# CLASSIFYING #METOO HASH-TAGGED TWEETS BY SEMANTICS TO UNDERSTAND THE EXTENT OF SEXUAL HARASSMENT

by
Claire Elise Hubacek

A thesis submitted to the faculty of The University of Mississippi in partial fulfillment of the requirements of the Sally McDonnell Barksdale Honors College.

Oxford
May 2018

Approved by

Advisor: Dr. Naeemul Hassan

Reader: Dr. Kirsten Dellinger

Reader: Dr. Dawn Wilkins

DEDICATION

This thesis is dedicated to all of the survivors of sexual assault and sexual harassment, regardless of age, gender identity, race, or sexual orientation. No matter how mild or severe you perceive your struggles to be, your feelings are valid, and this thesis is for you. In particular, this work is for all the survivors with the courage to take their experiences to social media and post "#MeToo", who are all much braver than I.

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Naeemul Hassan for his continued support and knowledge while guiding me throughout my thesis. I cannot overstate how thankful I am for the degree of patience and understanding demonstrated with me throughout this process, nor could I overstate the contribution Dr. Hassan brings to the computer science department here at Ole Miss.

In addition to my advisor, I would also like to thank my other readers: Dr. Kirsten Dellinger and Dr. Dawn Wilkins for their insight, questions, and flexibility. I am fortunate to have such accommodating and supportive readers.

My sincerest thanks also go to the computer science department, for accepting an art student with no background in computer science and helping me become the competent programmer that I am now. Thank you to the Sally McDonnell Barksdale Honors College faculty and staff for the opportunities and challenges given to me these past five years. I would also like to thank the Title IX office and Department of Violence Prevention for the accommodations, help, and support given to me during my own struggles throughout this past year. Thank you for empowering me to be able to finish my last year, and thank you to The University of Mississippi for "the best five or six" years of my life.

Thank you to everyone who participated in this project and took the time to read my research on sexual harassment, read my categorization rules, and assist me in manually categorizing a set of tweets: John Joseph Angel, Ainsley Ash, Kevin Carugati, Tim Dolan, Cami Duncan, Jason Hale, Lily Hassan, Gracie Hubacek, Karen Hubacek, Krish Lamba, Will Lewis, Mallory Loe, Ethan Luckett, Jake Wooley, Lina Ye, Dr. John Wiginton, and Dr. Debra Young. Your support and participation are appreciated beyond what I can express.

I would like to thank my family and friends who have supported me throughout this thesis as well as my undergraduate career. Reflecting upon on my transcript and resume as I prepare for graduation, I only see the contributions of my support system and how my achievements would not have been possible alone. Without your patience and support, I would not have finished. Lastly, I extend a special thank you to my pet rabbits who have continued to inspire and motivate me every day since I have loved them: Ellie, Ezra, and Buddy.

ABSTRACT

Classifying #MeToo Hash-tagged Tweets by Semantics to Understand the Extent of Sexual Harassment

(Under the direction of Naeemul Hassan)

This thesis contains a program that will process tweets from Twitter that use the hashtag "#MeToo" and categorize them by their relevance to the movement, their stance on the movement, and the type of sexual harassment expressed (if applicable). Being able to work with a narrowed set of tweets belonging to a specific category creates the capacity to do more in-depth research and analysis, exploring Twitter as a special platform for discussing these sensitive topics and showing that this online space for expressing personal experiences has delivered unprecedented potential avenues of study. This thesis also contains research into additional solutions towards addressing sexual harassment online, exploring the needs of society through the results to a questionnaire that was administered to university students asking for opinions on how sexual harassment is addressed on social media as well as through a literature review of current obstacles for victims.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

**CSV** Comma Separated Values

**EEOC** Equal Employment Opportunity Commission

**ML** Machine Learning

**NLP** Natural Language Processing

**NLTK** Natural Language Toolkit

**SVC** Support Vector Classifier

**SVM** Support Vector Machine

**TF-IDF** Term frequency–inverse document frequency

**UI** User Interface

# CHAPTER 1

# INTRODUCTION

> "If all women who have been sexually harassed or assaulted wrote 'Me too.' as a status, we might give people a sense of the magnitude of the problem." - Alyssa Milano

Beginning in October 2017, victims of sexual assault or harassment began using the hashtag "#MeToo" to illustrate the magnitude and prevalence of sexual assault and sexual harassment. Since its inception, the movement has expanded beyond its original scope and purpose, establishing itself as a global forum for debating and discussing the issue. The #MeToo movement's inherently public nature has created an avant-garde platform to vocalize experiences as victims and survivors of sexually based crimes, pushing the subject of sexual assault and harassment to accrue cultural significance and traction in academia. This thesis contains a basic classification implementation for examining tweets that use the #MeToo hashtag and classifying the type of sexual harassment described. Furthermore, this thesis distributed a survey that shows many ways sexual harassment could be addressed and mitigated online. Through research into existing journals, projects, and legal constructions surrounding sexual harassment, this thesis also details possible avenues through which this program might contain utility

## 1.1 Project Overview

This thesis uses Natural Language Processing (NLP) and machine learning algorithms to categorize tweets of a harassing nature. NLP is a domain of computer science that applies computer or machine intelligence to understand, process, and use language that has developed naturally in human use. NLP can deal with written, typed, or spoken language, but this thesis only concerns language as it has been typed and posted in tweets on the social media site Twitter.

Machine learning refers to the field of computer science in which programs and algorithms autonomously improve their decision-making abilities based on the initial data they use to "learn." The algorithms, by design, have not been programmed with the improvements; rather, they have been programmed to incorporate the data to improve themselves. Supervised and unsupervised machine learning are two different approaches to the ML process. Unsupervised machine learning does not have outcomes fed to the algorithm to "teach" the program how to respond to data and instead relies on techniques such as clustering to make inferences regarding the structure behind the data. Conversely, supervised machine learning uses a set of data with the desired output labeled by humans prior to the training process, and the algorithm is left to determine the mapping function to achieve an accurate generalization for any unseen input. This thesis uses a substantial dataset of human-labeled tweets (regarding the relevancy, stance, and category of harassment of the tweet) to train the machine learning algorithms employed in the classifier project.

### 1.1.1 Classification Model

A classification model generalizes predictions for unseen data based upon the mapping function it has derived from the labeled data it has already observed. For each sample, this classification model makes generalizations about the relevancy of the tweet to the #MeToo movement, about the stance of the tweet, and about the type of sexual harassment or assault experienced. A comprehensive definition of each class was researched and established to develop a reliable model for categorizing types of harassment. Approximately 5,000 tweets were manually hand-tagged by humans according to the classification rules to train the classifier model. Afterward, this model was applied to two algorithms with varying degrees of accuracy and success.

# CHAPTER 2

# PREPARATORY WORK

This chapter addresses the research and work performed before the development process with regards to designing the rules for the classification model as well data collection. This chapter describes sexual harassment as it has developed legally and contains a discussion of the literature regarding sexual harassment categorization. This research was fundamental to designing and applying an exhaustive, hermetic categorization topology to the classification model that avoided bias.

## 2.1 Legal History of Sexual Harassment in the United States

Despite steady, incremental progress, defining different types of sexual harassment is a task that remains incomplete primarily because "sexual harassment" didn't exist as a criminal offense until case law first sanctioned it in the 1970's. Since its implementation, the Title VII of the Civil Rights Act of 1964 makes it illegal to discriminate against potential or actual employees based upon gender (and Title IX later upholding the same philosophies within education). Throughout the 1970's several women used Title VII provisions to sue their employers for coercing, or attempting to coerce, them into sexual acts, and their successes within the courts enshrined sexual harassment as a criminal offense under these provisions. Since these cases, appropriately defining sexual harassment continues to be refined through improvements in research, laws, and court decisions and the nuances vary across state lines. As a controversial topic with many nuances and diverse perspectives, proper categorization of sexual harassment necessitates a comprehensive legal understanding alongside a psychological one.

The first Supreme Court case ruling in favor of the victim alleging sexual harassment was in the 1974 case *Barnes v. Train*, in which the Supreme Court found the harasser at fault for firing a female employee for refusing his sexual advances, although the term "sexual harassment" was not used yet at this time. A few years later, the Supreme Court upheld that this type of behavior from employers was a violation of Title VII; subsequently, the Equal Employment Opportunity Commission (EEOC) refined the rules to cover gender-based harassment as a form of discrimination. The first case to explicitly grant relief for sexual harassment under Title VII provisions was *Williams v. Saxbe* (1976). The plaintiff in the landmark case *Meritor Savings Bank v. Vinson* (1986) sued and won against her previous employer for forcing and coercing her into sexual acts, effectively establishing *quid pro quo* behaviors as a form of sexual harassment. This case also recognized that verbal remarks and questions of a *quid pro quo* nature were in violation of Title VII even if the victim suffered no tangible consequences because the offender's language itself creates a "hostile work environment" [Pierce, 1989]. Now, legally recognized instances of sexual harassment continue to fall under the domain of the EEOC and Title VII, and the accepted definitions categorize sexual harassment as either being actual or attempted *quid pro quo* acts or behaviors that contribute to a "hostile work environment" [York and Brookhouse, 1988]. Through a myriad of rule changes, public statements, and Supreme Court decisions, the Department of Education's Office for Civil Rights has upheld these same principles as they pertain to education, rallying behind the principle that sex-motivated violence or harassment to a student creates a hostile environment and encroaches upon their civil rights.

What behaviors qualify as sexual harassment on a criminal level are continuously revisited through new court decisions and research. Under U.S. law, *quid pro quo* sexual harassment, also referred to as sexual bribery, includes attempted and actual pursuits of a sexual nature against a person in a professional or academic environment when tangible benefits could be given or denied to the victim. For a long time, researchers overestimated the frequency of this type of harassment

compared to other forms because of its popularity in research. Historically, research appears to gravitate towards sexually explicit cases over work that is merely sex-related, and research into other domains of sexual harassment did not begin to pick up speed until the 90's [Fitzgerald et al., 1995]. Actions that contribute to a "hostile work environment" are the much more common category of sexual harassment. The rules do not ban less overtly sexual behaviors, such as offhand remarks, teasing or banter, and personal questions. However, if these acts occur with a sufficient degree of severity or frequency, they would then be considered as elements that create a hostile environment in work or school and therefore qualify as sexual harassment.

When evaluating behaviors that do not take place in an educational or professional setting, the actions themselves cannot be evaluated under Title VII or Title IX regulations because it lacks the context of equal rights with regards to having access to work and education. These instances of sexual harassment that don't occur in a professional setting must occur with such a degree of frequency or severity that they can be considered (with context) as falling under other, similar criminal statutes such as harassment, stalking, cyberstalking, or sexual assault. This legal technicality makes the proper evaluation of sexual harassment complex because the behavior in question, while it might be considered sexual harassment in a professional setting, does not have an equal opportunity regulation to dictate it as such when the action occurs between peers. Furthermore, the general public lacks sufficient knowledge regarding this topic, causing an extreme variance in individual perception of sexually harassing behaviors.

### 2.2 Literature Review

As the courts first began to address more and more claims of alleged sexual harassment, the need for a proper categorization grew to appropriately assess the degree and severity of the actions. From both a legal and social research perspective, the lack of consistency among categorization definitions caused problems when trying to compare and use instances or research from one context when evaluating another. Frank J. Till designed the first widely used standardization in 1980. In his work, Till defines five major categories that consist of generalized sexist remarks or behavior, inappropriate and offensive but essentially sanction-free sexual advances, solicitation, coercion, and sexual crimes [Till, 1980]. This categorization was the most frequently used throughout the 1980's, although many who used his categorization either rephrased the categories or consolidated them into three categories to suit their needs as appropriate.

In 1992, James E. Gruber, a legal consultant, made a significant contribution to categorizing sexual harassment by defining three overarching categories and 3-4 subtypes for each category, for a total of 11 distinct categories that are both exhaustive and mutually exclusive. Gruber's first category "verbal requests" encompasses verbal expressions of desire from one person to another, even though the goal or target expressed may be implied or vague. The subtypes include sexual bribery, sexual advances, relational advances, and subtle pressures/advances. Subtle pressures and advances generally refer to statements or questions that can be ambiguous and not directed towards the victim but are still inappropriate, such as accidentally "thinking out loud" or double entendres. These are all behaviors that are said directly to the victim with the intent of a sexual or an increased personal or social relationship goal even if the relationship sought isn't openly sexual.

The second category, "verbal comments," is defined by spoken or written language of a sexual nature that doesn't pursue any goal or target with regards to another individual. The subtypes include personal remarks, subjective objectification, and sexual categorical remarks. Sexist verbal remarks, rumors, inappropriate comments regarding someone's body or perceived level of attractiveness, inappropriate jokes or rumors, and bystander harassment are also classified here. These categories generally refer to statements of a "nonsolicitory" nature directed either to a woman, about a woman, or about women in general.

The third category, "nonverbal displays," includes sexual assault, sexual touching, sexual posturing, and the possession or display of inappropriate sexual materials. Actual or attempted coercion is sexual assault regardless of the context or history between the offender and victim. Pinching, grabbing, groping are examples of sexualized touching. Sexual posturing includes violations of personal space and attempts to make physical contact, including behaviors like making an obscene sexual

gesture with one's hands. The possession or display of sexual materials includes personal items regarding sexuality (such as tampons, pads, birth control, sex toys) in addition to pornographic materials or sexist materials. Excluding assault, describing or proving an instance of nonverbal sexual harassment can be difficult as it often lacks evidence, and individual interpretations vary significantly.

Altogether, there are 11 distinct types of sexual harassment, and these categories are both mutually exclusive and reflective of the EEOC's guidelines [Gruber, 1992]. Gruber originally wrote these categories based upon reviews of sexual harassment that only included female victims and male harassers, but their continued application today shows that these definitions are not gender-exclusive. His contribution to this field is significant, and these sexual harassment types have been used in an incredible number of research and court cases since their induction. Gruber's original topology of sexual harassment types and his original table containing descriptions of them have been reproduced in Appendix A.

These mutually exclusive categories establish a fluid progression in how certain behaviors can contribute to a hostile work environment, including guidelines for assessing minor behaviors in their proper context over time. A review of past, present, and future directions for improving gender and minority diversity in professional environments written in 2005 and published in *Sex Roles* refers to Gruber's work as a cornerstone in developing a comprehensive legal definition of sexual harassment [Murrell and James, 2002]. Gruber's topology continues to guide research today, and its continued use over time creates a standardized framework for assessing and comparing different research. For example, a study performed in 2005 to evaluate the effect of an obscene television show on individuals' perception of what constitutes sexual harassment used Gruber's categorization in their participant surveys to do so [Ferguson et al., 2005]. Despite the significance of his contribution, there is a neglected space of defining sexual harassment beyond the scope of Title VII and Title IX that Gruber's categories do not accommodate entirely. At the time he wrote this topology, Gruber worked as a legal consultant and was only concerned with defining behaviors that existed under the law (*i.e.*, behaviors that fell under *quid pro quo* and hostile work environment). Consequently, many researchers continue to consolidate these 11 categories and, in doing so, claim a lesser degree of specificity as a concession for incorporating the nuances of peer to peer harassment.

A prevalent issue with classifying sexual harassment lies in the varied perceptions of what is and isn't sexual harassment among observers [Studzińska, 2015]. The victim's perception is integral not just for legal categorization but also to understand the degree and severity of psychological harm inflicted on the victims and others exposed to the behavior. However, some concerns with relying too heavily on the victim's perception include the fear that the victim might be biased and in some rare cases, deliberately dishonest. Psychological research into coping with sexual harassment suggest that this is not necessarily the case. A 2015 study by Aparna Pathak on sexual harassment and coping behaviors synthesized many different research publications over the past few decades in her work. This synthesis notes that a 1997 study found that experiencing sexual harassment, whether or not the victim can identify the behavior as such, will still have adverse outcomes on the victim concerning health and distress [Schneider et al., 1997]. Additional research into false claims shows that the prevalence of false allegations of sexual assault comprises between 2% and 10% of cases, which is the same statistical prevalence as false reporting of non-sexual crimes [Lisak et al., 2010]. These types of research support the philosophy of giving the victim the benefit of the doubt when categorizing controversial instances.

Additionally, a lack of research exists in sexual harassment as it pertains to instances that don't occur within title VII or Title IX, particularly on the streets. Pathak's review notes that (as of 2015, the time this was written) a study performed in 2000 is one of the only known, peer-reviewed attempts at documenting the extent of unwanted sexual attention from strangers. MacMillan et al. found that over 80% of women experienced unwanted sexual attention (ex. catcalls) and just under 30% of women experienced confrontation of a sexual nature from strangers when walking in public areas [MacMillan et al., 2000]. Altogether, these caveats indicate that while Gruber's categorization might be the most appropriate tool for evaluating sexual harassment in a professional capacity because of its adherence to EEOC guidelines, it does not necessarily scale towards including social environments outside of a workplace or school.

Shortly after Gruber published his categorization, Fitzgerald et al. developed a more straightforward, consolidated categorization architecture to develop a better questionnaire for measuring sexual harassment. Their categorization revolved around three categories: unwanted sexual attention and gender harassment (hostile work environment) and sexual coercion (*quid pro quo*). This categorization was an attempt to distinguish between sexual harassment "as a legal concept and a psychological construct" to better accommodate how a victim might perceive or label a behavior. Accommodating this "gray space" in victim perception allows for more consistency among responses, which guided their research goal of developing a more scalable questionnaire for surveying the frequency of sexual harassment [Fitzgerald et al., 1995]. Similarly, in a 2008 publication, Chamberlain et al. deviated from Gruber's 11 types upon the basis that these legal-driven approaches "underscore diversity" and "suggest substantial variation with regard to intent and severity" from the victims' perceptions [Chamberlain et al., 2008]. For their purposes, Chamberlain et al. uses the following sexual harassment categories: patronizing (sexist but nonsexual remarks, gestures, or condescension), taunting (sexual gestures, physical displays, and overly personal comments and queries), and predatory (encompassing sexual solicitation, sexual promises or threats, touching, and forced contact). These choices echo the three categories defined by Fitzgerald et al. in 1995 and mimic the scope of the three categories. For this thesis, I have borrowed the inclusive language of the categories *patronizing* and *predatory* from Chamberlain et al. and the term *unwanted sexual attention* from Fitzgerald et al. for the third category, and otherwise agree with the definitions and use of these three domains for sexual harassment research.

### 2.3 Data

Dr. Hassan initially collected a large number of tweets using the hashtag #MeToo and then assembled them into a single file with the tweet's unique ID and original text contents. These two attributes were compiled into a Google excel sheet in the first two columns. The next three columns contained headers for labeling the relevance, stance, and type of sexual harassment described in the tweet. If the correct classification could not be determined from the text available, or if it was not applicable, the index was left blank. The file originally contained 10,000 samples.

| | | Classification Labels | | |
|---|---|---|---|---|
| | | **Relevant**<br>1. Related<br>2. Unrelated | **Stance**<br>1. Support<br>2. Against<br>3. Neutral | **Harassment Category**<br>1. Patronizing<br>2. Unwanted Sexual Attention<br>3. Predatory<br>4. Not Enough Context |
| **ID** | **Tweets** | | | |
| 1 | I was 13 #MeToo | 1 | 1 | 3 |
| 2 | metoo faked being a Native American... | 2 | | |
| 3 | This is exactly why I'm not for #MeToo | 1 | 2 | |

Table 2.1: Categorization Header

To create a reliable training set of data, a sufficiently large amount of tweets needed to be manually tagged by human readers. To avoid bias, I had volunteer participants, including family and friends alongside strangers, contribute to the labeled dataset in sets of 500. These voluntary participants were given a comprehensive explanation of sexual harassment both legally and psychologically, examples of properly categorized tweets, category definitions as defined in this thesis, as well as the categorization rules for dealing with anomalous tweets.

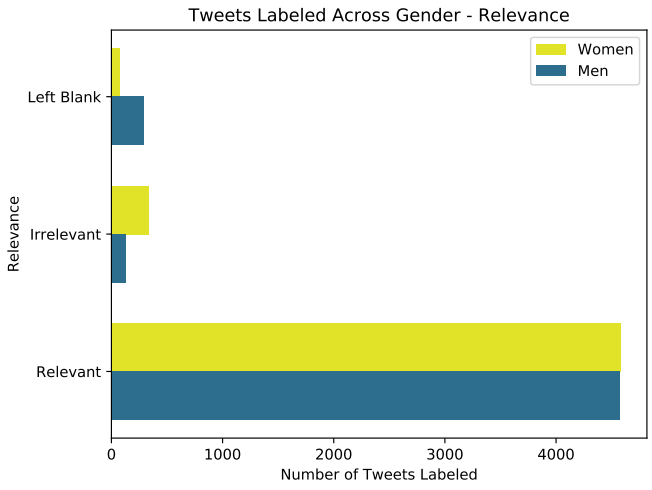
Participants were each provided 500 unlabeled tweets and asked to categorize them according to the rules provided in this thesis. One male and one female of a similar age and level of education shared each set of 500 tweets and categorized them without having access to any other responses

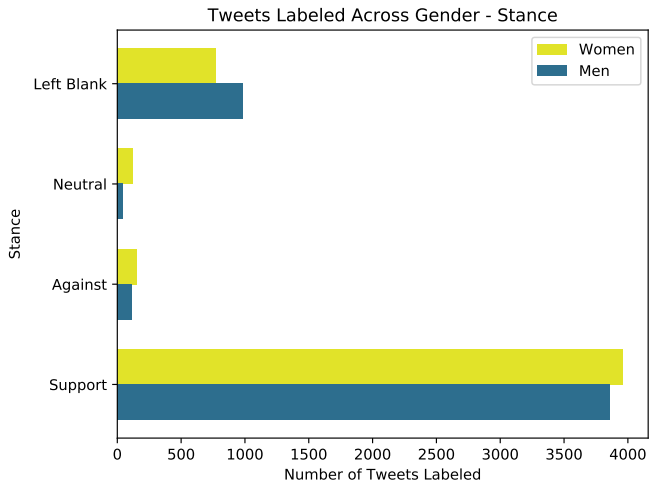
beyond the examples provided. Because research suggests that males and females assess sexual harassment differently, this process was performed to improve the integrity and consistency of the training set [Studzińska, 2015]. In total, 5,000 tweets were labeled by participants with each set of 500 being labeled by a male and female. Cumulatively, 10,000 categorizations were made by participants. Only the tweets that were categorized identically by all participants were considered in the training and testing sets of data. Table 2.2 contains the final results for the 5,000 tweets used in the training set. Answers that were invalid or left blank were discarded; thus the totals do not evaluate to an even 5,000.

| **Class Label** | **Female Classifications** | **Male Classifications** | **Overall Agreement** |
|---|---|---|---|
| Related | 4588 | 4576 | 4294 |
| Not Related | 340 | 132 | 76 |
| Support | 3961 | 3860 | 3382 |
| Against | 150 | 112 | 55 |
| Neutral | 119 | 44 | 12 |
| Patronizing | 96 | 60 | 16 |
| Unwanted Sexual Attention | 82 | 166 | 33 |
| Predatory | 372 | 370 | 219 |
| Not Enough Context | 2367 | 2161 | 1452 |

Table 2.2: Final Dataset Size

After receiving the final training set of all of the participants' labeling decisions, one label of each class existed with such a degree of prominence that it eliminated the accuracy of the lesser-represented classifications completely. After the hand labeling process, the "Related", "Support", and "Not Enough Context" categories vastly outweighed the others for each class. Figure 2.1 shows the distribution of the labeling by gender.



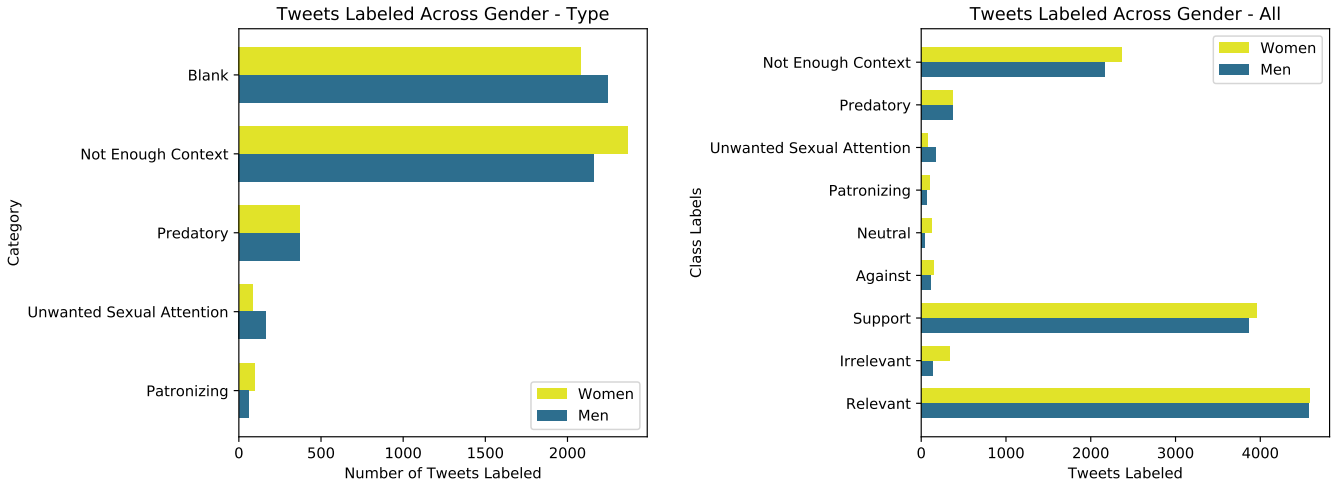



Figure 2.1: Male and Female Labeling Results

Because of the significant discrepancy, supplemental data were required. Consequently, more tweets from the remaining 5,000 were labeled by myself, and the categorizations from exclusively the minority class labels were taken and added to the existing training set. After using these additional tweets to flesh out the training set, some classifications were still close to 0. Further improvement on the accuracy was achieved through programmatic processes, such as by arbitrarily eliminating data from the more abundant categories and evaluating the subsequent accuracy. The quantity of samples added and removed as well as the corresponding improvements made to the classifier accuracy is discussed in Chapter 5.

Overall, women were less likely to leave tweets blank. Regarding relevance, women were more likely to label tweets as being irrelevant to the dataset and less likely to leave them blank. Across the stance category, Women labeled a stance for the tweet more often than men, who left them blank more often when they were uncertain. Regarding the type of sexual harassment, women were more likely to label a tweet as being patronizing; this is consistent with existing research that shows men are more likely to find sexist humor to be harmless and less likely to see it as a form of harassment [Studzińska, 2015]. While they were equally likely to label predatory behaviors as such, men were more likely to leave a tweet blank and more likely to label the tweet as being an instance of unwanted sexual attention.

Some of the participants who labeled the tweets preferred to have a phone call to discuss the guidelines while others preferred to read a document that explains the process. Some participants read through the examples and rules on the second and third sheets of the Google Sheets provided, and some did not. In particular, one participant mentioned afterward that she did not realize leaving tweets blank was an option, which may have made a significant difference in the comparisons. It is possible that men were more likely to leave answers blank because of an inherent hesitation to impose too much on what is viewed as a women's issue, which one participant stated upon completion. Not having this process streamlined may have contributed to some of the discrepancies among gender.

# CHAPTER 3

# CLASSIFICATION METHODOLOGY

For every tweet, the program will make three generalizations depending on the level of relevance, context, and detail. It will first determine whether or not the tweet using #MeToo is relevant to the movement. It will then determine the stance of the tweet regarding the movement. The last check is to determine the type of sexual harassment or assault described by the author if there is one.

Because each sample can only be assigned to one target label (*i.g.* a tweet cannot be both relevant and irrelevant at the same time) and each tweet maps to three classes (relevance, stance, and category of harassment), this model is a multi-class, multi-label classification. This section details the classification rules that were used to label the training set by hand before its use in training the ML algorithms.

## 3.1 Determining Relevancy

The majority of the tweets using the hashtag #MeToo are relevant to the movement in some way. However, sometimes they are not. Examples of irrelevant tweets are tweets written by bots where no meaningful determination can be made; unintelligible tweets; tweets where the entirety of the context can only be determined by following URL or image; tweets that use the hashtag #MeToo for the purpose of winning a giveaway or for increased visibility in a promotion; tweets where the hashtag #MeToo was used before Alyssa Milano popularized it and the tweet, therefore, refers to something else entirely; and tweets that are obviously misrepresenting or misusing the hashtag. Relevant tweets were assigned the number 1 and irrelevant tweets were assigned a the number 2. Table 3.1 illustrates examples of irrelevant tweets.

| **ID** | **Tweets** | **Relevant** | **Stance** | **Harassment Category** |
|---|---|---|---|---|
| 1 | https: soundcloud.comzay-hippy #askNiall #Bellator185 #MeToo #BadTimesToTellAJoke #FridayFeeling #RaiderNation #LouCity #bitcoin #WWEBuenosAires | 2 | | |
| 2 | #MeToo You too can achieve salvation by doing worship as per our Holly scriptures. To know more watch SADHANA TV 7:40 pm pic.twitter.com4wwhbEFZjM | 2 | | |
| 3 | @Der_Peemann check this crazy track out #hiphop you like it? “Yes really”? #MeToo WHAT? | 2 | | |

Table 3.1: Tweet Examples - Irrelevant

## 3.2 Determining Stance

Each tweet that is relevant to the #MeToo movement with enough context to ascertain a stance was assigned a 1, 2 or 3 when it was either in support of, against, or neutral to the movement accordingly. When manually tagging the data, tweets without enough context to determine the stance were left void. Tweets that are expressing a personal experience with sexual harassment or assault are considered to be inherently supportive tweets. Tweets that are expressing a supportive sentiment but not claiming victimhood are also considered supportive tweets. The distinction between tweets that are supportive of #MeToo in solidarity and supportive of #MeToo because of personal

experience is made in the third evaluation of the program. Tweets that are critical or against the movement are labeled as thus, and tweets asking a sincere question or making an unaffiliated remark regarding the movement are considered neutral. In the instances where the author made an antagonistic remark about a group of people in relation to the movement (e.g. a remark debasing all men), it was considered neutral to the movement if it only regarded a group of people not inherently linked to the movement. However, it was considered to be against the movement if the group of people could not be separated from the movement in the context of the tweet, such as a remark criticizing victims. Table 3.2 shows examples of the different stance categorizations.

| ID | Tweets | Relevant | Stance | Harassment Category |
|---|---|---|---|---|
| 1 | Calling In – Not Calling Out – Men ( #METOO BUT NOW WHAT?) Good men wondering what to do, this guide is for you. | 1 | | |
| 2 | The #MeToo Photo Going Viral on Instagram https: buff.ly2ysRfIe pic.twitter.comuyFONo00vA | 1 | | |
| 3 | Okay, first off, with all the #metoo stuff going around, what exactly are you classifying as sexual assault? | 1 | 3 | |
| 4 | the whole MeToo thing seems pointless tbh, like literally all 3.whatever billion women on this earth have experienced degrees of harrassment | 1 | 2 | |
| 5 | Just another trend started by idiots to get attention. If someone abused you,you should've slapped that cunt.Not cry on social media #MeToo | 1 | 2 | |
| 6 | This #metoo thing has me nearly in tears. It's not that I didn't know, but it's something else to confront the enormity of the problem. | 1 | 1 | |
| 7 | #MeToo | 1 | 1 | 4 |

Table 3.2: Tweet Examples - Stance

### 3.3 Classifying Types of Sexual Harassment

Many researchers who consolidated Gruber's categories did so while concluding that consolidation is ultimately the strongest approach to this particular problem, as it maintains consistency among the diversity in victims' perceptions. With regards to the integrity of the classification, Twitter's character limit prohibits users from providing an adequate context to classify every tweet with confidence. The personal bias, ignorance, or language choice of each author would result in improper categorization if the algorithm were to attempt to place each tweet within one of Gruber's eleven categories. Ultimately, three broad categories have been defined that consolidate Gruber's eleven types into each one: *patronizing*, *unwanted sexual attention*, and *predatory*.

#### 3.3.1 Category 1: Patronizing

The *patronizing* category aggregates the following categories from Gruber's work:

- Relational Advances
- Possession/display of sexual materials
- Subjective objectification
- Sexual Categorical Remarks

The *patronizing* category is designed to address behaviors that commonly fall into the "gray space" of sexual harassment. Comprehensively, this category is comprised of generally sexist remarks, gender-motivated harassment that is not necessarily pursuing a personal relationship with the recipient, nonverbal displays of harassment that are sexual, and non-sexual behaviors and remarks that the victim interprets as being sexual. Because many victims experience sexual harassment in contexts that are not protected from sexual harassment under Title VII or Title IX, there is a lot of controversy among victims' perception of these behaviors when they occur in a social setting. When deciding on a tweet's label, the tweet is considered *patronizing* if a neutral, unaffiliated third party were to witnesses the behavior and reasonably deem the behavior as potentially being non-sexual, yet the recipient still perceives it as thus. Additionally, this requires that the remark or behavior not contain an apparent sexual goal with the victim, which is often confusing when the remark is sexual but with the intent of degradation, not personal interest. Table 3.3 shows examples of the categorization of tweets describing patronizing behavior.

| ID | Tweets | Relevant | Stance | Harassment Category |
|---|---|---|---|---|
| 1 | Simply walking to class in normal, baggy, purposely-unattractive clothes still somehow warranted catcalls and unsolicited comments. #metoo | 1 | 1 | 1 |
| 2 | #MeToo When I was 17 my boss screamed @me in front of a store full of customers what's ur problem? R u on ur period or something? | 1 | 1 | 1 |
| 3 | #MeToo Never have i been treated with so much disdain and lack of respect due to my sex than in the past 3 years working in a record store | 1 | 1 | 1 |
| 4 | because having big boobs means "all the boys will like you" #MeToo | 1 | 1 | 1 |
| 5 | Men, don't use deterogating\belittling\demeaning words towards women or call men female words - you make it seem women are worth less #MeToo | 1 | 1 | 1 |

Table 3.3: Tweet Examples - Patronizing

Behaviors that are subject to common controversies, such as crossing the boundary between flirting and harassment, are likely to be categorized as *patronizing* because they are not consistently and objectively interpreted as having the intention of pursuing a sexual relationship with the victim. More examples of patronizing sexual harassment include but are not limited to sexist comments; obscene gestures or drawings about the victim; catcalling or ambiguously sexual behaviors and comments; teasing; banter; jokes; inappropriate comments regarding the victim's body (ex. weight or perceived level of attractiveness); and other behaviors that would contribute to a hostile work environment in a legal evaluation.

Some behaviors are ambiguous and therefore evaluated by the effect on the victim. Relational advances, such as repeated contact of a non-sexual nature, could reasonably make a victim fear that the advances in question might be sexually motivated and therefore cause them to interpret the behavior as harmful. In some cases, the situation reveals in hindsight that the offender had a sexually motivated goal with regards to the victim, but it could not be objectively determined at the time the behavior occurred. While these behaviors ought to be instances of *unwanted sexual attention*, an unaffiliated third party might witness the behavior without context and objectively think otherwise; therefore, these behaviors should be classified as *patronizing* behaviors instead as it maintains the integrity of the claim and the accusation.

Gruber's category of "sexual categorical remarks" (possession or display of sexual materials and other sexist behaviors and comments) belongs in this category because of the discrepancy that exists between issues that occur in professional versus casual environments. Subjective objectification,

which includes remarks made about a victim whether or not she is present (ex. rumors), is also evaluated as being within the patronizing category for the same purpose of accommodating the legal constraints of sexual harassment. When evaluating contexts that do not fall underneath Title VII or Title IX, these behaviors are not always present to a degree of severity or frequency that they could constitute a different criminal offense (such as stalking, cyberstalking, and harassment). However, the same behavior would easily qualify as sexual harassment in a professional setting because of the behavior's contribution to a hostile work environment, and therefore it is classified here. Consistency among this legal discrepancy is resolved by classifying these types of behaviors that are not explicitly pursuing a sexual goal or target as *patronizing* behaviors instead.

#### 3.3.2 Category 2: Unwanted Sexual Attention

The *unwanted sexual attention* category aggregates the following categories from Gruber's work:

- Sexual advances
- Subtle pressures/advances
- Personal remarks
- Sexual posturing

The category of *unwanted sexual attention* includes any behavior, language, questions, or comments of an explicitly sexual or romantic nature. These comments are an easy classification to make when they take place in a professional environment because they directly follow traditional legal classifications when evaluating a hostile work environment. When evaluating social and casual environments, the behaviors become more complicated to categorize because the harasser legally has the room to act within a reasonable degree of respectful, personal interest and flirtation with the victim. Table 3.4 illustrates examples of tweets describing this type of sexual harassment.

| ID | Tweets | Relevant | Stance | Harassment Category |
|---|---|---|---|---|
| 1 | "Sleep in my bed I'll take the floor" "You're cold, let me in with you" "forget your boyfriend" "Come on I could make you so happy" #MeToo | 1 | 1 | 2 |
| 2 | Three diff. men,same day, same line "My wife & I are getting a divorce so..." Military women have experienced harassment since day 1 #MeToo | 1 | 1 | 2 |
| 3 | Was 18. New apt, new unlisted ph #. Applied 4 job. Mgr starts sexual, obscene calls 2 me. I called the business. Calls stopped. #MeToo | 1 | 1 | 2 |
| 4 | #MeToo I personally witnessed gay men shocked to the core when mildly harassed on the streets. I almost laughed. This is women's reality. | 1 | 1 | 2 |
| 5 | #MeToo . On my bus just now. Men twice my size sitting at the back of a bus proclaiming what they'd do to my body given the chance. | 1 | 1 | 2 |

Table 3.4: Tweet Examples - Unwanted Sexual Attentions

Regarding instances where the perpetrator claims to be merely flirting in such an environment, the comments or questions made would fall under the category of *unwanted sexual attention* if the victim has communicated their lack of interest or if the comments or questions were egregiously sexual. If the alleged harasser has a reasonable defense for claiming their statements were innocent, and a neutral observer would agree, the incident would instead be categorized within the previous category as *patronizing* instead.

In general, any form of non-consensual physical contact is *predatory* excluding socially and culturally acceptable forms of physical contact, such as shaking hands or using hugs as a greeting. However, the victim in some cases still interprets these acceptable forms of contact as harassment. Sometimes, this is because the socially acceptable contact is coming from someone who has previously exhibited sexual misconduct or another valid reason that could make the victim uncomfortable, and the sexually harassing nature of the physical contact is only evident with context over time. To categorize these behaviors accurately, the standard of evaluation is, again, considering how an objective third party would interpret the scene had they walked in to witness the behavior. If the neutral party would reasonably interpret the physical contact as socially appropriate, yet the victim maintains that the contact was sexually motivated, the harasser's behavior is classified as *unwanted sexual attention*. If the neutral third party would reasonably interpret the behavior as strange, unusual, or inappropriate, it would be categorized as *predatory* instead.

Bystander harassment is also considered *unwanted sexual attention* within this thesis. According to Gruber's classification, a victim of bystander harassment (an individual witnessing harassment that happens to another person) would fall within his category of sexual categorical remarks, which in this thesis has been aggregated into *patronizing* behavior. However, for this thesis's applied purpose of categorization, it is more appropriate to classify bystander harassment as a form of *unwanted sexual attention*. Gruber's topology was developed regarding legally upheld forms of sexual harassment and consequently does not consider some impacts of sexual assault as opposed to sexual harassment. Gruber's description of bystander harassment does not address witnessing an instance of sexual assault, possibly because the likelihood of witnessing sexual assault is minimal in general and even more so with regards to professional environments, that it was not considered in at the time of his original publication. However, because this thesis needs to consider victims who witnessed an assault, bystander harassment (inclusive of both harassment and assault) is classified as being *unwanted sexual attention* as opposed to *patronizing*. This method of considering bystander harassment is the only significant philosophical or logical deviation from Gruber's classification rules.

Overall, the category of unwanted sexual attention is comprised of any behavior that is explicitly sexual when the victim has expressed their lack of interest. It also includes bystander harassment, any explicitly sexual behavior that occurred within a workplace or academic environment, any language or interaction that is beyond what is culturally accepted as appropriate flirting, and forms of conventionally accepted physical contact that the victim has claimed to be sexual.

#### 3.3.3 Category 3: Predatory

The *predatory* category aggregates the following categories from Gruber's work:

- Sexual bribery
- Sexual assault
- Sexual touching

*Predatory* behavior includes all forms of non-consensual physical contact, excluding the socially acceptable forms of contact addressed within the previous category. *Predatory* behaviors also include attempted or successful rape, sexual assault and battery, and *quid pro quo* arrangements. Table 3.5 demonstrates examples of tweets that fall under the *predatory* category.

Some nuances regarding authority exist within the *predatory* behavior category. If the harasser is exhibiting behavior classified as *unwanted sexual attention* but exists in a position of authority over the victim (ex. boss or teacher), the behavior should be labeled as *predatory*. This is due to the nature of the relationship between the harasser and victim, similar to the way statutory rape is regarded under the law. Sending sexually explicit photos to someone without their permission or consent is also *predatory* behavior. While not every instance of a victim being drugged results in actual or attempted assault or rape, being drugged should be considered *predatory* if the victim expresses the fear of assault or rape, particularly if bystander intervention is the reason the attack was unsuccessful.

Any and all forms of sexual comments, remarks, questions, and inappropriate physical contact between an adult and a minor should fall into this category. Even if the behavior would be considered *patronizing* or *unwanted sexual attention* if the two parties were of legal age, it is predatory by nature of the age discrepancy; however, if the behavior is in fact *patronizing* or *unwanted sexual attention* and the action occurs from one minor to another minor, it is not inherently *predatory* (*e.g.* middle school children making sexually degrading remarks). Aside from these specifications, this type of behavior is generally any physical violation (actual or attempted) between the perpetrator and victim. Table 3.5 contains examples of tweets from this category.

| **ID** | **Tweets** | **Relevant** | **Stance** | **Harassment Category** |
|---|---|---|---|---|
| 1 | Or the time I woke up to a friend assaulting me when I was drunk but never said anything because he made me think it was my fault? #MeToo | 1 | 1 | 3 |
| 2 | Latest incident breasts grabbed from behind at a concert. Man laughed at me when I spun around and hit! #MeToo thank you @Alyssa_Milano | 1 | 1 | 3 |
| 3 | I have had my ass slapped at two different jobs by two different men. #metoo | 1 | 1 | 3 |
| 4 | I was 14. Mom protected him, shamed me, kept me from help. Took me into my 20's to figure out I wasn't to blame. You aren't either. #MeToo | 1 | 1 | 3 |
| 5 | Denied an A in HS Chemistry at age 16 for refusing to go into the storage room with my pedophile Chemistry teacher. #metoo #manyexamples | 1 | 1 | 3 |

Table 3.5: Tweet Examples - Predatory

#### 3.3.4 Category 4: Not Enough Context

The majority of tweets that use #MeToo to express an experience do not describe it with a compelling level of detail. Many users do not describe their experience at all and merely attest their victimhood by writing "#MeToo" by itself. Anything that is claiming to have experienced sexual assault or sexual harassment but does not have a significant enough level of detail to determine the type of harassment falls into this category. Tweets that were classified as having a supportive stance but did not claim victimhood are left blank in this evaluation. Examples of tweets that claim victimhood but without enough detail for a categorization decision are shown in Table 3.5.

| ID | Tweets | Relevant | Stance | Harassment Category |
|---|---|---|---|---|
| 1 | #MeToo | 1 | 1 | 4 |
| 2 | I felt so ashamed when I understand what has happen. I thought it was my fault bc I never said anything. It's NOT #MeToo | 1 | 1 | 4 |
| 3 | #MeToo I could horrify you with the details, trust me, I see them everyday, but I prefer to show you how strong I've become in spite of it. | 1 | 1 | 4 |
| 4 | Talking to him still makes me feel furious and i want to throw up but im trying to forget. I need help #MeToo | 1 | 1 | 4 |
| 5 | Mine was the son of my mom's best friend. 20 years later, I still don't know what happened. #MeToo | 1 | 1 | 4 |

Table 3.6: Tweet Examples - Not Enough Context

Assembling Gruber's 11 types of sexual harassment into the three categories *patronizing*, *unwanted sexual attention*, and *predatory* successfully accommodates the ambiguous and controversial areas created by each Twitter user's interpretation of the behavior that they experienced and described. As these users' individual accounts of their experiences have not necessarily been reviewed (such as by the police or an HR department), this is the most philosophically ideal approach to this problem. It also avoids claiming a degree of accuracy that cannot be determined with confidence. The *not enough context* label is only for tweets where victimhood is clear; otherwise, it is left blank.

### 3.4 Categorization Rules

Some rules exist to accommodate the parameters of the algorithm. This list of categorization instructions covers all labeling decisions that may not be intuitive.

1. Remove tweets that are not written in English from the dataset.
2. Ignore URLs entirely, even if the rest of the context could be retrieved at that destination. Do not parse and consider words within the URL.
3. Do not consider words that are contained within username mentions beginning with the @ symbol.
4. Consider words that are contained within other hashtags.
5. Context that cannot be confidently determined because of sarcasm should be left blank.
6. Advocacy on behalf of a friend is considered to be only supportive and not a personal experience if the author did not personally witness the assault or harassment.
7. If the tweet only says "#MeToo", the author should be given the benefit of the doubt and assume that they have indeed experienced a form of sexual harassment or assault. The appropriate label is *not enough context*.
8. Tweets that imply a personal experience but do not explicitly claim one with language are categorized as having a supportive stance but left blank when determining the type of harassment. The only exceptions are tweets that contain only "#MeToo."
9. If the categorization is difficult to determine, the decision should be made by giving the benefit of the doubt to the author of the tweet. If the victim is claiming to have experienced something or claiming to have been affected significantly by an event, we will assume that is true.

# CHAPTER 4

# IMPLEMENTATION

## 4.1 Design and Development

This section discusses design considerations and key development decisions.

### 4.1.1 Classifier Design Objectives

The objective of this program is to classify tweets that use the #MeToo hashtag and return a categorization that is as accurate as possible. The harassment classification model and the ground truth dataset help the program ascertain three different classifications regarding each tweet. This thesis also contains an evaluation of the classifier model. The last requirement is to build a UI such that anyone can upload their tweet sample and, if appropriately formatted, receive the categorization results of the classifier or the results of an individual tweet.

### 4.1.2 Data Processing

Initially, all of the tweets in the training set were labeled within Google Sheets so that participants could easily access them. These were exported as individual XLSX excel sheets where the formatting was then removed. Then, the datasets were consolidated, and the complete set of labeled samples were saved as a CSV file. In Python, this data is loaded into a DataFrame object from which point the male and female responses are compared. Only the tweets that have been labeled identically by both parties are saved and used to train the algorithm. Because of the discrepancy between male and female responses from participants, I labeled more tweets to reach a dataset size substantial enough to even out the representation of each category. The final dataset is stored and used as a DataFrame object from Python's pandas library for the text to be pre-processed.

The pre-processing phase involves the translation of the words from the documents into a more normalized form so that associating them with their labels is more consistent. This process includes converting the text to lowercase; removing stop words (such as "he", "is", "at", "which", and "on"); removing URLs and links; removing username mentions that begin with the @ symbol; removing characters that are invalid in UTF-8 encoding; and removing other punctuation.

This phase also separates the tweets into tokens that ultimately create the vocabulary. An important stage of this phase is TFIDF, or term frequency-inverse document frequency. This process uses the term frequency of a feature within a document to evaluate how important a word is by considering how many documents that feature appears in out of the total collection of documents. The feature and its associated weight are called vectors. The TFIDF process is integral to having a successful classifier because it ensures words commonly used in dialogue without significant meaning, or unusual words and mis-spellings by one author, do not skew the data. Python's scikit-learn library provides functions to perform these translations and processes automatically. Stemming and lemmatization were not performed at this stage, but were tested later on.

Because of the way the tweets were mined, the majority of the URL's were broken up into parts. Because of this, complete removal of the URL's was not always possible. Additionally, the tweet's context and meaning were often contained within a picture or URL, the content of which cannot be included in a strict text analysis [Juditha and Kominfo, 2015]. Consequently, random strings of numbers and letters are frequently included in the vocabulary.

### 4.1.3 Architecture

The supervised machine learning process requires the raw data to be formatted and labeled by humans to form a reliable training set. The original labeled set of data is broken up into two sets

of roughly 20% (test set) and 80% (train set) of the original. The training set is used alongside the classification rules and machine-learning algorithm employed to develop the model. The test set is used twice: once with the classification removed to which the model is applied, and then again with the labels maintained in order to compare the accuracy of the classifications made on those samples.

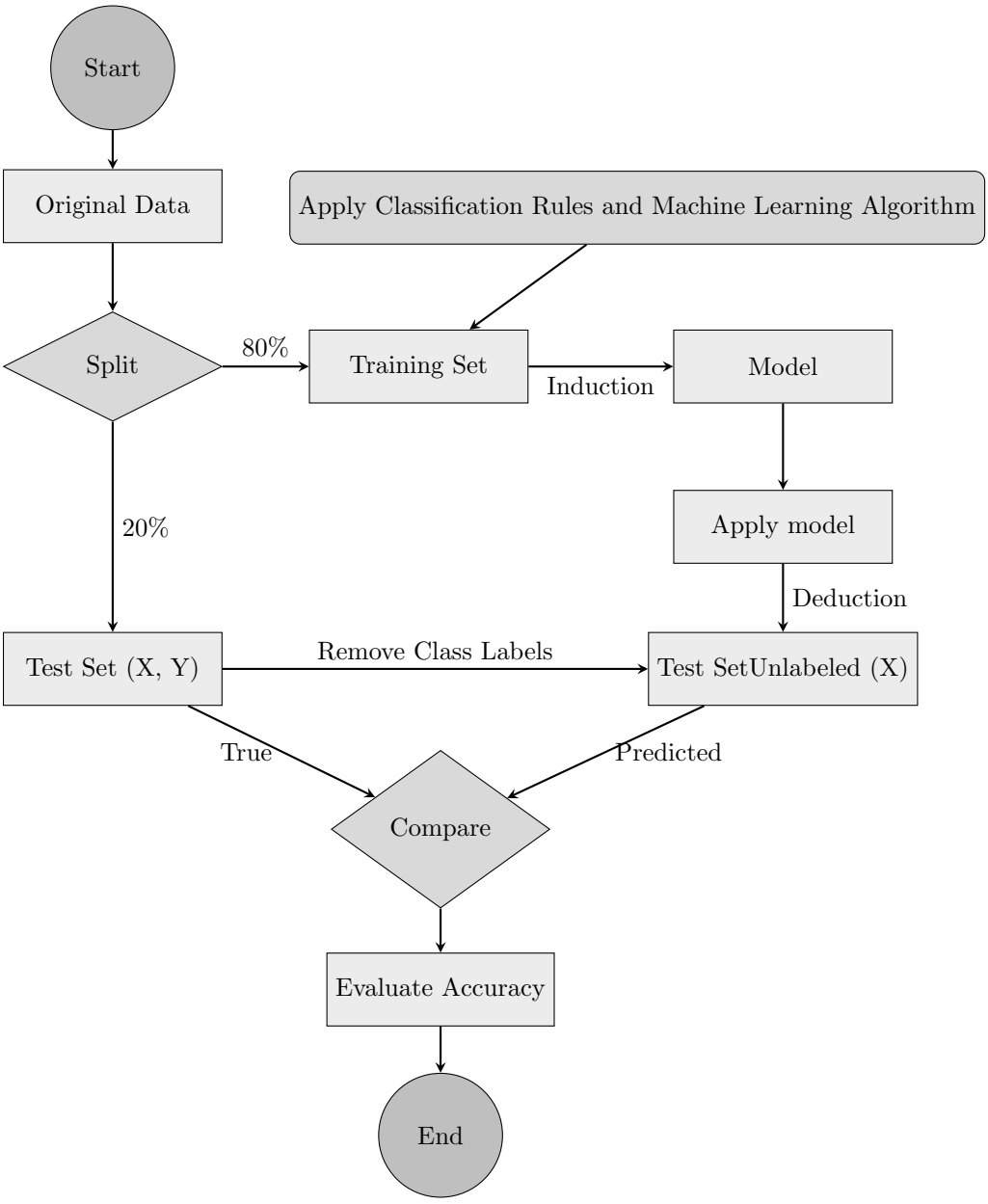



Figure 4.1: Supervised Machine Learning Architecture

This thesis compares the accuracy of Naive-Bayes and Support Vector Machine (SVM), two different but very common algorithms used in classification problems.

#### 4.1.4 Naive Bayes

The Naive Bayes classification technique, named after the Bayes' Theorem regarding conditional probability, is termed as such because it assumes an independent relationship between every feature in the text document. The scikit-learn library has three different implementations of the Naive Bayes model. Because the feature vectors in this training set are not likely to have a normal distribution, nor are they binary, the Gaussian and Bernoulli implementations were not appropriate choices. Multinomial Naive Bayes models are very commonly used for NLP tasks, and they perform very well for the level of effort needed to implement them successfully.

Bayes' Theorem is used to discern the probability that an event will occur based on the knowledge of conditions that have previously led to the event. The theorem's formula permits updates upon being provided new evidence, and its relation to conditional probabilities allows for explanations of counter-intuitive results. This centering on conditional probability lends itself well to problems involving natural language.

Because natural (human) language is not spoken with a series of unrelated words, this assumption of independence is considered a "naive" approach; however, Naive Bayes classifiers achieve a high level of accuracy and are very commonly used for classification. Naive Bayes classifiers do not require as much training data, and the simplicity of their implementation can make it the most appropriate implementation choice for small NLP tasks such as spam filtering, sentiment analysis, and others, the extent of which is dependent upon the training data and is sensitive to parameter tuning.

#### 4.1.5 Support Vector Machine (SVM)

SVM's methodology finds a line or hyper-plane in multidimensional space that makes classifications based on which side of the line the sample falls on. In dealing with only one plane, SVM represents the features as points in space with as much distance between them as possible and attempts to draw a line through them that will separate the data points into two classes (one for either side of the line). When plotting a new sample, its classification is determined by which side of the separating line the sample falls on.

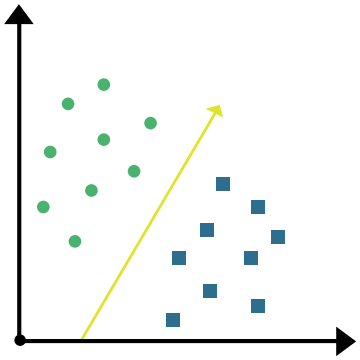


Figure 4.2: SVM Separation Line in One Dimension

An essential characteristic of the SVM classification model is the margin or the distance between the data points and the separation line. The best classifiers will have a significant margin on either side of the line between the line and the points of data. The larger the margin is on both sides of the line, the higher the accuracy will be when predicting values that have not yet been seen by the classifier.

Because the points of data in the space are unlikely to ever fall in space in such a way that an even line can be drawn between them, additional planes are added that allow "lines" to be drawn through other dimensions until an accurate division can be made to separate the points. Functions, called kernels, are used to map the data to a new plane in space with more dimensions so that the

data is separable, and the resulting separation line is called the hyper-plane. Once the base accuracy has been achieved, tuning parameters exist for optimization.

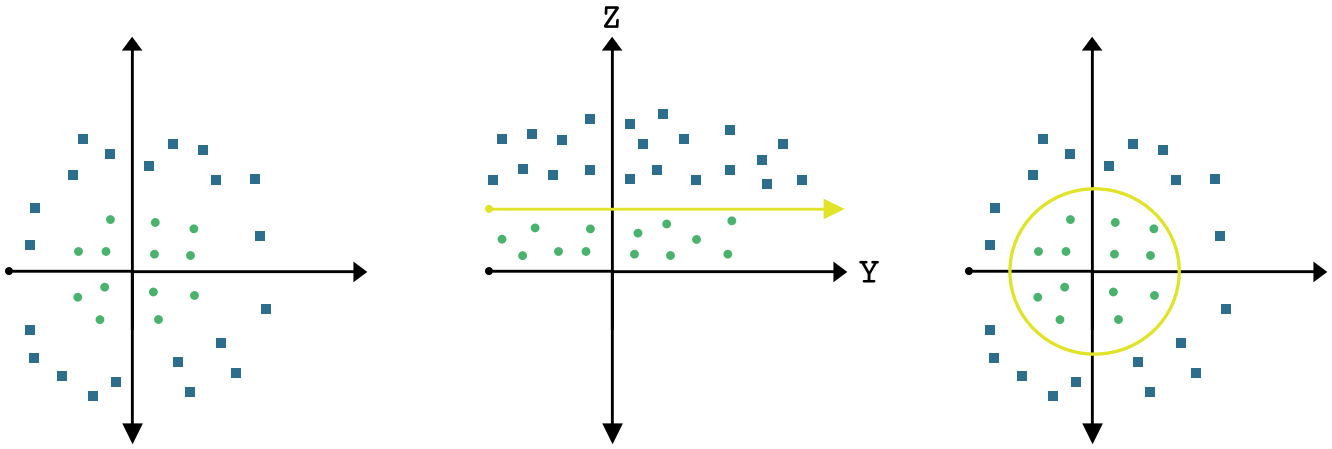



Figure 4.3: SVM Separation Line in Multiple Dimensions

Formulas employing linear algebra are used to incorporate the supporting feature vectors into the mathematical mapping of new, unseen data to its classification. Depending on the distribution of the data set used in training, different formulas may perform better than others. For instance, scikit-learn's implementation of SVM allows developers to pick between these formulas by changing the parameter named kernel. The different kernel options perform better depending on the distribution, bias, variance, and more. Performance evaluations suggest that choosing a kernel should be done through an automated process comparing the results between trials of different kernel usage to avoid selection bias [Cawley and Talbot, 2010].

However, there is also an option to implement a Linear SVC as opposed to conventional SVM with a linear kernel. Linear SVC employs a linear kernel by default and additionally takes an inherently different approach to classification. SVM employs a "one-vs-one" (OVO) approach, such that for each sample the algorithm builds several classifiers, two at a time. It creates a binary classifier for each pair of classes and compares them. If $n$ is the number of classes for the sample, then the number of classifiers produced will be $n$ * ($n$ - 1) / 2. When applying an OVO methodology to an unseen sample, a vector of individual classifications is produced, and the class predicted for the unseen sample will be the most frequent class from the results of each pair of classes. Alternatively, the "one-vs-the-rest" approach will create a classifier for each class (except for instances of two classes, for which one binary classifier is created). In general, these two produces similar results.

In scikit-learn's implementation of these two classifiers, Linear SVC uses a different loss function that can be changed from a squared hinge loss to a hinge loss, which is the default in SVM. SVM can also be implemented using a linear kernel. The other parameter to change to make these two implementations resemble each other is intercept-scaling; however, Linear SVC applies inherent regularization to the intercept scaling, and therefore they can only be exactly equal if there is no bias present in the problem to which they are being applied. That is not the case in this problem, and LinearSVC was used due to its higher accuracy.

A couple parameters give control over how much to let the distribution of the data influence where the separation line is drawn. One of those parameters is regularization, termed as C in scikit-learn's implementation. When separating out the classes in space, the hyper-plane employed can either prioritize the size of the margin between points or prioritize classifying more points accurately. The larger the value of C, the better the hyper-plane will do at classifying all points

correctly. Regularization helps control the sensitivity to outliers.

Similarly, the gamma parameter controls how much the points of data lying furthest from the hyper-plane influence where the separation line ultimately is placed. A low gamma considers points far from the line, and a high gamma restricts the points considered to those that fall closest to the line. However, LinearSVC does not have a parameter for gamma; therefore, only C was tuned in this implementation.

SVM classifiers continue to be a useful model for text classification because they can deliver accurate results even when dealing with a large quantity of features (which is inherently associated with natural language), allowing the model to generalize well. Additionally, the parameter tuning can be automated, which eliminates human error within that process.

#### 4.1.6 Other Classifiers

Naive Bayes and SVM are two of many classification algorithms. Other algorithms such as Decision Trees, K Nearest Neighbors, LDA topic clustering, NGrams, Neural Networks, and more are used for classification and clustering problems. Speaking generally, a single best algorithm does not exist because their performances depend on the data that they are being used with; for example, some algorithms perform better with numerical data, and some are better for text classification problems.

Frequently researchers will compare the relative accuracy of different algorithms on a single set of data to analyze performance. A 2013 comparison between six classification algorithms (Naive Bayes, Support Vector Machine, NGrams, K-Nearest Neighbor, Genetic Algorithm, and Back-propogation Network) found that NGrams algorithm is consistently faster while still producing a high degree of precision, with SVM and Naive Bayes producing higher precision in some trials [Ramasundaram and Victor, 2013]. A 2017 study used the classification of Amazon product reviews to compare the effectiveness of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression methods for multi-class text categorization. This study found that Logistic Regression had the highest accuracy and Decision Tree had the lowest, while the remaining three had a statistically insignificant difference in their accuracy. This study also found that increasing the training data set from 5,000 to 75,000 did not yield a significant improvement in accuracy for Naive Bayes, SVM or Random Forest classifiers and that the accuracy is related more to the n-gram properties than to the quantity of the training data [Pranckevicius and Marcinkevicius, ].

The best way to achieve high accuracy for a classifier is to compare the success of different algorithms and choose the one that is performing better for the given data set. This thesis only compares Naive Bayes and SVM because of their overall reliability and ease of implementation.

### 4.2 Testing and Integration

The accuracy of the classifier can be evaluated by reserving a percentage of the labeled data, using the classifier to categorize it, and checking those results against what had been previously labeled. By manually categorizing a sufficiently large data set, it can be determined with a high degree of accuracy how successful the algorithm and project will be. Figure 4.4 shows the algorithmic flow in optimizing and testing the accuracy of a model.

The model is repeatedly tweaked and optimized until a satisfactory level of accuracy has been achieved. In the training phase, the dataset is split into a train and test set. A machine learning algorithm is applied to the training set with the new vocabulary established during the pre-processing phase, and the test set without labels is used to evaluate the accuracy of that model once compared to the testing set with labels. If the desired accuracy is not achieved, optimization techniques are implemented. The optimization that made a significant difference in this implementation are discussed in Chapter 5.

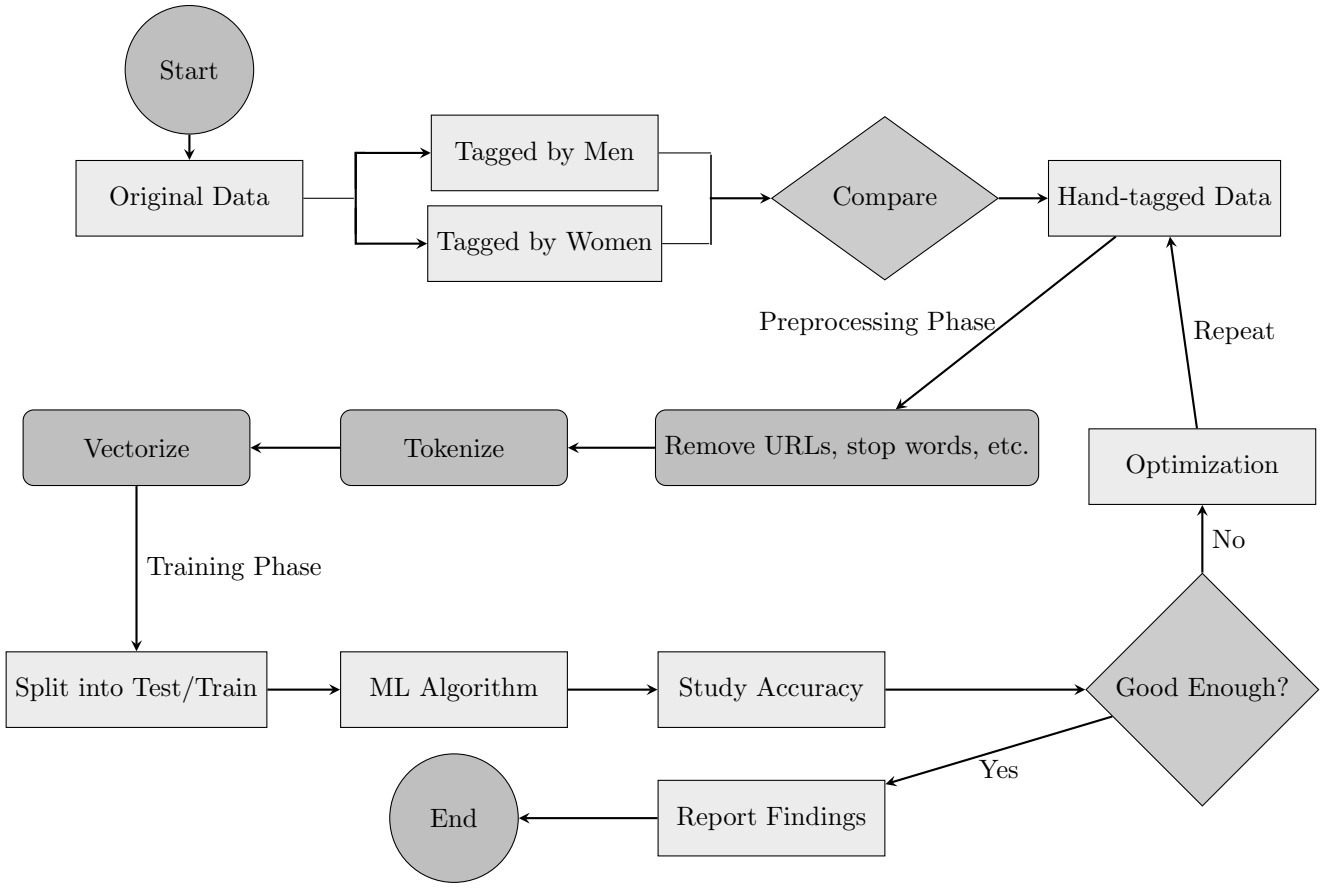



Figure 4.4: Testing and Integration Flow

### 4.3 User Interface

There are three ways to apply the model, the first being through the traditional use of the command line. If the user has installed Python and all of the libraries implemented in this thesis, they can process and classify a set of tweets on their local machine. To do so, the user must have a CSV file with the set of tweets they wish to classify formatted so that the first column contains the tweet's unique ID and the second column contains the original text. The program will then output a new file with the predicted categorizations in the third, fourth, and fifth columns. To avoid forcing users to follow the cumbersome process of installing all of the implemented Python libraries, an online UI was built that allows the user to use this program remotely. Uploading a CSV file online and displaying the results within a web UI is the second application of the model. Alternatively, the third application allows a user to enter one tweet at a time into a text box and see the classification results immediately. This UI is also a testing feature so that users can classify an individual tweet and see the results one at a time. Figure X and Y show the web UI.

### 4.4 Language and Libraries

This project was written in the Python3 programming language, chosen for its flexibility and the high quantity of libraries available to assist in natural language processing and supervised machine learning. Python and R are the two most competitive languages in the data science industry. Python was ultimately the most appropriate choice for this thesis because of its intuitiveness, lower learning curve, and more straightforward integration with web interfaces.

The library pandas was used to load the CSV file's data into DataFrame objects, which is a two-dimensional data structure that can have row and column labels as well as different data types.

Functionally, DataFrame objects resemble a standard database and can be queried and manipulated reliably. When used in conjunction with other libraries, very powerful manipulation can be done relatively easily.

The library Natural Language Toolkit (NLTK) is the most commonly used platform for Python programmers working on NLP related tasks. The NLTK contains a wide range of functionality such as lexical analysis and tagging parts of speech. The library SpaCy, which is commonly used in software development due to its superior speed and efficiency, was considered but ultimately not needed for this thesis. The Flask library was used to implement the web UI due to its simplicity in setting up a web server and support for Python3.

Scikit-Learn is the most integral library used in this project due to its existing implementations of classification algorithms and ability to use DataFrame objects fluidly. Once the training and testing sets are formatted properly as DataFrame objects, they can be passed to various algorithms through scikit-learn's API and the results easily determined. Because of this ease, different classification algorithms were employed and their accuracy compared.

# CHAPTER 5

# RESULTS

## 5.1 Accuracy

Different classification algorithms were used to compare the relative accuracy of the classifier, and the individual results of each of these are within their subsections.

### 5.1.1 Accuracy Improvements

When evaluating the accuracy of a classifier, there are four possible results for the classification attempt: True Positives, True Negatives, False Positives, and False Negatives. These parameters are more descriptive than the general term "accuracy" and are used to compute the precision, recall, and F1-Score when evaluating the reliability of the classifier.

| | | **Predicted Class** | |
|---|---|---|---|
| | | Class=Yes | Class=No |
| **Actual Class** | Class=Yes | True Positive (TP) | False Negative (FN) |
| | Class=No | False Positive (FP) | True Negative (TN) |

Table 5.1: Positive and Negative Classifier Terms

Accuracy is defined as the sum of the True Positives and True Negatives over the number of classification possibilities (Accuracy = (TP+TN)/(TP+TN+FP+FN)). Precision deviates slightly in that it describes the ratio of true results over the number of positive classification attempts made (Precision = TP/(TP+FP)). Recall is similar to precision, but describes the ratio of true positives over all of the positive observations in the class (Recall = TP/(TP+FP)). The F1-Score is the harmonic mean, or weighted average, of the precision and recall values and takes both false positives and false negatives into the evaluation (F1-Score = 2*(recall*precision)/(recall+precision)).

The following table contains the size of the original dataset as well as the final sizes once the supplemental tweets were added.

| **Class Label** | **Original Total** | **Percent of Category** | **Improved Total** | **Percent of Category** |
|---|---|---|---|---|
| Related | 4294 | 98.3% | 4874 | 97.5% |
| Not Related | 76 | 01.7% | 127 | 2.5% |
| Total | 4370 | | 5001 | |
| Support | 3382 | 98.1% | 3877 | 96.2% |
| Against | 55 | 01.6% | 88 | 02.1% |
| Neutral | 12 | 00.3% | 64 | 01.5% |
| Total | 3449 | | 4029 | |
| Patronizing | 16 | 00.9% | 251 | 11.3% |
| Unwanted Sexual Attention | 33 | 01.9% | 260 | 11.7% |
| Predatory | 219 | 12.7% | 253 | 11.4% |
| Not Enough Context | 1452 | 84.4% | 1452 | 65.5% |
| Total | 1720 | | 2216 | |

Table 5.2: Final Dataset Size - After Additional Collection

Each label had at least one category that occurred significantly more frequently than the other categories, causing the representation of the less prevalent categories to all comprise less than 2% of the class (except for *predatory* behavior, which was 12%). These results were not proportional training sets and did not yield any reasonable accuracy with the text classifier. Consequently, I categorized more samples and supplemented the existing set by adding only the tweets labeled as being irrelevant, against, neutral, patronizing, or unwanted sexual attention. Afterward, the predominant category was reduced partially and the subsequent changes in accuracy, precision, recall, and F1-score were observed and reported. The reduction of the category was done using a unique random seed for relevance, stance, and sexual harassment category but the seed kept the same between each reduction call for that category to keep the results consistent. The following tables show the changes in the F1-score.

| **Percentage** | **Quantity** | F1-Score | | |
|---|---|---|---|---|
| **Removed** | **Remaining** | **Relevant** | **Irrelevant** | **Average** |
| 0% | 4874 | 0.99 | 0.27 | 0.97 |
| 15% | 3940 | 0.99 | 0.32 | 0.97 |
| 35% | 2866 | 0.99 | 0.44 | 0.96 |
| 45% | 2370 | 0.98 | 0.45 | 0.95 |
| **55%** | **1905** | **0.99** | **0.56** | **0.97** |
| 65% | 1470 | 0.97 | 0.35 | 0.94 |

Table 5.3: Improved Accuracy from Reducing the *Relevant* Category

Although the overall performance declines by removing relevant tweet samples, the recall and F1-score increase. Reducing the dataset by 55% yielded the best overall performance with the least sacrifice regarding the average. For this reduction, the overall accuracy was 97.2%, the precision was 0.97, the recall was 0.97, and the average F1-score was 0.97. The most significant weakness in this classifier is the recall of irrelevant tweets, which was only 0.39. Randomly reducing the number of relevant tweet samples prevents the same number of features from being preserved across each trial, but for approximately 1,985 total samples, there were 4,642 features produced.

| **Percentage** | **Quantity** | F1-Score | | | |
|---|---|---|---|---|---|
| **Removed** | **Remaining** | **Support** | **Against** | **Neutral** | **Average** |
| 0% | 3877 | 0.98 | 0.19 | 0.00 | 0.95 |
| 10% | 3391 | 0.98 | 0.20 | 0.00 | 0.95 |
| **20%** | **2944** | **0.98** | **0.27** | **0.00** | **0.94** |
| 15% | 3167 | 0.98 | 0.10 | 0.00 | 0.94 |
| 25% | 2731 | 0.97 | 0.00 | 0.00 | 0.93 |
| 30% | 2521 | 0.98 | 0.0 | 0.00 | 0.93 |

Table 5.4: Improved Accuracy from Reducing the *Support* Category

The results of the stance classifier were very inconsistent. Neutral and Against tweets were interpreted very differently among participants, and a relatively small amount of each sample was used. For example, some made a distinction between being against the movement and being against people involved in the movement, and some did not. Some people found constructive criticism to be supportive, but others saw constructive criticism to be against or neutral to the movement. There was also a discrepancy between how the author and labeler of the tweet interpreted the author's stance, such as when an author claimed to be in support of the movement but in practice expressed values that ran contrary to the movement, which caused inconsistency when labeling.

Because the context of the neutral and against samples varied significantly along with their labels, the performance of the classifier was affected more by *which* samples were being removed rather than by how many. At significantly higher reduction percentages (75-90%), the average F1-scores and accuracy increased for neutral and relevant tweets, but the overall accuracy decreased significantly. On average, the best performance for the stance classifier was reducing the supportive samples by 20%, which yielded an overall accuracy of 95.4%, an average precision of 0.93, an average recall of 0.95, and an average F1-score of 0.94. This ended with an average of 3,069 samples and 5,292 features.

| | | F1-Score | | | | |
|---|---|---|---|---|---|---|
| **Percentage Removed** | **Quantity Remaining** | **Patronizing** | **Unwanted Attention** | **Predatory** | **Not Enough Context** | **Average** |
| **0%** | **1452** | **0.12** | **0.24** | **0.65** | **0.87** | **0.69** |
| 5% | 1379 | 0.15 | 0.19 | 0.68 | 0.86 | 0.68 |
| 10% | 1307 | 0.07 | 0.30 | 0.65 | 0.86 | 0.67 |
| 15% | 1234 | 0.11 | 0.23 | 0.61 | 0.86 | 0.65 |
| 20% | 1162 | 0.03 | 0.31 | 0.53 | 0.85 | 0.64 |
| 25% | 1089 | 0.17 | 0.24 | 0.60 | 0.76 | 0.63 |
| 30% | 1016 | 0.16 | 0.24 | 0.48 | 0.83 | 0.61 |

Table 5.5: Improved Accuracy from Reducing the *Not Enough Context* Category

The category of sexual harassment classifier suffered from similar problems to the stance classifier. The interpretations for patronizing and unwanted sexual attention varied to a significant degree within the labeling process and among the authors of the tweet. For example, many individuals who used #MeToo in their tweet to describe being groped in a private body area self-identified as having experienced mild harassment, especially when this happened on public transportation or at a nightclub or bar. Sexism was particularly hard to characterize, as the other things described in the tweet were almost always employing similar language and situations to situations described in unwanted sexual attention samples, such as summing up the experience as 'harassment'. Predatory was relatively unique in the situations and language used, causing it to have significantly higher precision and recall than the preceding two categories despite similar sizes. As the "not enough context" category caught all tweets that didn't belong in one of the others, removing these samples and consequently their features generally appeared to lower the overall accuracy of the classifier. Without any reduction, the classifier's average accuracy was 71.3% with an average precision of 0.67, recall of 0.73, and F1-score of 0.69. Using 2,216 total samples yielded 3,329 features.

## 5.2 Comparison of Algorithms

The above results were found by arbitrarily eliminating a portion of the data; however, a specified random seed was used so that the results would be generally reproducible. The testing set was 20% of the samples sklearn's stratification parameter was applied when splitting. The C parameter that controls regularization was set to 1.0 and the loss was changed to "hinge" from "squared-hinge." These sizes and features were used to establish the baseline accuracy for each classifier. The following section shows the maximum accuracy achieved from each classifier with individual parameter tuning and adjustments of testing and training set sizes.

### 5.2.1 Support Vector Machine

The following tables detail the success in accuracy of the SVM classification algorithm.

| **Classification** | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
|---|---|---|---|---|
| Relevant | 0.97 | 1.00 | 0.99 | |
| Irrelevant | 1.00 | 0.39 | 0.56 | |
| | | | | |
| Average/Total | 0.97 | 0.97 | 0.97 | 97.23% |

Table 5.6: SVM Accuracy - Related

| **Classification** | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
|---|---|---|---|---|
| Support | 0.96 | 0.99 | 0.98 | |
| Against | 0.38 | 0.20 | 0.26 | |
| Neutral | 0.00 | 0.00 | 0.00 | |
| | | | | |
| Average/Total | 0.93 | 0.95 | 0.94 | 95.23% |

Table 5.7: SVM Accuracy - Stance

| **Classification** | **Precision** | **Recall** | **F1-Score** | **Accuracy** |
|---|---|---|---|---|
| Patronizing | 0.26 | 0.12 | 0.16 | |
| Unwanted Sexual Attention | 0.30 | 0.27 | 0.29 | |
| Predatory | 0.72 | 0.71 | 0.71 | |
| Not Enough Context | 0.83 | 0.93 | 0.88 | |
| | | | | |
| Average/Total | 0.69 | 0.73 | 0.71 | 73.4% |

Table 5.8: SVM Accuracy - Harassment Category

LinearSVC had fewer parameter options than SVM. Increasing the C parameter for regularization had a negative impact on relevance and stance, but yielded a marginal increase for the sexual harassment category. Hinge loss, which is standard for SVM but not for SVC, increased the overall accuracy by around 1-2% for each category. Stemming the language decreased the performance; however, lemmatization had a small positive impact on the stance and sexual harassment categories.

### 5.2.2 Naive-Bayes

The following tables detail the success in accuracy of the Naive-Bayes classification algorithm.

| Classification | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Relevant | 0.97 | 1.00 | 0.98 | |
| Irrelevant | 1.00 | 0.47 | 0.64 | |
| | | | | |
| Average/Total | 0.97 | 0.97 | 0.97 | 97.1% |

Table 5.9: Naive-Bayes Accuracy - Related

| Classification | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Support | 0.96 | 1.00 | 0.98 | |
| Against | 0.00 | 0.00 | 0.00 | |
| Neutral | 0.00 | 0.00 | 0.00 | |
| | | | | |
| Average/Total | 0.92 | 0.96 | 0.94 | 95.3% |

Table 5.10: Naive-Bayes Accuracy - Stance

| Classification | Precision | Recall | F1-Score | Accuracy |
|---|---|---|---|---|
| Patronizing | 0.00 | 0.00 | 0.00 | |
| Unwanted Sexual Attention | 0.33 | 0.08 | 0.12 | |
| Predatory | 0.80 | 0.08 | 0.14 | |
| Not Enough Context | 0.68 | 1.00 | 0.81 | |
| | | | | |
| Average/Total | 0.58 | 0.67 | 0.56 | 67.3% |

Table 5.11: Naive-Bayes Accuracy - Harassment Category

Although not pictured, the reduction process was indeed re-applied to Naive Bayes to verify that the sizes used in the SVM were also the optimal sizes to be used for Naive Bayes. The F1-Score for the stance and sexual harassment category remained the same, but the F1-Score for relevance was decreased by an additional 10% for a total of 65%, with 1,530 samples and 3,988 features. Naive Bayes's performance also decreased when applying lemmatization and stemming. Despite this optimization, SVM outperforms Naive Bayes for every class.

## 5.3 Discussion

This section discusses several points of concern and limitations that exist as well as things revealed through the processes in this thesis.

The LinearSVC and SVM discrepancy eliminated the ability to tune several parameters, as LinearSVC is not truly linear due to the inherent penalization of the intercept. This potentially lost accuracy and precision in those parameters that could not be tuned, such as the gamma.

Another significant limitation in this thesis is the minimal dataset. More tweets should be labeled in order to bring each category to an amount equaling several thousand at least for better results. Furthermore, all tweets should be labeled by a small group of people with a sufficient

understanding of the categories, or an assessment should be performed on the participant to avoid the vast discrepancies that were present in their labeling.

Another significant problem with the labeling process that was not considered preemptively was the tweets' authors having an incorrect perception of their own experience harassment or assault, which frequently influenced the interpretations of labeling. Frequently, the authors of tweets experienced a severe type of sexual harassment or even assault, yet discussed it more casually or directly identified as only having experienced a lesser type of harassment. Some participants were confused by this interpretation, causing many to disagree when labeling tweets as against and neutral as well as whether the tweet was patronizing or unwanted sexual attention. This project also revealed that authors of tweets predominantly used the hashtag to discuss forms of assault as opposed to forms of harassment.

In addition to increasing the sample size of the less frequent categories with data that has higher integrity, the thesis could be improved by employing more classifiers and comparing the results. SVM and Naive Bayes often perform similarly, and these two alone do not give significant insights to the data.

# CHAPTER 6

# FURTHER SUPPORT OF #METOO

## 6.1 Mitigating Sexual Harassment

This chapter discusses ways sexual harassment in society can be mitigated, and victims can be assisted. Because legal action is both a way to prevent the victim from suffering further and a means through which to end the perpetrator's behavior, understanding how "sexual harassment" relates to legal statutes for stalking, harassment, and cyberstalking is essential for victims as well as researchers (sexual harassment as a crime in and of itself has been discussed in Chapter 2). Additionally, some characteristics of cyberbullying and cyberstalking are unique to the online nature of the misconduct, and some of these problems are discussed in this chapter alongside some third-party solutions that are in progress. Lastly, this chapter contains a survey that was administered to aid in the research for this thesis and a discussion of the results.

### 6.1.1 Stalking and Sexual Harassment

The rise of the internet and social media has brought with it new mediums of communication, and some of these avenues have been abused by harassers, predators, and others with malicious intentions. Criminal charges such as "stalking" are now accompanied by the prevalence of "cyberstalking," the latter of which is generally comprised of the same antagonistic pursuit of an individual but through electronic methods. Despite its merits, the internet and cellular data have provided a new means through which offenders can sexually harass and terrorize victims online.

What falls in the domain of "harassment" varies everywhere, but in all considerations within the United States, this concept is addressed as a criminal offense contingent upon a credible threat and reasonable fear for the victim's safety, pursuant to the federal statute 18 USCS §2261A. As defined, the federal statute does not consider damage to mental health and stability from minor to moderate annoyances. While victims of "substantial emotional distress" can file criminal charges, victims of mild or moderate stalking behaviors can file civil charges if damages can be proven. More specific considerations vary significantly from state to state through the charges of stalking, menacing, and harassment; some states include all three under one label while others make careful distinctions to establish two or three independent charges.

The Stalking Resource Center, a branch of the National Center for Victims of Crime in the United States, provides regularly updated resources on navigating which legal statues apply to an individual situation. In the states that do recognize harassment as a criminal offense, the particular characteristics of stalking revolve around repeated attempts at communication that is meant to alarm, annoy, and terrorize victims. When penal statutes for stalking are written independently of harassment, stalking is typically defined as behavioral patterns that instill a reasonable fear for the safety of the victim yet may not have a direct line of communication to the victim. The criminal offense of menacing differs from stalking and harassment in that it does not have to be a pattern of behavior, and a single dire offense can give the victim room for advocacy. Many states criminalize all of these behaviors while only using one of these terms; thus, the concept of both stalking and harassment is generally discussed in research using only the term "stalking." Regardless, all variations of these statues contain the same pitfall in that valid charges require reasonable fear of a physical threat before the victim has any legal options for recourse [sta, 2005].

While not all forms of stalking and harassment are inherently motivated by gender, the majority of victims are women and the majority of offenders are men. In 1998, one in every twenty U.S. women was estimated to have been stalked in their lifetime, and more than three-fourths of all U.S. stalking cases were related to failed intimate relationships [Abrams and Robinson, 1998]. In 2005, one in 12 women and one in 45 men were predicted to have been victims of stalking [sta, 2005].

Across the majority of these cases, the victim and offender are neither employed together nor do they attend school together, which eliminates sexual harassment as an option for legal recourse even if the behavior is indeed sexually motivated. Victims and survivors of sexual harassment are often left with only filing harassment charges, filing stalking charges, and filing for a protection order as their only way to prevent the onslaught of unwanted attention.

Sometimes, a victim of stalking or harassment does not have sufficient documentation to meet the criminal burden of proof, or the victim opts not to file criminal charges for personal reasons. In such cases, victims still have the option to file for a civil protection order against the offender, commonly called a "restraining order." Because of their civil nature, protection orders only need to be proven by a preponderance of the evidence, which means "more likely than not." While this option is a viable way for victims to prevent further harassment, some individual state implementations limit the scope of people that can file for one. As of 2016 with the most recent state's update to protection order regulation, fifteen U.S. states require that the victim must be or have been in an intimate relationship with the offender or be directly related as family. Four states do not have an option for a civil protection order at all. Some states require physical contact to have been made between the offender and victim. Other states require criminal charges to be filed to be granted the order, which eliminates the utility of a "civil" protection order. In some cases, states require the victim to have been physically assaulted if the offender is not a direct family member. In many cases, the states require the charges to be filed in the specific city the offender exhibited the behavior and therefore requires that the victim make sacrifices to travel. Many other specifications that serve as obstacles for victims that are not described here exist in individual states and continue to be a problem for those seeking help [sta, 2005]. Civil protection orders are not an adequate solution for the majority of victims of stalking and sexual harassment.

Including solutions for stalking and harassment is an appropriate choice when considering solutions to sexual harassment because these offenses are not only typically gender-motivated but also the only recourse available to victims of sexual harassment in many circumstances. Cyberstalking statutes are defined in some places and are typically handled under the law the same way as stalking laws, with the provision that the communication or threats occurred through specific electronic mediums. Many sexual harassment perpetrators also choose to use these same mediums to exercise their harassing behaviors as well, making all of these criminal statutes appropriate considerations to include when evaluation sexual harassment.

### 6.1.2 Current Problems

Whether or not social media has adequate tools to handle online harassment is a frequent topic of research and discussion, partially because the details of each instance and how the victim wants to resolve the issue varies so significantly. When perpetrators use these mediums, victims are often put into the position of having to reduce or eliminate their online presence in order to avoid their offender, which can unfairly cause them to miss out on other opportunities or information in life [Fox and Tang, 2017]. Many other issues exist that have not been studied to the degree that researchers have a quantifiable level of impact or severity, but these issues have been identified through journalism, discussion, and research and are actively being investigated.

A group of researchers from MIT developed a product called SquadBox to work towards mitigating online harassment. In their research, they interviewed 18 candidates who had been victims of sexual harassment, and this insight guided the design of SquadBox. They found that in most cases, existing privacy features of social media were not sufficient to deter persistent and determined perpetrators. The ability to block accounts was continuously ineffective because the harasser could quickly and easily make new accounts. Many interviewees white-listed an approved list of contacts and blocked all others to thwart their harasser; however, in one case a participant said that this caused them to miss out on a job because the white-listing approach blocked a potential employer from extending an interview offer. Many participants were forced to withdraw from social media altogether, which unfairly causes them to miss out on other opportunities. Another respondent removed her ability to appear in Facebook searches, which caused her to miss a friend from decades prior attempting to reach out. All of these participants had different ways they wanted to respond

to harassment as well; some did not want to know that their harasser had contacted them at all, some wanted to be able to review it for reputational damage control, and some wanted it kept as evidence [Mahar et al., 2018]. SquadBox endeavors to improve the state of harassment online by allowing victims to assemble a "squad" of friends to review their messages before the victim sees them. The product is still being developed further before release.

Women, Action, and the Media (WAM!) created a team in 2015 to assist Twitter in evaluating harassment that took place on the website. A team of individuals worked through WAM! to review messages reported for harassment and escalate them based on priority, allowing Twitter to make the final decision. This report revealed severe security flaws in the lack of maintenance on archiving and storing all tweets and messages posted by users, particularly those of a harassing nature. Twitter cannot remove a tweet of a harassing nature if the evidence submitted is a screenshot–the tweet must still be available to be viewed live to be used as evidence when determining whether or not to take action against a Twitter user, which might make it harder for the victims to seek help. Additionally, because Twitter does not store the original tweet's contents permanently, victims are left without evidence to take to law enforcement even if Twitter does respond and remove the offending messages [Matias et al., 2015].

Furthermore, this report also reveals that when a harasser deletes his or her own tweet, or Twitter removes the offending tweet, that tweet is no longer available to law enforcement agencies [Matias et al., 2015]. Most social media companies, including Twitter, track metadata such as who the sender and receiver of messages are when messages are sent, and when they are opened even if the platform does not store permanent logs of the messages themselves. Sometimes in court, screenshots can be corroborated by metadata to validate the screenshots to be used as evidence. How social media evidence is handled changes across the different states, courts, and districts without a conclusive answer; however, if a victim does not take screenshots of misconduct, it makes proving the misconduct a problematic endeavor. Both administrative and user removal of messages contributes to this difficulty. WAM!'s report indicated that the lack of persisting data and messages over social media platforms is a disservice to victims as it often leaves them without evidence to take to the police and without evidence to have their harasser removed.

Twitter is not the only social media platform that has this problem. Instagram users have the option of deleting their comments and direct messages as well as hiding their presence from a user entirely, including their username. Instagram's public policy, like that of all social media platforms regarding their cooperation with law enforcement, states that they will comply with valid subpoenas, court orders, or warrants under outlined circumstances. However, if a victim were to be harassed over Instagram and the perpetrator removed their messages and presence from the perspective of the victim, the victim would have no evidence to bring to law enforcement to file charges and proceed with a legal request for the data. Students who were previously involved with Nikolas Cruz, the shooter in the Stoneman Douglas High School incident, have gone on record to say that Cruz had made violent threats against them and others through Instagram direct messages before the mass shooting. However, because Cruz used Instagram's unsend feature to remove these messages, the students were left with only their word when reporting his behavior to the administration [Wright, 2018].

This lack of persisting data disenfranchises victims of all crimes that occur on social media. Within the past year, several other messaging apps have enabled users to delete messages. Line rolled out this feature at the end of 2017 but requires that users delete their message within 24 hours [Lomas, 2017]. WhatsApp implemented the same policy earlier in the year but requires the messages to be unsent within only a few minutes [Chowdhry, 2017]. While these time limits prevent individuals from returning to the conversation to hide evidence of misconduct, it does not protect the victims who must deal with distress if they open and read the message before it is unsent.

After Mark Zuckerberg was found to have removed his personal direct messages sent over Facebook's messaging app Messenger, Facebook announced in April 2018 that it would soon be providing all users the option to unsend messages as well [Price, 2018]. Security and legal experts are concerned that this feature will make it even harder for many victims to find justice, particularly in the case of harassment victims. Often, harassment and stalking behaviors can only be substantiated with a series of evidence over time to prove the persistent nature of the behavior [Dwyer, 2018].

While many are actively working on improving the state of harassment online, technology continues to evolve fast enough that finding solutions is a challenging endeavor.

## 6.2 Survey

To further develop an understanding of how to mitigate sexual harassment, a questionnaire was developed. The questionnaire was reviewed by IRB and found no need for approval, as it dealt with voluntary participants above the age of 18, no identifying information was collected, no harm was posed to the respondents, and responses were completely anonymous. All parties involved in this thesis were CITI certified before distribution. The survey was sent to the presidents of college Greek fraternity and sorority organizations of different campuses and distributed among their members. The majority of the responses are from college students; however, some participants forwarded the survey link to their organization in ways that included alumni, causing some responses to be from older adults.

### 6.2.1 Questionnaire

The only demographic information requested was age and gender identity, but the questions were optional in the event the participant did not want to share this information. The survey then gave a list of behaviors and asked the respondent to mark whether they thought the behavior was a mild, moderate, or severe form of sexual harassment. Of the listed behaviors, seven would be classified according to this thesis as *patronizing*, four as *unwanted sexual attention*, and two as *patronizing* examples. The last behavior was a box for the participant to type another behavior not listed and classify it. Because not every survey respondent will not have had the same (or any) experience with sexual harassment, this question was included primarily to get the respondent to contemplate behaviors that they might not otherwise consider (ex. "Messages on social media from strangers") while they continue the survey to answer the questions regarding sexual harassment on social media.

The survey asks the respondent to indicate which social media platform they use the most regularly, including an option to enter a platform not listed within a text box. In the first distribution of the survey (240 responses), the social media platform Snapchat was not included on this list, and many wrote 'Snapchat' in the other option. Upon noticing this, Snapchat was included as an option in other distributions of the survey. The next two questions asked the respondent to consider the platform that they had indicated before and to evaluate its performance in preventing and reporting sexual harassment.

Lastly, the survey asked participants to answer what ways that social media platform could be designed to prevent sexual harassment from happening. Five examples were given where the respondent could check as many that applied, and there was a text box for adding the respondent's opinion. The full questionnaire administered in the study is available in Appendix B.

## 6.3 Survey Results

In the survey, the respondents select the social media platform they use the most and then evaluate how good that platform's mechanisms are at preventing and reporting sexual harassment.

Figure 6.1 shows a comparison between the responses for different platforms regarding how good those platforms are at **preventing** sexual harassment. In this graph, the results for "Tumblr" and "Other" have been omitted due to their low response rate. One respondent indicated GroupMe as their most used platform and evaluated it as having adequate mechanisms. Of the five respondents who selected Tumblr, one said that it was adequate, one said that it was inadequate, and three said that they do not know.

In addition to rating the social media platform's mechanisms for preventing sexual harassment, respondents also evaluated the platform's ability to **report** sexual harassment, shown in Figure 6.2.

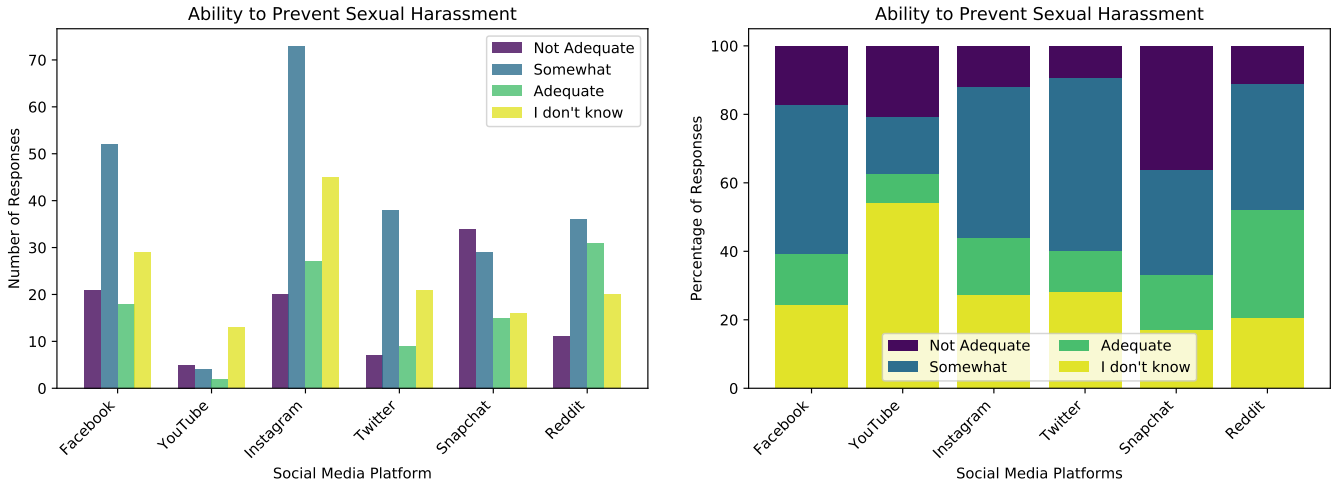



Figure 6.1: Evaluation of Social Media's Ability to Prevent Sexual Harassment

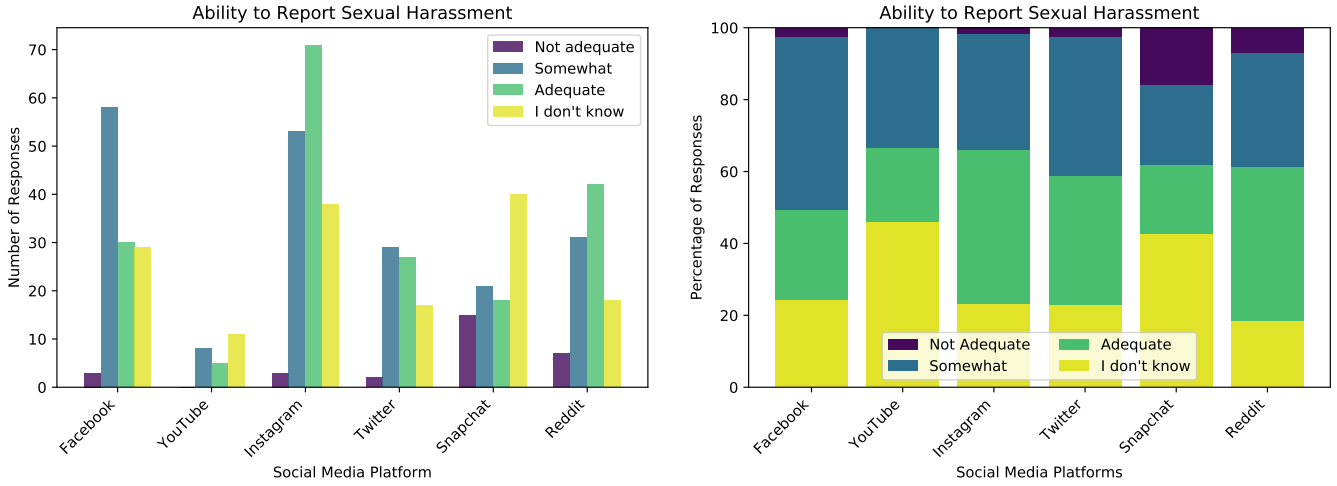



Figure 6.2: Evaluation of Social Media's Ability to Report Sexual Harassment

The suggestions to respondents were long and have been abbreviated in the figures below according to the following list:

**Advertisement -** Increasing the prevalence/priority advertisement which addresses sexual harassment

**Stricter Regulation -** Stricter regulation of posts/content on social media sites

**News Feed Adjustment -** Adjustments to news feed algorithms to prioritize content regarding victims of sexual harassment

**Automatically Archive -** Automatically archiving or documenting instances of sexual harassment online for evidence

**Connect Victims -** Connecting victims on social media with help (local authorities, other victims, or local resource options)

**Other -** Please describe any other thoughts you have on ways technology could mitigate sexual harassment

The final question from the questionnaire asks the respondents for their opinion on ways that social media platforms could be designed or improved to mitigate and prevent sexual harassment. Respondents could select as many options as they wanted, including the option to write in their own suggestions in addition to selecting them. Figure 6.3 shows the frequency each improvement suggestion was selected, distributed by gender.

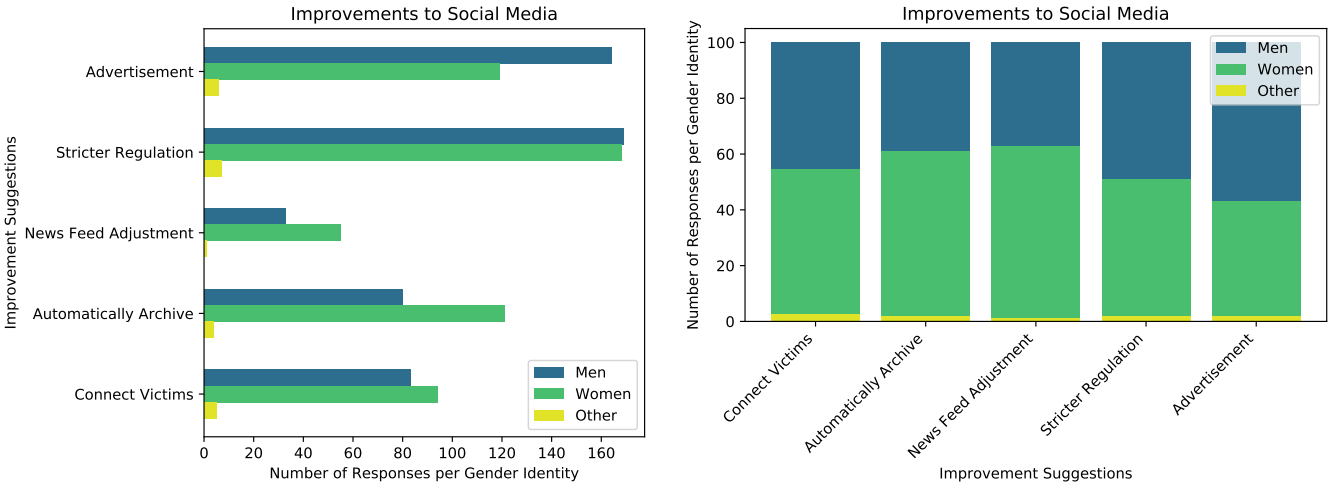



Figure 6.3: Improvement Suggestions to Social Media - Multiple Choice Selection

Lastly, 53 users opted to write in a suggestion. To organize them, they were grouped by the sentiment they were trying to express. Below is a brief description of each group alongside the abbreviation used in the following figures.

**Report to Authority -** Forward the instance from within the social media platform to some legal authority

**Stricter Disciplinary Action -** Enforce stricter bans, blocking, or punishments for offenders

**Improved Account Privacy -** Improve individual account privacy features

**Stronger Administrative Action-** Increased administrative action, such as increased response time to bans or greater attention spent to individual thread moderation

**Improved Reporting Systems -** Improved reporting systems that are not contingent upon administrative action

**Increased Public Awareness -** Increased public knowledge or awareness regarding sexual harassment facts as well as available options for defense

**No Action Due to Privacy Concerns -** Take no action due to concerns regarding individual users' privacy

**Victim's Responsibility to Act -** It is the victim's responsibility to avoid harassment

**Not Possible -** Nothing can be done, social media cannot accomplish this goal, or social media shouldn't be the one working on this goal

**Do Not Care -** Respondent does not care or does not know

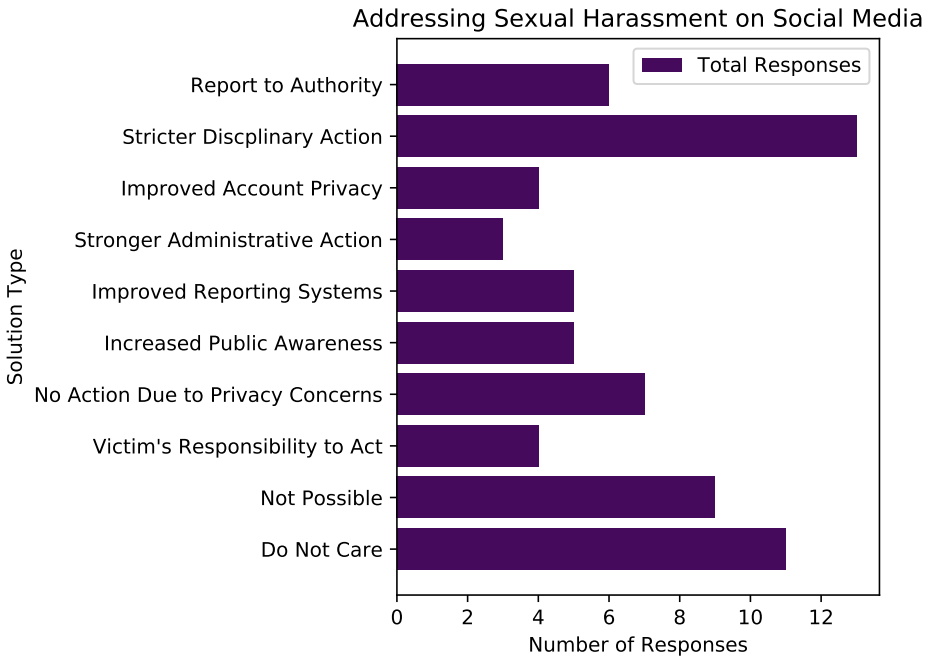



Figure 6.4: Total Responses of Each Solution Type

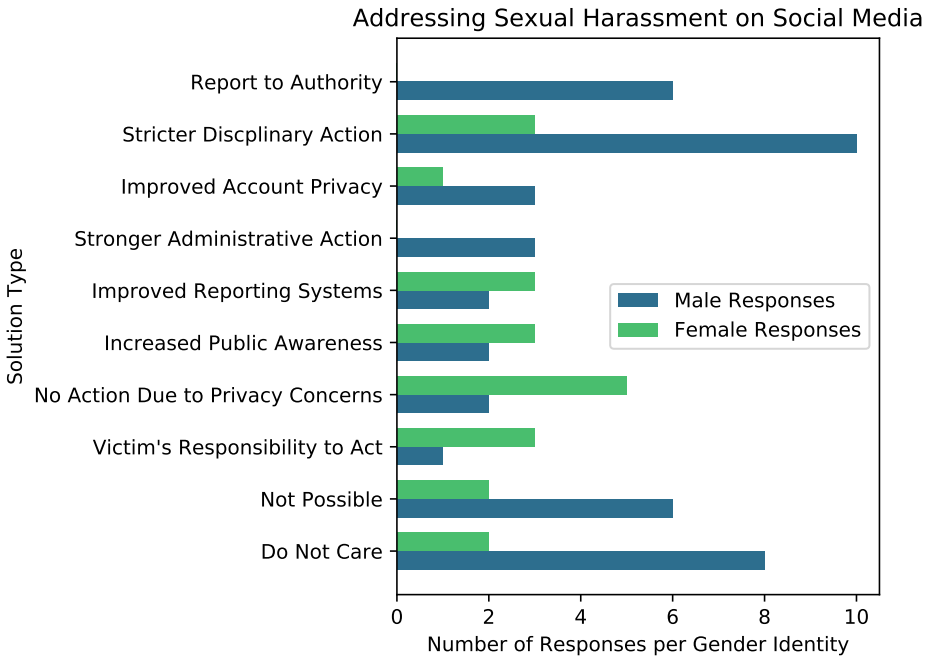



Figure 6.5: Responses of Each Solution Type per Gender Identity

Some text responses included more than one opinion in their text and were labeled as belonging to more than one group accordingly. Consequently, the sum of responses for each group is slightly higher than the total responses given. The full list of every response and the group it was assigned to

is available in Appendix C. Figure 6.4 shows the total distribution of the responses, and Figure 6.3 shows the distribution of responses by gender identity. Two respondents who identified as "other" were not included in the male-female gender comparison figure, but their responses were included within the total.

## 6.4 Discussion

The most successful platforms in preventing sexual harassment were Reddit, as well as Facebook and Instagram by a lesser degree. It is possible that this is because the moderators of individual boards are also users of the site and are not paid, and therefore have a more active roll in their individual communities. An individual on Reddit knows whom to contact when a reported instance is not being handled because moderators and admins are listed by username in an obvious manner. Moderators are also able to implement specific rules across domains of the site and enforce them with bots as well as reviewing cases, and they can answer mail from individuals. The least capable platform for preventing sexual harassment was found to be Snapchat, as well as Facebook and YouTube to a lesser degree. When using Snapchat, there is not a valid way to prevent someone from being exposed to inappropriate pictures other than preemptively blocking the individual, which could be the reason for this evaluation.

With regards to the social media platforms' capacity to report sexual harassment, Instagram was most favorable. Instagram has some privacy features on by default, such as not allowing strangers to direct message a user until their request to do so has been approved by the user. When reporting or blocking an individual, an option to do both is given. The UI for this process is fluid and simple, requiring only a couple button taps for any report. It is also straightforward to adjust privacy settings on Instagram as they are not hidden behind layers of navigation, which also might contribute to its evaluation. Reddit was also seen favorably for reporting sexual harassment, which could be for the same reasons listed above.

The least adequate platform for reporting sexual harassment was Snapchat by a significant margin. Snapchat's reporting system requires an individual to hold down on a person's name, select settings, and chose report. While this is a simple process, it only allows someone to report the user overall and does not allow someone to report individual messages. Cases for specific incidents do not exist, and it is not possible to see if the situation is resolved without filling out Snapchat's ticket form online which does not appear to communicate the status of the issue consistently.

The majority of responses came from men using Reddit. In general, men were more likely to say that they do not care about solving this problem or that there is not a solution at all. Men predominantly advocated for reporting verifiable offenders and frequent offenders to an authority, such as law enforcement, as well as harsher punishments on social media such as IP banning. Women, on the other hand, gravitated towards more mild solutions, such as taking no action due to their concerns about user privacy, improving the reporting system, and solving the problem by addressing what the general population is aware of. More women than men felt that addressing sexual harassment online was more the responsibility of the victim to take appropriate measures to ensure their own safety than it was for the social media platform to protect them; this answer makes sense alongside the preference to maintain privacy. There are many reasons why this might be the case, correlating with the under-reporting of sex crimes and frequent occurrence for women to feel embarrassed or anxious regarding their situations. The overall apathy from respondents selecting that there is no solution or that they do not care is indicative of the lack of knowledge regarding the issue of sexual harassment, which is consistent with the guiding sentiment behind the entire #MeToo movement.

Regarding the multiple choice "improvements to social media" question, all genders preferred addressing the issue through more strict regulation and moderation online. More men than women thought the problem might be improved with targeted advertisement, and adjusting news feed algorithms to prioritize the presence of victims was consistently the least favorable choice. A significant amount of women chose the option to automatically archive instances of sexual harassment for evidence, which might be because of the hesitation to report crimes and stigmas around survivors of sex crimes. It is consistent with the need to record evidence and maintain privacy regarding their experiences.

# CHAPTER 7

# CONCLUSION

This thesis is proof of concept for the development of and the utility in a classifier for sexual harassment online. The ability to identify tweets that discuss sexual harassment or assault could address the needs of individuals expressed within the survey results. Stricter disciplinary action and regulation of content, the most preferred solution by all genders in both ways the question was posed, could be assisted by escalation or corroboration from an algorithm that identifies harassing messages. It could also improve the ability to provide information and resource options to users which were heavily indicated on the survey responses.

The archiving of instances online could help a growing problem regarding using social media as evidence in legal battles against perpetrators. While this was not a heavily discussed answer, the survey results do indicate that it was favorable, and the review of this issue legally certainly demonstrates a need for more evidence.

### 7.0.1 Limitations

The most prominent limitation of this research is a small sample size regarding tweet samples and survey responses, as well as maintaining the integrity among tweet labels. Text processing limitations with URLs, chatspeak, and pictures hindered the efficacy of the vocabulary development as well. The categories were defined independent of knowing the trends in the tweets themselves, but their definitions and boundaries are confusing for the layperson and the tweet authors often self-identify as having experienced a different category. The classifier model only compares two different algorithms and does not achieve a high degree of accuracy when differentiating the stance of the tweet.

### 7.0.2 Future Work

The classifier could be improved by diversifying the implementation with different algorithms, tuning more parameters, and comparing them. The most effective model of the classifier could be deployed online to begin testing the applicability of it in accomplishing goals. More survey responses need to be collected as well as more tweets labeled from the smaller categories.

Additionally, there might be utility in contacting the authors of tweets using #MeToo, or leaving an option for survey respondents to be contacted, for more information regarding their experiences with sexual harassment to form a more comprehensive view of possible solutions and their limitations. It is also possible that gathering a collection of tweets along with their metadata and associated user profile information could show more interesting trends among the demographics of survivors. Even without that, further results could be explored through an analysis of a subset of the data made by the classifier, such as examining tweets with predatory behavior to determine the prevalence of abuse towards minors.

# BIBLIOGRAPHY


[sta, 2005] (2005). Stalking resource center. *Reference Reviews*, 19(7):33–34.

[Abrams and Robinson, 1998] Abrams, K. M. and Robinson, G. E. (1998). Stalking part i: An overview of the problem. *The Canadian Journal of Psychiatry*, 43:473–476.

[Cawley and Talbot, 2010] Cawley, G. c. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning research*, 11:2079–2107.

[Chamberlain et al., 2008] Chamberlain, L., Crowley, M., Tope, D., and Hodson, R. (2008). Sexual harassment in organizational context. *Work and Occupations*, 35(3):262–295.

[Chowdhry, 2017] Chowdhry, A. (2017). Whatsapp adds the ability to 'unsend' messages.

[Dwyer, 2018] Dwyer, D. (2018). Facebook "unsend" message option could increase cyber bullying, warn experts.

[Ferguson et al., 2005] Ferguson, T., Berlin, J., Noles, E., Johnson, J., Reed, W., and Vincent Spicer, C. (2005). Variation in the application of the "promiscuous female" stereotype and the nature of the application domain: Influences on sexual harassment judgments after exposure to the jerry springer show. *Sex Roles*, 52:477–487.

[Fitzgerald et al., 1995] Fitzgerald, L. F., Gelfand, M. J., and Drasgow, F. (1995). Measuring sexual harassment: Theoretical and psychometric advances. *Basic and Applied Social Psychology*, 17(4):425–445.

[Fox and Tang, 2017] Fox, J. and Tang, W. Y. (2017). Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. *New Media  Society*, 19:1290–1307.

[Gruber, 1992] Gruber, J. E. (1992). A typology of personal and environmental sexual harassment: Research and policy, implications for the 1990s. *Sex Roles*, 26:447–464.

[Juditha and Kominfo, 2015] Juditha, C. and Kominfo, K. (2015). Cyberstalking on twitter @triomacan2000 at election 2014. *Jurnal Penelitian Komunikasi*, 18:2460–0172.

[Lisak et al., 2010] Lisak, D., Gardinier, L., Nicksa, S. C., and Cote, A. M. (2010). False allegations of sexual assault: An analysis of ten years of reported cases. *Violence Against Women*, 16(12):1318–1334.

[Lomas, 2017] Lomas, N. (2017). Line adds unsend for recalling missent messages.

[MacMillan et al., 2000] MacMillan, R., Nierobisz, A., and Welsh, S. (2000). Experiencing the streets: Harassment and perceptions of safety among women. *Journal of Research in Crime and Delinquency*, 37(3):306–322.

[Mahar et al., 2018] Mahar, K., Zhang, A. X., and Karger, D. (2018). Squadbox: A tool to combat email harassment using friendsourced moderation.

[Matias et al., 2015] Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., and DeTar, C. (2015). Reporting, reviewing, and responding to harassment on twitter.

[Murrell and James, 2002] Murrell, A. J. and James, E. H. (2002). Gender and diversity in organizations: Past, present, and future directions. *Sex Roles*, 45:243–257.

[Pierce, 1989] Pierce, M. R. (1989). Sexual harassment and title vii – a better solution. *Boston College Law Review*, 30(4):1071–1101.

[Pranckevicius and Marcinkevicius, ] Pranckevicius, T. and Marcinkevicius, V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, 5(2):221–232.

[Price, 2018] Price, R. (2018). Facebook's plan to let users 'unsend' messages could boost harassment and bullying, experts warn.

[Ramasundaram and Victor, 2013] Ramasundaram, S. and Victor, S. (2013). Algorithms for text categorization : A comparative study. *World Applied Sciences Journal*, 22:1232–1240.

[Schneider et al., 1997] Schneider, K. T., Swan, S., and Fitzgerald, L. F. (1997). Job-related and psychological effects of sexual harassment in the workplace; empirical evidence from two organizations. *Journal of Applied Psychology*, 82:401–415.

[Studzińska, 2015] Studzińska, A. (2015). Gender differences in perception of sexual harassment. (2015TOU20052).

[Till, 1980] Till, F. J. (1980). Sexual harassment: A report on the sexual harassment of students. pages 4–22.

[Wright, 2018] Wright, S. (2018). Three former classmates attempted to alert school authorities, after repeated threats from nikolas cruz.

[York and Brookhouse, 1988] York, K. M. and Brookhouse, K. J. (1988). The legal history of work-related sexual harassment and implications for employers. *Employee Responsibilities and Rights Journal*, 1(3).

# CHAPTER 8

# APPENDIX

## 8.1 Appendix A

| **Harassment Types** | **Descriptions** |
|---|---|
| A. Verbal Requests (more to less severe) | |
| 1. Sexual bribery | A request with a threat and/or promise of reward: a *quid pro quo*. Examples: offers of money for sex; offers of better working conditions for sex. |
| 2. Sexual advances | A request without a threat or promise that seeks sexual intimacy. May be in the form of questions or statements expressing a sexual interest. Examples: "Will you be my lover?" "How would you like this (beard) between your legs?" Statements indicating intentions or desires ("I'd like to . . . or "When I see you I want to...") of a sexual nature. |
| 3. Relational Advances | Request without a threat or promise that seeks a social relationship. Sexual desire or intent is not stated or implied. These are harassful by virtue of *repetition* (e.g., "nagging", "badgering"). Women often complain of men who "won't take no for an answer." |
| 4. Subtle pressures/ advances | Statements in which the *goal* or the *target* of the request is implicit or ambiguous. Their harassing nature is seen most clearly through an analysis of the full *context* of the interactions. Examples: double entendres; "Wishing out loud" comments ("I need some TLC"; "I'm really horny today"); or inappropriate personal questions ("Would you ever date a married man?" "Have you ever had an affair?"). |
| B. Verbal Comments (more to less severe) | |
| 1. Personal Remarks | Comments or questions of a nonsolicitory nature directed *to* a woman: includes jokes, teasing, questions about sexuality or appearance, and semantic derogation. Some solicitations are "remarks" because of the context. Examples: requests for body measurements; comments attributing a state of sexual arousal to the harassed; sexual slurs. |
| 2. Subjective objectification | Remarks *about* a woman either in her presence or in the form of rumors. The recipient is "invisible" or physically absent from the sexual discussion of her. Examples: rumors of alleged lesbianism or sexual promiscuity; public discussion about the harassed's body or sexuality. |
| 3. Sexual categorical Remarks | Sexually based comments about other women or women "in general." Includes comments which slur womanhood or particular groups or categories of women. Also *bystander harassment*, witnessing the requests or remarks of a harassing nature addressed to specific other women, is categorized here. Examples: "Women are whores"; women in a department are called the "cunt brigade"; a female co-worker's sexual anatomy is discussed in front of other women. |
| C. Nonverbal displays (more to less severe) | |
| 1. Sexual assault | A prolonged or intense and aggressive form of sexual contact involving coercion. Examples: actual or attempted intercourse; fondling recipient's sexual anatomy. |
| 2. Sexual touching | Includes both *sexual* and *sexualized* touching. The former is more brief and more spontaneous than sexual assault (e.g., a pinch, a grab) so that the terms "coercion" and "resistance" do not apply well. The latter requires greater understanding of the *context* of the interaction than the former. |
| 3. Sexual posturing | Includes violations of personal space and attempts to (or threats to) have physical contact. Distinctions should be made as to whether the posturing involves her directly (e.g., following or cornering; attempted grabs), as an audience member (e.g., men feigning masturbation with each other), or as a bystander (recipient observes other women being approached directly, touched, or assaulted). |
| 4. Sexual materials | Pornographic materials or objects which sexually debase women or womanhood (movies, magazines, pictures, "sex toys"). Also, the *profanation* of women's sexuality or personal items (e.g., underwear, menstrual cycle). |

Table 8.1: Jame's E. Gruber's Topology of Sexual Harassment

## 8.2 Appendix B

Answer choices marked with an asterisk (*) indicates that the answer choice had a text box where the participant could type an answer.

**Survey for Understanding the Extent of Sexual Harassment**

1) What is your age?*
2) What is your gender identity?
   - Male
   - Female
   - Other*
   - Prefer not to respond
3) Select all of the following behaviors that you consider to be forms of sexual harassment and the degree of severity you consider the infringement to be.

| **Action** | **Mild** | **Moderate** | **Severe** | **Not Sexual Harassment** |
|---|---|---|---|---|
| Staring or leering | ○ | ○ | ○ | ○ |
| Whistling, catcalling, or winking at someone | ○ | ○ | ○ | ○ |
| Pinching or poking | ○ | ○ | ○ | ○ |
| Sexist comments | ○ | ○ | ○ | ○ |
| Inappropriate drawings | ○ | ○ | ○ | ○ |
| Messages on social media from strangers | ○ | ○ | ○ | ○ |
| Making lewd/sexual remarks about someone's looks or body | ○ | ○ | ○ | ○ |
| Obscene gestures or sounds | ○ | ○ | ○ | ○ |
| Sending repeated messages, calls, or other forms of contact after the receiver expresses disinterest | ○ | ○ | ○ | ○ |
| Asking overly personal questions | ○ | ○ | ○ | ○ |
| Stalking or Harassment | ○ | ○ | ○ | ○ |
| Groping or other touching | ○ | ○ | ○ | ○ |
| Sending unsolicited photos of private body areas | ○ | ○ | ○ | ○ |
| Other (please specify)* | ○ | ○ | ○ | ○ |

4) What social network platform do you use the most frequently?
   - Facebook
   - Youtube
   - Instagram
   - Twitter
   - Snapchat
   - Reddit
   - Tumblr
   - Other (please specify)*

5) Considering the social media network designated above, how would you rate this social network platform in terms of its ability **to prevent** sexual harassment from happening?
   - It does not have any mechanism to prevent sexual harassment from happening
   - It has some mechanisms but these are not adequate
   - It has a good set of mechanisms to prevent sexual harassment from happening
   - I don't know
6) Considering the social media network designated above, how would you rate this social network platform in terms of convenience of **reporting sexual harassment**?
   - It does not have any mechanism to report an instance of sexual harassment
   - It has some mechanisms but these are not adequate
   - It has a good set of mechanisms to report sexual harassment
   - I don't know
7) In your opinion, in what ways could this social network platform be designed to prevent sexual harassment from happening? Mark all that apply.
   - Increasing the prevalence/priority advertisement which addresses sexual harassment
   - Stricter regulation of posts/content on social media sites
   - Adjustments to news feed algorithms to prioritize content regarding victims of sexual harassment
   - Automatically archiving or documenting instances of sexual harassment online for evidence
   - Connecting victims on social media with help (local authorities, other victims, or local resource options)
   - Please describe any other thoughts you have on ways technology could mitigate sexual harassment:*

## 8.3 Appendix C

### Questionnaire Responses Grouped by Type of Solution

| Solution Type | Responses |
|---|---|
| **1. Report to Authority** | **Forward the instance from within the social media platform to some legal authority**<br>- I think most cases of sexual harassment through social media is not anything serious as you can block people. I think the only way there can be sexual harassment charges is if the reciever is under aged or they threaten to hurt someone.<br>- If an excessive number of posts are reported, submit the name to an authority<br>- Not really much other than banning sexual harassment users and reporting to authorities if its severe<br>- actually holding perpetrators accountable<br>- companies that save/maintain/sell your info to 3rd parties should also save/maintain anything that could be used as evidence |
| **2. Stricter Disciplinary Action** | **Enforce stricter bans, blocking, or punishments for offenders**<br>- Zero Tolerance Policies that, upon having proof of sexual harassment, lock further messages from the IP address of the sender<br>- Could be quicker to use harsh punishments like account banning. Also, steps could be taken to prevent malicious users from creating new accounts and repeating the behavior.<br>- moderate threads<br>- Blocking accounts that make lewd or obscene posts<br>- Have an option to block and report someone for sexual harassment<br>- Not really much other than banning sexual harassment users and reporting to authorities if its severe<br>- Stricter punishments for repeat offenders who use social media platforms to stalk/harass others<br>- IP bans, shadow bans, etc. |
| **3. Improved Account Privacy** | **Improve individual account privacy features and options**<br>- More privacy settings in terms of who can DM you (applies to all social media platforms)<br>- Somehow flagging people who send inappropriate messages so that they can be easily removed from the platform. Also, some people will tag random other people in inappropriate posts, so maybe a way to only allow certain friends to tag you in posts would help<br>- Regulating who can send messages to other users.<br>- Ability to block trolls, harassers by IP instead of by user name. |
| **4. Stronger Administrative Action** | **Increased administrative and moderation action, such as response time to bans or greater attention spent to individual thread moderation**<br>- Requiring moderators to ban users for instances of sexual harassment.<br>- quick responses to reporting of comments/posts<br>- let finstas not be allowed to be a thing<br>- allow anonymous ways to report sexual harassment on the platform, but only if the reporter has a mechanism to provide documentation directly from the site so as to avoid false accusations. should be reviewed by humans<br>- Admin behind the tech needs to take complaints seriously.<br>- Actually doing something when people report abusers<br>- Consistent punishment of offenders and thenpunishment made public |

Table 8.2: Questionnaire Responses Pt. 1

| Solution Type | Responses |
|---|---|
| **5. Improved Reporting Systems** | **Improved reporting systems that are not contingent upon administrative action**<br>- AI that detects harrasment<br>- Ability to report harassment, which would lead to immediate blocking and temporary account freeze till user is educated about their behaviour<br>- Somehow flagging people who send inappropriate messages so that they can be easily removed from the platform. Also, some people will tag random other people in inappropriate posts, so maybe a way to only allow certain friends to tag you in posts would help<br>- allow anonymous ways to report sexual harassment on the platform, but only if the reporter has a mechanism to provide documentation directly from the site so as to avoid false accusations. should be reviewed by humans<br>- Being able to report to all subreddit's mods, or admins if they do nothing. Not all mods care at all.<br>- A way to report perpetrators as opposed to simply blocking or unfriending them |
| **6. Increased Public Awareness** | **Increased public knowledge or awareness regarding sexual harassment facts as well as available options ofr defense**<br>- It could be easier if they designed concrete rules on what constitutes sexual harassment, and then had a mechanism for reporting which actually worked. Even in PMs.<br>- Tech can't mitigate it, maybe if you really wanted to you could put some informative ads but thats about it sadly. Some people just don't get social cues and/or just really thirsty<br>- I think this pertains to the option above, but when you report something on Instagram, Facebook, etc, it could pop up with a list of national, state and local resources based on your location.<br>- increase awareness that men are often victims and people need to start standing up for them<br>- if people actually knew what was and wasn't sexual harassment it would be a lot easier |
| **7. No Action Due to Privacy Concerns** | **Take no action due to concerns regarding individual users' privacy**<br>- I don't really think social media can strongly fight sexual harassment. Users can block harassers, and the website can suspend reported harassers; but that's about it. Anything further requires removing anonymity on social media so that their actions can be permanently linked with them, but that also risks exposing the personal information of victims.<br>- None of the above are good suggestions for limiting cases of sexual harassment on social media given how the way they are worded could imply encroaching on the First Amendment and freedom of information. The best way is for the user being "sexually harassed" on social media is for them to continuously block the individual and completely disintegrate contact with the harasser.<br>- We must be careful in how sexual harassment is defined if we are to implement technology that, in effect, acts as the judge and jury over that person's ability to use the said platform. It is also worth noting that if any such algorithm(s) were to be effective, the definition of sexual harassment must be constrained considerably to reach the confidence intervals necessary to identify valid threats while minimizing the false positives resulting from unintended consequences of modeling a lot of otherwise common behavior that may happen to actually constitute harassment in certain contexts considering the sexual orientation of parties involved. Only the most flagrant of harassment would have a unique enough behavior profile to be segmented and eradicated effectively, and It should be. But the engineers of such software have an ethical responsibility to be very thoughtful in how these models are trained and come to be defined. And as an engineer you have a responsibility to portray the concrete and abstract complexity which such models entail. That's my 2 cents for what it's worth. |

Table 8.3: Questionnaire Responses Pt. 2

| Solution Type | Responses |
|---|---|
| **8. Victim's Responsibility to Act** | **It is the victim's responsibility to avoid harassment**<br>- You don't have the right to not be sexually harassed online anymore than you have the right not to be harassed online.<br>- None of the above are good suggestions for limiting cases of sexual harassment on social media given how the way they are worded could imply encroaching on the First Amendment and freedom of information. The best way is for the user being "sexually harassed" on social media is for them to continuously block the individual and completely disintegrate contact with the harasser.<br>- You can't prevent someone else's behavior. You can only control your response to their actions.<br>- I'm not particularly concerned with cyber sexual harassment. When one is too offended by a particular social media platform, it does not always point to a fault on the part of the platform. Use a different platform or learn to deal with uncomfortable situations. |
| **9. Not Possible** | **Nothing can be done, social media cannot accomplish this goal, or social media platforms shouldn't be the ones working on this goal**<br>- None ways to prevent it, maybe ways to stop it when it starts.<br>- I think most cases of sexual harassment through social media is not anything serious as you can block people. I think the only way there can be sexual harassment charges is if the reciever is under aged or they threaten to hurt someone.<br>- none. definitely don't want any increased sexual harassment ads, stricter regulations, more news about sexual harassment, etc.<br>- No way<br>- There's really nothing that can mitigate sexual harrassment short of removing a substantial amount of control from the user base.<br>- You can't<br>- Tech can't mitigate it, maybe if you really wanted to you could put some informative ads but thats about it sadly. Some people just don't get social cues and/or just really thirsty<br>- I think Reddit hardly qualifies as a social network and is this much more removed from the idea of sexually harassing an individual.<br>- None<br>- I don't really think social media can strongly fight sexual harassment. Users can block harassers, and the website can suspend reported harassers; but that's about it. Anything further requires removing anonymity on social media so that their actions can be permanently linked with them, but that also risks exposing the personal information of victims.<br>- Not a way<br>- Sexual harassment is the fault of the one doing the harassing. Technology is not responsible for it.<br>- The block feature that's available on every social media platform |
| **10. Do Not Care** | **Respondent does not care or does not know**<br>- They do all of this<br>- I don't know<br>- Who cares just block them<br>- I don't know and really don't care just block someone who does it<br>- It's why snap chats around. Send nudes<br>- I'm not particularly concerned with cyber sexual harassment. When one is too offended by a particular social media platform, it does not always point to a fault on the part of the platform. Use a different platform or learn to deal with uncomfortable situations. |

Table 8.4: Questionnaire Responses Pt. 3

## 8.4 Appendix D

My pet rabbits, who inspire me every day in my endeavor to make the world a better place. Buddy, Ezra, Ellie, and the newest member "Baby Bun."