platforms who wish to implement an image classification program to detect harassing images would improve their results by only collecting images that originate from their platform. Otherwise, it may choose to focus on irrelevant details in the environment which will weaken the results and using the program to detect vulgar images in external platforms would be unnecessary to them.

The training data for this project comes from a SocialVR platform, which features a three-dimensional pen most prominently in two of its environments. These environments are where all of the images used in training and testing were collected.

The number of images for the first training/testing set was 700, 350 images for each label. I produced the images for the training data and, in the interest of privacy, they were done alone in private rooms. The final images were two-dimensional images, i.e. screenshots of the completed genitalia drawing, which were sized and cropped identically. Using two-dimensional images of three-dimensional objects has been successfully done in other studies (Burnap, 2015) and it is especially appropriate in this project for reasons covered later. The images were separated into two categories, genitalia and not genitalia, and an equal number of images were included in each category. Because of the environmental variety within the room, one of the environments was subdivided into five sections and an equal number of images were included in each subsection. The number of images in the second environment was double that of a single subsection of the first environment. In machine learning for image processing, it is ideal to have images sourced from multiple people, but this was not an option due to ethical concerns. For this solution, the images also needed to originate from the first-person perspective of the drawer, so awaiting images to be drawn by harassers in a naturally occurring environment would not work either regardless of the impracticality of waiting that period of time. As compensation for the single source, I intentionally varied the arc, the size, and relative dimensions of the images. Each drawing of genitalia included the universal features listed above alongside one or more of the optional features; the optional features varied from image to image.

Both groups, genitalia and non-genitalia, were horizontally mirrored to increase the number of training images from 700 to 1400, they were converted to grayscale from RGB, and they were resized to save space. In a later model, Gaussian blurring was done

at a five-pixel radius and added to the training data, raising the total number of images to 2800, but the resulting model had a considerably weakened performance, so the previous model was used. Some non-genitalia images were drawn to include individual features that are from the set of vulgar images. Among these images, there were rainbows that had a topmost arc similar to the corona of male genitalia, images of two cherries connected at the stem which was similar a pair of testicles, and images of keys which looked similar to a shaft with a single testicle.

7.3 Convolutional neural networks

The classification in this project was done by a convolutional neural network (CNN) because of its established success rate in whole image processing projects, such as AlexNet (Krizhevsky, 2012). In short, CNNs assign values to each pixel in an image, which may derive from RGB values, but the images are often converted to greyscale in preprocessing since it reduces the pixels to one channel, minimizing the computational costs. CNNs considers regions of pixels, i.e. kernels, within the image and, if max pooling is used, the most representative value among those pixels is assigned to that region (Cireşan, 2011). After this, the region of interest moves to the next region, which may overlap with the previous, and the amount of overlap will depend on the CNNs stride, i.e. the number of pixels away from the previous kernel's furthermost points it will move. As the image processing proceeds, CNN produces a feature map of the images (see Figure 7). These features are given weights based on the degree to which they appear in the image data. Ideally, these features will not be found in the non-target classification, which is why having a large and varied group of control images is

important to avoiding false positives. Once the CNN has been fed all the training and testing data, the best model is produced and may be used in the classification of new images, which may also be added to the model later. Faster, more efficient models such as Fast R-CNN may also be applied to the task of image classification, but this neural network and others are designed for object detection, rather than the whole image classification needed to accurately label vulgar images (Girshick, 2015). For the thesis, Keras (https://keras.io/), the high-level neural network API, provided a sequential model for organizing the network layers. These choices were made for the sake of simplicity and potential scalability since a SocialVR platform's CNN will be added to and summarily improved over time.

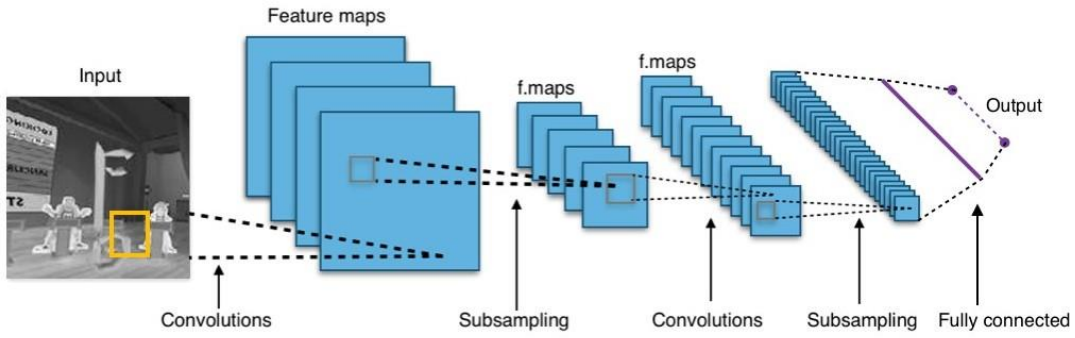

Figure 7 Full convolutional neural network – By Aphex34 [CC BY-SA 4.0], via Wikimedia Commons, input image from Rec Room

7.3.1 Results

The *loss*, a measurement of how well the model is behaving, recorded for the model built on the data set described above was 0.0494 and the *accuracy* of the model was

98.65%. If these percentages held up in the presence of new images, the CNN would have been sufficient for the task of image classification.

A hundred more images, fifty assigned to each label, were tested with this CNN model to check its performance and the confidence scores for the images were recorded to get more details about the performance. The combined accuracy of both genitalia and non-genitalia images in the model was 78%, which is high enough to be useful in image classification, but still far too low to allow for unsupervised classification.

As stated, these were the accuracy scores from the first high performing CNN and numbers are expected to continually improve as they are given more data. In its current state, the model can be used for a supervised harassment classification in which a human looks through the images classified as male genitalia, confirming and disconfirming them. At the moment, this model had a true positive to false positive ratio of approximately 3:1, but it is assumed that the false positive ratio would be substantially higher because one would not expect half of all drawings in a SocialVR platform to be male genitalia. Alternatively, the true to false classification percentages in which the model was nearly certain (over 99.9%) was 31% TP and 5% FP, which is a ratio of just over 6:1, which would help a platform only considering positive classifications with this high level of confidence filter out many of the images, saving time in the supervised approach.

7.3.2 Implementation of CNN

A SocialVR platform wishing to implement CNN image classification would collect images in a fashion similar to methods in the training data in this project. When the instrument is in hand, drawing in three dimensions requires holding a trigger on the controller and releasing it. When the trigger is held long enough (i.e. 1.0 seconds) and released, a two-dimensional image of the drawing will be captured, much like someone taking a photograph, and the image will be saved to the platform's database, where it can be evaluated. Many SocialVR platforms already include a camera that allows users to take pictures (*Rec Room*, *AltspaceVR*, *TheWaveVR*, *High Fidelity*), so this camera may be repurposed for invisibly and inaudibly capturing the users' drawings; the platform would only need to find the ideal vantage point.

As data is collected from users who are abusing the drawing tools, the performance will become more robust and it may become possible to have unsupervised classification. New CNNs can be build and expanded to include other vulgar or offensive images. In the qualitative data, derogatory drawings of female breasts were included and once a taboo word was written out with a three-dimensional pen, and image classification should be equally capable of removing them and restoring a non-hostile environment.

7.4 CNNs for NLP

CNNs are a powerful tool for their ability to evaluate machine data that can be represented numerically. As seen, this is true for images, it is true for audio and, given

an adequate data set for training, it is also true for representation in NLP. Neural networks have the ability to find lexical items, nominal pairs, and sentence features that regularly appear in discourse instances already labelled as harassment (Shen, 2014). They can analyze them in apposition to other words and give a probabilistic answer regarding a phrases harassment status. However, these surface level representations of meaning (i.e. the words as they are spoken) may also go through a semantic-level analysis, which serves as a type of lexical preprocessing that helps the CNN uncover characteristics of the words and speakers' underlying intent (Gao, 2014). Using semantics analyzers and CNNs may turn out to be an improvement over the NLP solutions developed in the current project insofar as it reduces false negatives in harassment, but these CNNs will require much more data than is currently available.

CHAPTER 8

CONCLUSION

Having provided methods for gathering data and the methods for using that data to create harassment detection tools, we can go into detail on their methodologies, design considerations, how they could be integrated into the SocialVR platform, and how they might be improved upon when more data is available.

8.1. Discussion of qualitative analysis methods

At the time of collecting data for this project, little or no formal discourse analysis work had been published in modern SocialVR, so many of the standards had yet to be established. In the time since data collection, a full multimodal critical discourse analysis was published in the thesis work of Claudia Maneka Maharaj (2017) where the focus was individual and group representation in SocialVR. The desire for naturally occurring data must be maintained to make any observations about the discourse but, as both a research and user, where to place yourself in order to overhear a conversation is not immediately clear; you want to remain close enough to listen to the speakers, but not so close that you get drawn into the conversation. Statements made in conversation in a common area within SocialVR are like those made in a public chat room because the speakers have no control over who hears it. SocialVR platforms maintain their right to share anything that happens in SocialVR and, likewise, other users maintain a right to report on their experiences. This information, however, is not explicitly understood by

all users and they may at times wrongly assume they are having a private conversation. For this reason, I did not make any attempt to physically hide from other users, instead staying close enough hear what was being said and visible to anyone who valued their privacy enough to look around themselves before speaking.

When arriving in the virtual environment, it may be appropriate to engage with other users politely, but I found it helpful to remain generally reserved in conversation. Discourse analysts would not want to inadvertently trigger a harassment event since it would not be considered naturally occurring. Among larger groups of people, five or more, it is generally easy to avoid engaging anyone in conversation, but it is more difficult to avoid speaking when the group is small since the presence of each user becomes more obvious. Whenever users initiated conversation with me, I maintained a polite demeanor, but gradually withdrew from the conversation when more people joined in and remained completely silent, not even laughing while they conversed since it may be seen as condonation of the activity (Revi, 2015). There were instances during the qualitative analysis when I became the target of a harassment event. In these cases, it was most helpful to remain expressively neutral, neither approving nor disapproving. If possible, it helps add actionable data to the qualitative analysis if you ask for clarification when being harassed, asking questions, such as, ‘What do you mean?’ or ‘What is that?’ This may give clues as to the motives of the speakers, whether they are being intentionally harassing or not.

Spending hours in SocialVR collecting discourse analysis data will also teach you which personality type to watch for. In this observation, users who moved around often, spoke loudly, and broke into other users’ conversations were more likely to harass

also. There were also instances of users observing a harasser's proclivities and joining their antics in a supporting role. There were harassers that do not fit this description but, in this study, following the loudest, most mobile users produced the most data.

Transcribing from SocialVR is made difficult by not being able to source speech, not knowing to whom speech is directed, the large amount of crosstalk, and difficulty recognizing physical gestures. Therefore, many of the noises made in SocialVR might be ignored because they will be impossible to source to an individual user unless someone reacts to them. When multiple conversations are occurring, the researcher may choose one of them and ignore everything that seems to belong to a separate conversation. However, if both conversation are of interest or the conversations start to blend as speakers move between conversational groups, two separate transcriptions can be made.

When the data is collected, all usernames or references to usernames are changed to protect their privacy. References to their geographic location are also given an alias. The gender of the speaker, their political affiliations, and, if applicable, the youthfulness of their voice is included in the transcript. If a harassed user mentions or confirms their race, this is also kept since it is likely to be important to the conversation, especially where hate speech is involved.

8.2 Data collection for lexically based analyses

Before enacting an automated process of harassment classification which will also take unsupervised action against the harassers, the program should be tested by running it on

current users. Their discourse may be recorded in text files, highlighting the language thought to be vulgar or hateful. The surrounding context in their conversation may also teach key words and structures determinate in the detection of harassment, which may be added to the NLP program. Data on the offending users may also be used to improve the user profiling scores if they are used. Scoring simulations may also be run on users, which will allow the platform to informedly modify its scoring to eliminate false positives and false negatives. When the program has proven its ability to run unsupervised, it may be implemented in the platform.

For the sake of data protection, SocialVR platforms should strongly consider anonymizing text data either by assigning a user number or encrypting the names of users. They might also consider deleting the text files after a pre-determined period. The text files should be searchable according to time and user in the eventuality that someone lodges a harassment complaint against another person and a review of the case is required. Mention of the SocialVR platform's right to keep a record of events taking place in their platform should be expressed in the terms of service, but it should not be explicitly mentioned under any other circumstances.

There are challenges found in this program which are familiar to all natural language processing, especially sentiment analysis (Srivastava, 2017). From the text alone, it may be difficult to determine whether users are being combative or joking. Users who are playing a game may be invoking *trash talk*, in which players will insult each other and their abilities, but the decision to do so might be mutual and lighthearted in this context (Rainey, 2010). It is important raise or remove thresholds between friends since they are more likely to speak with familiarity and that could be mistaken for

harassment. Videogame culture also presents a similar problem since many of the participants, speaking in the first person, will speak about violent acts, which would be horrific outside of that context.

Another challenge comes from mimicry or repetition of harassing statements. It has been found in the qualitative data that harassed users will sometimes repeat harassing statement to express shock, to direct the statement back at the harasser, or to report the harassing statement to a neutral third-party user. A positive classification that is false would be especially undesirable in these cases since the harassed user is being wronged by both the harassing user and the SocialVR platform. This makes an especially strong case for maintaining a supervised program until collecting adequate data.

The lack of human error in speech-to-text processing is one advantage it has over text processing. Harassers and trolls who do not wish their speech to be filtered out can easily manipulate text to make it difficult for natural language processors to comprehend (Srivastava, 2017). They may do this through using phonetic misspellings for vulgar language, approximate spelling, and swapping similar looking characters. The same techniques cannot as easily be done when users' speech is faithfully transcribed.

8.3 Data collection for image processing

There were no publicly available vulgar images within the major SocialVR platforms. An inquiry about image data was made to a few SocialVR platforms, but they either denied having any or considered the data confidential. The potential legal jeopardy of

asking a third party to draw vulgar images for the project meant that nobody else could. Therefore, it would be necessary for me to produce the images.

The benefit of building a CNN model of vulgar images and using it for supervised classification is that, at the time of writing, the userbase of SocialVR platforms is small, relative to social media platforms, numbering in the thousands or tens of thousands of monthly active users. Using Rec Room as an example, there would typically be two or fewer users drawing with the 3D pen in a public area at a given time. After being classified by the CNN model, thumbnails of the images that are then placed in a folder could be scanned for vulgarity quickly and the true positives could be confirmed. The true positives and true negatives could be added to the training data and help improve the image classification in the interest of creating an unsupervised system that is prepared for a rapidly expanding userbase.

8.4 Interventions against positive harassment classification

The majority of SocialVR platforms include one or more tools for ending a harassment situation, but their implementation requires the harassed person's knowledge of the anti-harassment tool, their access to the tool, and their willingness to take initiative against harassers. Platforms can try to inform users of their harassment tools in the tutorial and sending users updates, but there will always be those who forget about the tools when they need them. Granting access to anti-harassment measures takes great skills in user interfaces because, when needed, users should be able to get to them quickly, but the tool should not be obtrusive to the user experience. Having harassed users leave a

location or sign out, thinking it is easier, is undesirable since they would have left with a negative experience and the harasser may freely continue in their anti-social behavior. Finally, harassed users may not wish to create conflict with a harasser by reporting on their behavior. If someone is the sole target of harassment, they may believe other users have sided with the harasser and will not see the point of taking action against them. Therefore, effective anti-harassment tools would overcome each of these problems in the interest of protecting harassed users.

8.5 Lexically based scoring

If the scoring methods from this project are adopted, then users will have a harassment score upon signing into the SocialVR platform and their score will change with regard to the environment and not their actions. The lexically based score increases based on the users' behavior and this project includes one threshold which, once reached, will trigger a response from the SocialVR platform, but there may be good reason to increase the number of score-based thresholds to three. For example, the first threshold could be 80 points and it flags a user's text file for review by the platform's staff. Reaching the second threshold, set at 100 points, could trigger an in-game response to the supposed harassment which would require an action by users near the potentially harassing users. Reaching the final threshold with a score of 120 points could lead to the supposed harasser's removal from the environment. This is only an example and SocialVR platforms should test different responses within their applications; for example, the actions taken in the first and second thresholds may be combined to fall under a single threshold score.

The in-game response mentioned in the second threshold would be a notification sent directly to the screen of people near the harasser, but not the harasser themselves. The notification would ask for confirmation that the harassing user is, in fact, engaging in harassing behavior. This notification compensates for the shortcomings of the existing tools by being quick and teaching non-harassing users that it is acceptable to stand up to users who engage in anti-social behavior. Confirmation of their behavior will result in the harasser being removed from the environment while disconfirming the harassment will end in no actions against the users, apart from a member of the platform staff reviewing the transcript of the supposed harassing event. In case of disconfirmation, notifications to the potential harassment victim asking about the event should not be sent again regarding the same harassing user.

Removal from an environment, whether initiated by a threshold or the report of another user, could fall under a few different types. Users can be immediately suspended from a platform, the duration depending on the severity of the harassment event, or they can be permanently removed from the platform. SocialVR platforms have a user limit for rooms in their environment and new, identical rooms are set up and filled whenever the user capacity is reached. If the number of users is high enough, offending users could be temporarily or permanently moved to a room in the platform where they may have more limited contact with other users. For example, users will only have contact with their friends, they will only have contact with other harassers, or they will be kept separate from users that are considered more vulnerable, such as new users. The platform could also remove their rights to public or common areas in the platform,

meaning they would be restricted to private areas and only encounter people in the platform they have personally invited.

As the SocialVR session continues, if a lexically based harassment score accrues but a threshold is not reached within a period of time, the harassment score coming from user behavior should gradually decrease. For example, someone who uses the same swear word twenty times in a two-minute period may be considered worse than someone swearing twenty times in two hours. It is not the intention of SocialVR platforms to discourage swearing or the discussion of emotionally charged topics, but their dense usage may be an indicator of abuse and not informal discussion.

8.6 Handling positive classifications of vulgar images

When an image is created and positively identified as being vulgar, the platform has a few good options for minimizing the damage and preventing harm in the future. If the image classification system is to the point of being unsupervised, the platform can erase the image, thereby limiting who sees it and for how long. If the user persists by drawing a vulgar image a second time, the platform should be warned against erasing the next image. Harassers are often known by their persistence and it is unadvisable to openly challenge their anti-social behavior because it will backfire, encouraging them to draw many more vulgar images of increasing complexity (Revi, 2015). It is preferable for them to believe that some glitch has removed the image and not a censorship tool on the platform, so either they will draw the image again, not notice the disappearance, or give up. The result is that the number and duration of vulgar images will be reduced, not

eliminated, but this action will also not lead to a net positive in the number of images when the harasser understands what is happening.

Once the user produces a vulgar image and action is taken upon it, the platform must also choose what action to take against the user drawing it. As with lexically-based offences, users may be suspended or permanently banned. The platform could continue to permit the user to the platform, but also choose to disallow use of drawing tools by those users. For example, an offending user would no longer be able to lift the three-dimensional pen they used to draw the vulgar object. This would allow the platform to maintain their user numbers, but also mitigate any harm to the entire user base due to anti-social behavior.

It would take a great deal of time to remove every vulgar image since the platform is competing with the vast ingenuity of all harassers, but lessening the perception that vulgar imagery is an aspect of the cultural norms within the SocialVR platforms will greatly accommodate a wider user base. The lessened incidences of disgust will lead to a greater attraction to the platform while leaving the harassers more isolated and ready to conform to the platform's standards of behavior.

8.7 Other anti-harassment tools

The existing tools to prevent or report harassment may be kept in place, but the difficulty inherent in them is their potential for abuse by the harassers themselves. If a user is known to abuse the anti-harassment tools by wrongfully flagging or reporting other users, the abuser, in this case, can have their ability to report other users covertly

removed; this means the buttons may still be in place, but they have no real effect. If one such case requires review, the platform may decide that the reporting user should have punitive actions taken against them. These false reports are known to happen in all circumstances, including when a player is disgruntled about the outcome of a game.

8.8 Future research

This project is intended to be a framework for studying SocialVR and the application of these methods of discourse analysis need not be confined to harassment. Many questions that have been posed regarding interpersonal relations and the use of physical cues during in-person speech can be asked again in SocialVR. This project has raised the question of how much personal information people are willing to share in SocialVR and how that compares to speaking with new people face-to-face. There have also been cases of people responding physically to a change in the environment that did not require any such response, for example, ducking when an object is thrown at a user or walking around objects when one could walk through them. Psychological studies on the effects of user and environment customization, and how much people associate with their own avatars, may also be done. Subjects can also be assigned to cooperative and competitive tasks to learn the sociolinguistic qualities of their interactions. Current hypotheses dealing with the adverse effect of anonymity on behavior may be also be tested in the new medium.

More relatedly to the aims of this research, the implementation of the proposed speech processing programs will enable the collection of harassment data which was

largely absent when the project began. Once there is adequate data available, the algorithms for understanding large amounts of discourse data can be improved upon. There will be an ability to apply sentiment analysis and more effectively find patterns in user speech, and responses to harassing speech, which better show that harassment is taking place. Researchers will be able to study sentiment analysis of spoken discourse mediated by SocialVR to the analysis of online texts, such as tweets, and find points of comparison. This practice can also apply to the development of market research, political and consumer focus groups, and the like.

At the start of this project, there were also scarce image data available. As with text data, the collection of more images from drawings in SocialVR platforms will increase the effectiveness of new and existing neural networks. More neural networks can be tested for performance, both speed and accuracy, and they may be compared until an optimal network is found. Other methods of data collection may be attempted, such as mapping the movement of the controller as it produces a drawing as is done in two dimensions with Sketch RNN (Ha, 2017). The method of collecting a single still image for image processing may also be expanded to include multiple angles, revealing the depth of an object alongside the height and width, which may more accurately classify it (de Vos, 2016).

Future research can work on acoustic event detection in SocialVR in the interest of detecting inappropriate sounds, such as simulated orgasm or the pre-recorded sounds of orgasm from played from pornographic material (Phan, 2016). This harassment event was found in the qualitative data and the current methods rely on lexical content to identify it, rather than moans or gasps. This approach would also need to distinguish

these sound from other involuntary sounds such as laughter, but this type of audio analysis through neural networks is already being performed in other domains (Amiriparian, 2017).

Given a more complete understanding of user behavior, researchers would also be able to consider user movement as it applies to harassment classification. They can analyze the character of movements being made to evade a harasser; likewise, they may understand when a harasser is chasing another user in the interest of abuse. More complex coordinated movements may also be looked at, such as how conversational groups will form clusters as harassers move into and around the perimeter. In the cases of multiple harassers, swarming behavior, where the harassers will approach and crowd a single target, is known to occur. All of these movements were found generally in the qualitative analysis, but the data was insufficient for both recognizing it as a pattern and describing the movements in great detail. Movement data alone may not be enough to classify a harassment event, but it could be used in combination with other data, such protestations from the harassed user, for harassment classification, or certain sequences of movement may simply be added to the harassment score.

Finally, a SocialVR platform with the ability to collect data on every harassment event in their platform will have the ability to better define and understand harassment itself. A qualitative analysis can reveal which types of harassment have occurred and infer patterns from those occurrences, but a quantitative analysis will be able to test those hypotheses more completely. Researchers may learn how frequently sex and race-based harassment occurs, and how these forms of harassment are expressed linguistically. They may see how people respond to harassment and the most effective

strategies for dealing with them, both interpersonally and through platform tools. The study of street harassment relies heavily on surveys and self-report because collecting a sizable amount of data comes with practical and ethical limitations. SocialVR, to an extent, is a laboratory-like setting for the study of human behavior since the visual and auditory environment is controlled. All verbal data may be collected as it is transferred between users and the qualities of their discourse becomes an abstraction which may be applied to the “real world”.

APPENDIX A

SUMMARIES FROM THE QUALITATIVE ANALYSIS

These are descriptions of the harassment behavior observed during the qualitative analysis and serve to summarize the harassment information as it appears in the transcript. These instances informed the NLP program written to classify harassing behavior. They illustrate the types of harassment that can and do happen in SocialVR.

Session One:

- Justin mimes ejaculation with the aid of a stick-like prop available in the environment. The action is accompanied by Justin's mimicry of sexual sounds. The performance is not easily avoidable by others within the virtual environment due to its high volume and their largely unobstructed view.
- Justin uses a derogatory word for little people in a public setting.
- Justin harasses Steve by remarking on his height as being ideal for the performance of oral sex. When Steve rejects participation in the taboo act and moves away, Justin demands that he return to the spot, allowing Justin to receive oral sex from him.
- Chris attempts hip thrusts towards Steve, which would mime a sexual act. Steve had already withdrawn his consent both verbally and physically.
- Chris and Justin make noises, such as howls and moans, which could only be associated with sexual activity. It is done loudly enough so that anyone in the virtual environment could not avoid hearing it. Additionally, the recipient of the sexual behavior had already said that he did not want to participate.
- Justin demands that Steve perform a sexual act on him although Steve had already expressed his disinterest.
- Chris discusses the best method of trapping Steve, most likely for the sake of performing sexual acts on him.
- Justin references a serial killer while using a strange, high-pitched voice. This is most likely intended to make the victim of harassment uneasy.
- Justin stands behind Skylar and speaks to her softly, which is a physically intimidating stance. Furthermore, Justin uses a diminutive term for blonde-haired women to refer to her, which suggests malicious intent.

- Justin demands that Skylar not move and be a recipient of something. It is likely a threat of forced sexual activity.
- Justin mimics sounds related to sexual activity while miming the fondling of his own genitalia in front of Skylar who has already expressed a disinterest.
- Justin discusses touching parts of Skylar's anatomy in an erotic or potentially abusive manner.
- Justin invites Chris to perform a sexual act on him. Since Chris has expressed his openness to discourse regarding sexual topics, the statement is only considered harassment because of its public manner. Said privately, this statement might not have been considered harassment.
- Justin discusses putting sexual excretions on Skylar's body. He is prompted by the innocuous use of the word '*come*' in Skylar's conversation, but he responds using the sexual homophone.
- Justin, Chris, and Sid physically intimidate Fado by surrounding her.
- Justin openly discusses alterations to his own genitalia and describes the acts that could be performed with that alteration.
- Justin threatens violence against a female avatar, Fado, while using a derogatory term for women. He mimes punching the female avatar and mimics the sound punching her would make. She moves away from him as a sign that the interaction is unwanted and later tells them to stop what they are doing.
- Justin and Sid make unwanted demands of Fado, which pertain to a diminutive role for women and wives in general. The harassment is towards Fado because of their discriminatory treatment and also listeners who are offended by the subjugation of women. This is repeated multiple times.
- Justin refers to the role of child bearing by women in a diminutive manner. He discusses violence as a means of ending a pregnancy, a topic that may be considered taboo and disturbing by some listeners. Sid also attempts to support Justin in his line of harassment.
- Justin indicates that committing violence against Fado is permissible because she is his wife. Fado is being harassed in that the title is forced on her by someone she is actively avoiding. Likewise, this may be seen as the endorsement of husband-to-wife spousal abuse, which has the high potential to trigger victims of physical abuse.
- Sid references Fado's poor economic status, which serves to reinforce a stereotypical narrative about her that had been entirely fabricated by Justin earlier in the dialogue.
- Justin repeats earlier allegations about the legality of David's immigration status and expands upon it by alleging that, financially speaking, he is a non-contributing member to society. Justin follows up by using vulgar language in an aggressive manner against David.

- Justin directs unusual, loud, and uninterpretable sounds at David. This limits David's ability to respond to the harassment and the sound is likely to be irritating to the hearers.
- Justin mimes physical violence against David. Meanwhile, Chris narrates the interaction, which is not yet harassment, but Chris becomes a participant in the harassment by verbally intimidating David.
- Chris ridicules David for his Latino accent.

Session Two:

- Biff is advocating two forms of violence against a racial minority within the dialogue. The first instance is against a racial group and the second instance is against a gender. The particular form of violence is irrelevant since they could be used interchangeably and still remain harassing statements. The approval of the listeners is also irrelevant because these harassing statements were said in a public forum and also belong to the sub-category *hate speech*.
- Biff also advocates that Mac performs a sexual act upon himself. This is not considered harassment against Mac, who has already proved comfortable with breaking taboos regarding sexual acts in the presence of Biff. The harassment is towards the bystanders listening to the sexual statements being made by Biff who may consider the statements to be unwanted.
- Mac harasses Billy by attempting sexual contact with Billy after he had warned Mac that the contact was unwanted. The fact that Billy is laughing and seems to think it is funny is irrelevant since Billy does not reciprocate the sexual activity through his words or actions. It is possible that Billy laughed as a means of conflict avoidance or smoothing over social awkwardness.
- Biff creates a hostile environment by expressing an intent to perform general violence within the space.
- Biff advocates discriminatory violence against another avatar due to their apparent height. Biff uses a word considered to be derogatory against short people and hearing the term could be unwanted by the listeners as well as the referent in the dialogue.
- Mac uses a term for homosexuality in a derogatory manner. The harassment is not against Biff since he had already instigated the breaking of multiple social norms, making him a willing participant. Instead, the harassment is against bystanders who may not wish to hear the term used in a derogatory context.
- Biff advocates violence against a targeted racial group, which is hate speech.

Session Three:

- The harasser introduces a sexual topic among people with whom he lacks adequate familiarity. The harasser suggests that the two people near him engage in sex. When the suggestion goes unanswered, he persists in the questioning. Since the pair of harassed people leave without speaking soon after hearing the harassing questions, it is evident that the questions were unwanted.

Session Four:

- B-Man's verbal harassment comes from expressing displeasure at being around too many members of an ethnic group.
- B-Man's non-verbal harassment comes from lewd and repetitious movement of his body in proximity to Bubba. Due to Bubba's short stature and youthful voice, it's possible that he was underage and may not have understood the significance of the gesture. However, Bubba's response to the second series of thrusts was an inquiry into the motivation of the action and it was not answered. The response to the third series of thrusts was finally a request to stop, perhaps meaning that the action was unwanted from the beginning.

Session Five:

- Derek repeatedly uses vulgar language, but the fact that it is reciprocated among other users keeps it from becoming harassment. There is a moment of harassment at the end of the dialogue where Derek withdraws from the group to hold a conversation with someone outside of VR. Mona continues to engage with him playfully, but Derek had already signaled that the interaction had shifted from wanted to unwanted, making it a brief instance of harassment on Mona's part.

Session Six:

- Julian uses vulgar language aggressively towards people he had not yet spoken to. This is an immediate disregard for standards of politeness. Julian does not wait for feedback on the vulgar language to gauge the feelings of the listeners. Pike asks why he is getting such negative treatment from Julia, which could be a sign that it is not wanted, but Julia does not respond to the question.

Session Seven:

- Kootie verbally harasses listeners by subjecting them to a song that is likely to discriminately offend member of religions that consider Jesus to be a holy figure. This may also be considered discriminatory towards homosexuals since they are treated as being apart from other sexualities.
- Commty harasses Lako by sketching and displaying a vulgar image to him. Commty had not spoken to Lako, much less gained consent, before showing the image to Lako.

Session Eight:

- Sole independently steers a conversation to sexual topics in a public setting without knowing what is being discussed.
- Sole uses gestures to mimic the sexual act. Since the action is public and performed without warning, it is unlikely to be wanted by Ren, the recipient. The fact that the act is jokingly responded to by Brown does not mean that the act itself was wanted.
- An unknown user exposes all users within the vicinity to the sounds of simulated orgasm without consent or warning. The exposure is likely unwanted by at least one listener given that Joey responds with seeming disapproval or disbelief while Harold's response might be interpreted as confusion or curiosity.

Session Nine:

- Bill and Walt participate in creating vulgar imagery and then mimic a sexual act using the imagery after its creation. They continue performing the sexual act after Navi confirms that the performance is unwanted.

Session Ten:

- Sammy mimics sexual activity with Shania without her consent and initially without her knowledge. He also makes noises associated with physical intimacy. Tex points out that this is harassment and he gives Shani instructions on how to respond.

- Tommy comes close to Shania and speaks softly to her, which can be perceived as threatening. When asked about it, Tommy gives an answer that seems purposefully difficult to interpret, giving it the potential to escalate further.
- Bob violates Shania's personal space by enveloping her avatar in his own.
- Peter repeatedly feeds Shania, compounding the already unwanted attention that is taking place.

Session Eleven:

- Evelyn, who inhabits a female avatar, seems to have left VR, but her avatar is still in the environment. A group of four avatars, each of them with male voices, discuss perform sexual acts on Evelyn as a group. Evelyn becomes physically active once more, but it is unclear whether she heard the conversation the four avatars were having. Seeming uncomfortable, Evelyn leaves the virtual environment.

APPENDIX B

CATEGORIES OF HARASSMENT CONSTRUCTIONS

1. Lexically Based Lists

- Controversial Topics: immigration, immigrant, abortion, porn, pornography,
- gay, homosexual, lesbian, gun, sex, holocaust, furry, fetish, cougar, prostitution
- Swear Words: fuck, shit, bitch, ass, bullshit, pain-in-the-ass
- Abusive Terms: cocksucker, fucker, bitch, whore, asshole, slut, retard, motherfucker
- Taboo Terms: c***, n*****, f**
- Intensifier: fucking

2. Harassing N-Grams

- Fuck N-grams: you, off, up
- Blow N-grams: me, you, job
- Shit N-grams: tough, eat, head
- Jerk N-grams: circle, off

  Examples:

  - Blow me.
  - Eat shit.
  - Jerk me off.

3. Name Calling

- Offensive Adjectives: stupid, ugly, fat, dumb, idiotic, brainless, nasty, retarded, filthy

Examples:

- You prick.
- You are a fucking ugly prick.

4. Threats of Violence and Hate Speech

- Violence Verbs: choke, kill, murder, strangle, slaughter, massacre, annihilate, destroy, stab, lynch, hang, shoot, hit, punch, kick, torture, decapitate, behead, rape, beat
- Auxiliary Verbs: want, going, will, should, would
- Exclusion Participles: banned, blocked
- Groups: black, white, chinks, women, bitches, f**s, dikes, homosexuals, Muslims, Jews, kikes, c***s, retards, gays, n*****s, democrats, republicans, yellow, Arabs, whores

Example:

- Women should be punched.
- In my country, homosexuals would be burned.

5. Unfavorable Comparisons of Groups

- Comparing Verbs: are, is, look, like
- Bad Adjectives: stupid, lying, smelly, evil, sinister, filthy, ugly
- Bad Nouns: thieves, dogs, monkeys, pigs, apes, maggots, shit, scum, trash, garbage

Example:

- Democrats are maggots.
- White people are ugly fucking dogs.

6. Promotion Self-Directed Harm

- Self-Harm Verbs: kill, shoot, hang
- Reflexive Pronouns: yourself, themselves, myself

Examples:

- Republicans should kill themselves.
- Why don't yellow people cut themselves.
- Kill yourself.

7. Discussing Death of a Group

- Death-Related Words: die, dead, death

Example:

- Jews should die.
- Die Jews

8. Demanding Sexual Activity

- Sexual Abuse: blow, ass, anal, lick, rape, molest, jack, ride, spank, finger, bang, suck, touch, feel, kiss, girth, jab, thrust, poke, ram, fuck, pound, fist
- Sex Descriptors: dirty, hardcore, hard, wet, throbbing, sweet, filthy
- Body Parts: vagina, pussy, tit, nipple, asshole, c***, breast, tongue, scrotum, testicle, ass, cock, prick, dick, penis, throat, cherry, crotch
- Body Parts2: head, eye, nose, ear, knee, shoulder
- Body Parts3: leg, thigh, face, mouth, finger, hand
- Prepositions: in, on, between, up, with
- Ejaculation Verbs: come, ejaculate, cream, spurt

Example:

- I want to suck your finger.
- I'm going to ride your hardcore prick.

REFERENCES

Amiriparian, S., Gerczuk, M., Ottl, S., Cummins, N., Freitag, M., Pugachevskiy, S., Baird, A., & Schuller, B. (2017). Snore sound classification using image-based deep spectrum features. *Proc. of INTERSPEECH*, *17*, 3512-3516. Retrieved from http://www.isca-speech.org/archive/Interspeech_2017/pdfs/0434.PDF

[The Awesome Patman]. (2013, January 8). */b/ Habbo Hotel Raid – 2013* [Video file] Retrieved from https://www.youtube.com/watch?v=6TDgAfQJRYY&t=116s

D'Anastasio, C. (2016, October 26). VR developers add 'superpower' to their game to fight harassment. Retrieved from http://kotaku.com/vr-developers-add-personal-bubble-to-their-game-to-fi-1788237241

Barrett, L. F. (2006). Are emotions natural kinds? *Perspectives on Psychological Science*, *1*(1), 28-58. Retrieved from http://journals.sagepub.com/doi/abs/10.1111/j.1745-6916.2006.00003.x

Belamire, J. (2016, October 20). My first virtual reality groping [Blog Post]. Retrieved from https://medium.com/athena-talks/my-first-virtual-reality-sexual-assault-2330410b62ee

Binns, A. (2012). Don't feed the trolls! Managing troublemakers in magazines' online communities. *Journalism Practice*, *6*(4), 547-562. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/17512786.2011.648988

Sprague, W.C. (1915) *Blackstone's Commentaries Abridged* (9th ed.). Chicago, IL: Callaghan and Company.

Bleich, E. (2014). Freedom of expression versus racist hate speech: explaining differences between high court regulations in the USA and Europe. *Journal of Ethnic and Migration Studies*, *40*(2), 283-300. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/1369183X.2013.851476

Bowman, C. G. (1993). Street harassment and the informal ghettoization of women. *Harvard Law Review, 106*, 517-580. Retrieved from http://www.jstor.org/stable/1341656

Bozgeyikli, E., Raij, A., Katkoori, S., & Dubey, R. (2016). Point & teleport locomotion technique for virtual reality. *Proc. of the 2016 Annual Symposium on Computer-Human Interaction in Play*, 205-216. Retrieved from https://dl.acm.org/citation.cfm?id=2968105

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223-242. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/poi3.85/full

Cireşan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, *22*(1), 1237-1242. Retrieved from http://www.aaai.org/ocs/index.php/IJCAI/IJCAI11/paper/download/3098/3425

Cole, S. S., Denny, D., Eyler, A. E., & Samons, S. L. (2000). Issues of transgender. *Psychological Perspectives on Human Sexuality*. 149-195. Hoboken, NJ: John Wiley. Retrieved from http://psycnet.apa.org/record/2000-07452-004

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*, Cornell University Library, New York City, NY. Retrieved from https://arxiv.org/abs/1703.04009

Duggan, Maeve (2017, July 11). Online harassment 2017. Retrieved from http://www.pewinternet.org/2017/07/11/online-harassment-2017/

Ehrenkranz, M. (2017, February 9). Trolls keep outsmarting anti-abuse tools. Will Twitter's new system actually work? Retrieved from https://mic.com/articles/168041/trolls-keep-outsmarting-anti-harassmenttools-will-twitters-new-system-actually-work

Epstein, D. (1995). Can a ‘dumb ass woman’ achieve equality in the workplace? Running the gauntlet of hostile environment harassing speech. *Georgetown Law Journal*, *84*, 399-452. Retrieved from http://heinonline.org/HOL/LandingPage?handle=hein.journals/glj84&div=25&id=&page=

Fox, J., & Tang, W. Y. (2016). Women’s experiences with general and sexual harassment in online video games: rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media & Society*, *19*(8), 1290-1307. Retrieved from https://doi.org/10.1177/1461444816635778

Gao, J., Deng, L., Gamon, M., He, X., & Pantel, P. (2014). *U.S. Patent Application No. 14/304,863*. Washington, DC: U.S. Patent and Trademark Office.

Geen, R. G., & Stonner, D. (1975). Primary associates to 20 verbs connoting violence. *Behavior Research Methods*, *7*(4), 391-392. Retrieved from https://link.springer.com/article/10.3758%2FBF03201552?LI=true

Girshick, R. (2015). Fast r-cnn. *Proc. of the IEEE international conference on computer vision*. 1440-1448. Retrieved from https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf

Gitari, N. D., Zuping, Z., Damien, H., & Long, J. (2015). A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, *10*(4), 215-230. Retrieved from https://preventviolentextremism.info/sites/default/files/A%20Lexicon-Based%20Approach%20for%20Hate%20Speech%20Detection.pdf

Ha, D., & Eck, D. (2017). A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, Cornell University Library, New York City, NY. Retrieved from https://arxiv.org/abs/1704.03477

Han, D. T., Sargunam, S. P., & Ragan, E. D. (2017). Simulating anthropomorphic upper body actions in virtual reality using head and hand motion data. *Virtual Reality 2017 IEEE*. 387-388. Retrieved from http://ieeexplore.ieee.org/abstract/document/7892339/

Harris v. Forklift Systems, Inc., 510 US 17 (1993). Retrieved from https://www.oyez.org/cases/1993/92-1168

Herring, S. C. (1999). The rhetorical dynamics of gender harassment online. *The Information Society*, *15*(3), 151-167. Retrieved from http://www.tandfonline.com/doi/abs/10.1080/019722499128466

Higgin, T. (2013). FCJ-159/b/lack up: what trolls can teach us about race. *The Fibreculture Journal*. Retrieved from http://twentytwo.fibreculturejournal.org/fcj-159-black-up-what-trolls-can-teach-us-about-race/

Huff, C., Johnson, D. G., & Miller, K. (2003). Virtual harms and real responsibility. *IEEE Technology and Society Magazine*, *22*(2), 12-19. Retrieved from http://ieeexplore.ieee.org/abstract/document/1216238/

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097-1105. Retrieved from http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks

Lewicki, P., Hill, T., & Sasaki, I. (1989). Self-perpetuating development of encoding biases. *Journal of Experimental Psychology: General*, *118*(4), 323-337. Retrieved from http://psycnet.apa.org/record/1990-09019-001

Maharaj, C. M. (2017). *Embodiment and the boundaries between us in virtual reality-A critical analysis of inclusivity in social virtual reality environments*. (Master's thesis). Retrieved from http://muep.mau.se/handle/2043/23611

Mouttapa, M., Valente, T., Gallaher, P., Rohrbach, L. A., & Unger, J. B. (2004). Social network predictors of bullying and victimization. *Adolescence*, *39*(154), 315-335. Retrieved from https://search.proquest.com/docview/195934334?pq-origsite=gscholar

Mukhra, R., Baryah, N., Krishan, K., & Kanchan, T. (2017). 'Blue Whale Challenge': a game or crime?. *Science and Engineering Ethics*, 1-7. Retrieved from https://link.springer.com/article/10.1007/s11948-017-0004-2

Nielsen, L. B. (2000). Situating legal consciousness: Experiences and attitudes of ordinary citizens about law and street harassment. *Law and Society Review,* 1055-1090. Retrieved from http://www.jstor.org/stable/3115131

O'Halloran, K. (2004). *Multimodal discourse analysis: systemic functional perspectives*. London, England: A&C Black.

Onacle v. Sundowner Offshore Services, Inc. 523 US 75 (1998). Retrieved from https:// www.oyez.org/cases/1997/96-568

Peoples, F. M. (2008). Street harassment in Cairo: a symptom of disintegrating social structures. *The African Anthropologist, 15*(1-2), 1-20. Retrieved from https://www.ajol.info/index.php/aa/article/viewFile/77244/67691

Phan, H., Koch, P., Maass, M., Mazur, R., McLoughlin, I., & Mertins, A. (2016). What makes audio event detection harder than classification? *Signal Processing Conference, 2017 25th European.* 2739-2743. Retrieved from http://ieeexplore.ieee.org/ abstract/document/8081709/

Phillips, T. (2015, January 06). Lego MMO development dogged by "dong detection" software. Retrieved from http://www.eurogamer.net/articles/2015-06-01-lego-mmo-development-dogged-by-dong-detection-software

Rainey, D. W., & Granito, V. (2010). Normative rules for trash talk among college athletes: An exploratory study. *Journal of Sport Behavior, 33*(3), 276-294. Retrieved from https://search.proquest.com/docview/744221182?pq-origsite=gscholar

Revi, R. (2015). Understanding obscenity and offensive humour: What's funny?. *The European Journal of Humour Research, 2*(3), 98-114. Retrieved from https://www.europeanjournalofhumour.org/index.php/ejhr/article/view/61

Rheingold, H. (2000). *The virtual community: homesteading on the electronic frontier*. Cambridge, MA: MIT Press.

Rotundo, M., Nguyen, D. H., & Sackett, P. R. (2001). A meta-analytic review of gender differences in perceptions of sexual harassment. *Journal of Applied Psychology, 86*(5), 914-922. Retrieved from http:// psycnet.apa.org/record/2001-18662-009

Schourup, L. C. (1985). *Common discourse particles in English conversation*. New York, NY: Garland.

Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. *Proc. of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 101-110. Retrieved from https://dl.acm.org/citation.cfm?id=2661935

Steinicke, F. (2016). *Being really virtual: immersive natives and the future of virtual reality*. Cham, Switzerland: Springer.

Stanford University. (2016, August 2). Article 1.7.1 sexual harassment. *Administrative Guide*. Retrieved from https://adminguide.stanford.edu/chapter-1/subchapter-7/policy-1-7-1

Shriram, K., & Schwartz, R. (2017). All are welcome: using VR ethnography to explore harassment behavior in immersive social virtual reality. *Virtual Reality, 2017 IEEE*. 225-226. Retrieved from http://ieeexplore.ieee.org/abstract/document/7892258/

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., & Weber, I. (2016). Analyzing the targets of hate in online social media. *Proc. of the Tenth International AAAI Conference on the Web and Social Media*. 687-690. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM16/paper/download/13147/12829

Srivastava, R., & Bhatia, M. P. S. (2017). Challenges with Sentiment Analysis of On-line Micro-texts. *International Journal of Intelligent Systems and Applications, 9*(7), 31-40. Retrieved from http://www.mecs-press.org/ijisa/ijisa-v9-n7/IJISA-V9-N7-4.pdf

Suler, J. (2004). The online disinhibition effect. *Cyberpsychology & Behavior*, *7*(3), 321-326. Retrieved from http://online.liebertpub.com/doi/abs/10.1089/1094931041291295

United States Equal Employment Opportunity Commission. (n.d.). Sexual harassment. Retrieved from: https://www.eeoc.gov/laws/types/sexual_harassment.cfm.

University of California Berkley. (n.d.) Code of conduct. Retrieved from http://sa.berkeley.edu/student-code-of-conduct

Vance v. Ball State University. 570 US (2013). Retrieved from https://www.oyez.org/cases/2012/11-556

Weber, A. (2009). *Manual on hate speech*. Strasbourg, France: Council of Europe. Retrieved from http://icm.sk/subory/Manual_on_hate_speech.pdf

Wodak, R (2011). Critical Discourse Analysis. In C. Seale, G. Gobo, J. F. Gubrium & D. Silverman (Ed.), *Qualitative Research Practices*. 186-201. London, England: Sage.

Wong, J. (2016, October 26). Sexual harassment in virtual reality feels all too real – 'it's creepy beyond creepy' *The Guardian.* Retrieved from https://www.theguardian.com/technology/2016/oct/26/virtual-reality-sexual-harassment-online-groping-quivr

Woska, W. J. (2014). The continuing development of the law on sexual harassment. In R. R. Sims, W. I. Sauser (Ed.), *Legal and Regulatory Issues in Human Resources Management*. 207-228. Charlotte, NC: Information Age Publishing.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., & Edwards, L. (2009). Detection of harassment on web 2.0. *Proc. of the Content Analysis in the WEB*, *2*, 1-7. Retrieved from https://s3.amazonaws.com/academia.edu.documents/

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network. *Proc. of 23rd International Conference on Computational Linguistics*. 2335-2344. Retrieved from http://www.nlpr.ia.ac.cn/cip/~liukang/liukangPageFile/camera_coling2014_final.pdf