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

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my advisor Dr. Naeemul Hassan for the continued support and knowledge while guiding me throughout my thesis. I cannot overstate my gratitude for the degree of patience and understanding demonstrated with me throughout this process and the contribution Dr. Hassan brings to the computer science department.

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.

My sincerest thanks also goes to the computer science department at the University of Mississippi, for accepting an art student with no background in computer science and developing me into the competent programmer 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.

Thank you to everyone who participated in this project and took the time to read my research on sexual harassment, my categorization rules, and assist me in manually categorizing a set of tweets: John Joseph Angel, Ainsley Ash, Tim Dolan, 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.

I would also like to thank the Title IX office and Department of Violence Prevention for the accommodations, help, and support given during my own struggles throughout this past year. Thank you for empowering me with the knowledge and support I needed to finish my last undergraduate year, and for inspiring me to take this direction with my thesis.

Lastly, I would like to thank my family and friends who have supported me throughout this thesis as well as my undergraduate career. Looking back on my transcript and resume as I prepare to graduate, I see nothing but 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. In particular I would like to thank Dr. Kristin Davidson, Gracie Hubacek and Ethan Luckett for not only your help in all my academic challenges but also for your unending support as friends. I would like to 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)

My thesis advisor Dr. Naeemul Hassan has collected approximately half a million tweets that contain the hashtag #MeToo and completed an initial analysis regarding various aspects of this data. His work, as well as others', could be expanded if the tweets were categorized according to their semantics and if analyses could be performed on isolated sets of this data. This thesis contains a program that can process these tweets and categorize them by the type of sexual harassment expressed, if applicable. By using this tool to filter out extraneous tweets to isolate the ones that express a personal experience, researchers would be privy to more useful analysis of the data by working exclusively with tweets that express a personal experience with sexual harassment while knowing the type of harassment described.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CHAPTER 1

INTRODUCTION

1.1 Problem

> "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. With the growth of the movement, the hashtag #MeToo is used widely on social media from people who are in support of the movement but not victims themselves, from people who are antagonists or critics of the movement, in general discussion and news coverage as well as for its original purpose of victims communicating a personal experience. As the study of sexual assault and harassment grows in relevance and popularity, the #MeToo movement exists as an unprecedented platform to be vocal about personal experiences regarding sexual assault and harassment. This thesis explores this platform and endeavors to use it to draw new conclusions about the demographic groups who experience these problems.

1.2 Scope

This thesis only covers a basic classification implementation and also provides an online interface to allow others to use the classification tool without compiling and processing on their local machine. However, research indicates many other possible avenues through which this program might be applied. These possibilities and applications are discussed within the Supporting Materials chapter.

1.3 Classification Overview

blah

CHAPTER 2

PREPARATORY WORK

This chapter addresses the research and work conducted before the development process. It describes sexual harassment as it has developed and recognized both legally and psychologically as well as a discussion of the literature regarding sexual harassment categorization. This was fundamental to designing and applying a strong, comprehensive categorization topology to the dataset. This chapter also discusses the data used in the program, how it was collected, and how it was manually categorized for the supervised machine learning process.

2.1 Legal History of Sexual Harassment in the United States

Categorizing different types of sexual harassment is a task that remains incomplete despite steady, incremental progress. Researchers first began making significant contributions to the categorization of sexual harassment in the 1980's because "sexual harassment" wasn't legally recognized until the 1970's where it was initially established through case law. Initially, the Title VII of the Civil Rights Act of 1964 made it illegal to discriminate employment opportunities 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 today. As a controversial topic with many nuances and varied 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 harasser was found 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, and subsequently the Equal Employment Opportunity Commission (EEOC) refined the rules to explicitly cover this type of sexual harassment. Because of these changes, the plaintiff in the landmark case *Meritor Savings Bank v. Vinson* (1986) effectively established *quid pro quo* behaviors as a form of sexual harassment as they were in violation of the EOCC rule changes. This case also recognized that comments of a *quid pro quo* nature were in violation of Title VII even if the victim suffered no tangible consequences because they contributed to a "hostile work environment" [**?**]. Now, sexual harassment continues to fall under the domain of the the Equal Employment Opportunity Commission (EEOC) and Title VII, and the accepted definitions categorize sexual harassment as either being *quid pro quo* acts or behaviors that contribute to a "hostile work environment." 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.

What behaviors constitute sexual harassment as a criminal offense is continuously revisited through new court decisions. 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. Actions that contribute to a "hostile work environment" are the much more common category of sexual harassment. Lesser behaviors, such as offhand remarks, teasing or banter, and personal questions are not explicitly banned as acts of sexual harassment. However, if these acts occur with a sufficient degree of severity or frequency, they would then qualify as behaviors creating a "hostile work environment", thus qualifying as sexual harassment and by law must be evaluated comprehensively.

When considering behaviors that do not take place in a professional setting, the actions cannot be evaluated under Title VII or Title IX regulations. 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 harassment, stalking, cyberstalking, sexual assault, or other criminal laws. This technicality makes 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.

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 in order 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 as a guide when evaluating another. The first widely used standardization was designed by Frank J. Till 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 [**?**]. This categorization was the most frequently used throughout the 1980's, although most organizations generally either rephrased the categories or consolidated them into three categories.

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 one, for a total of 11 distinct, exhaustive, and mutually exclusive categories. The first category, "verbal requests", includes sexual bribery, sexual advances, relational advances, subtle pressures/advances. These are all behaviors that are said directly to the victim with the intent of a sexual or personal relationship goal. The second category, "verbal comments", includes personal remarks, subjective objectification, and sexual categorical remarks. These are statements of a nonsolicitory nature directed either to a woman (ex. teasing or jokes), about a woman, or about women in general. The third category of "nonverbal displays" includes sexual assault, sexual touching, sexual posturing, and possession/display of inappropriate sexual materials. Altogether, there are 11 distinct types of sexual harassment, and these categories are both mutually exclusive and reflective of the EEOC's guidelines [**?**]. 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. These mutually exclusive categories establish a fluid progression in how certain behaviors can contribute to a hostile work environment, and Gruber's work continues to influence many more recent approaches to this topic. In a 2005 review of past, present, and future directions of improving gender and minority diversity in professional environments, Murrell and James refer to Gruber's work as a cornerstone in developing a comprehensive legal definition of sexual harassment [**?**]. 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 in order to do so [**?**]. 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. Consequently, many researchers continue to consolidate these categories, and in doing so they claim a lesser degree of specificity to allow for the nuances of peer to peer harassment to be appropriately placed.

In order to accurately categorize a behavior as sexual harassment, a neutral, third party ought to be able to corroborate the victim's opinion through an objective analysis. The victim's perception is integral not just for legal categorization but also to understand the degree and severity of the harm inflicted on those who are exposed to the behavior; however, the victim might be biased and in some rare cases, dishonest. A 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 is aware of it, will still have negative outcomes on the victim [**?**] in terms of health

and distress. Furthermore, 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 attempt 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 direct confrontation of a sexual nature from strangers [**?**]. Altogether, this indicates that while Gruber's categorization might be the most appropriate tool for evaluating sexual harassment in a professional capacity because of it's 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 simpler, consolidated categorization architecture in order to consequently develop a better questionnaire for measuring sexual harassment. Their categorization was comprised of three types: 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" in order 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 [**?**]. In a 2008 publication, Chamberlain et al. deviated from Gruber's 11 types for similar reasons. Upon the basis that these legal-driven approaches "underscore diversity" and "suggest substantial variation with regard to intent and severity" [**?**] from the victims' perception. 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).

2.3 Data

A collection of 10,000 tweets were originally assembled into a single excel sheet with the tweet's unique ID and contents in two columns. The next three column headers contained the checks of relevancy, stance, and type of sexual harassment. If the correct classification could not be determined from the text available, the index was left blank.

		Classification Labels		
		Relevant 1. Related 2. Unrelated	**Stance** 1. Support 2. Against 3. Neutral	**Harassment Category** 1. Patronizing 2. Unwanted Sexual Attention 3. Predatory 4. Not Enough Context
id_str	**text**			
1	'. . . example . . . '	1	1	3
2	'. . . example . . . '	2		
3	'. . . example . . . '	1	2	

Table 2.1: Categorization Header

To create a reliable training set of data, a sufficiently large amount of tweets were manually categorized by human readers. These voluntary participants were given a comprehensive explanation of sexual harassment both legally and psychologically, examples of properly categorized tweets, category definitions, as well as the categorization rules.

Participants were each provided 500 unlabeled tweets and asked to categorize them according to the rules provided. 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.

5,000 tweets were categorized total, with each set of 500 being categorized by a male and female.

Cumulatively, 10,000 categorizations were made. Only the tweets that were categorized identically by both the male and female participant were considered in the training and testing sets of data.

CHAPTER 3

CLASSIFICATION METHODOLOGY

For every tweet, the program will make either one, two, or three determinations depending on the level of relevancy, context, and detail. It will first determine whether or not the tweet using #MeToo is relevant to the movement. If it is relevant, it will then determine the stance of the tweet regarding the movement. If there is enough context, the last check is to determine the type of sexual harassment or assault being described.

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 ones that are written by bots and no meaningful determination can be made, the tweets are unintelligible, the entirety of the tweet's context can only be determined through following a URL or image, tweets that use the hashtag #MeToo for the purpose of winning a giveaway or for increased visibility in a promotion, or 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	**Original Tweet Text**	**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 Discerning the Stance

When manually categorizing tweets within the excel sheet, each tweet that is relevant to the #MeToo movement with enough context to ascertain a stance was assigned a 1, 2 or 3 accordingly:

1. Support
2. Against
3. Neutral

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 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 a personal experience are made in the third check of the program. Tweets that are critical or against the movement are labeled as thus, and tweets asking a sincere question or making an ambiguous remark regarding the movement are considered neutral.

ID	Original Tweet Text	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

Following the school of thought of those researchers who consolidated Gruber's categories is ultimately the strongest approach to the 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 a tweet with confidence, and the personal bias of the author of each tweet could result in improper categorization if the algorithm were to attempt to place each tweet within one of Gruber's very specific categories. Ultimately, three broad categories have been defined that consolidate Gruber's 11 types into each one: *patronizing*, *unwanted sexual attention*, and *predatory*.

3.3.1 Category A: 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 in nature, and nonsexual behaviors and remarks that the victim interprets as being sexual. Because many victims experience sexual harassment in contexts that are not applicable to Title VII or Title IX but cannot otherwise cannot

legally qualify as harassment, there is a lot of controversy among perception of these behaviors. In general, if a neutral, unaffiliated third party witnesses the behavior and could reasonably deem the behavior as being nonsexual yet the percipient still perceives it as thus, that type of harassment would be classified as patronizing behavior; additionally, this requires that the remark or behavior does not have an obvious sexual goal with the victim.

ID	Original Tweet Text	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 To all the boys driving, yelling perverted things at me, and my mom who said its because of the clothes I wore...at 12 years old	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: Patronizing Tweet Examples

Relational advances, such as repeated contact of a nonsexual nature, could reasonably make a victim fear that the advances might be sexually motivated and therefore cause them to interpret the behavior as harmful. In some cases, it is revealed in hindsight that the offender had a sexually motivated goal in regards to the victim, but it could not be objectively determined at the time the behavior was exhibited.

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 between issues that occur in professional versus casual environments. When evaluating non-professional contexts, these behaviors are not always present to a degree of severity or frequency that they could constitute a different criminal offense (stalking/cyberstalking, harassment, etc.) where they would easily qualify as sexual harassment in a professional setting because of the behavior's contribution to a hostile work environment.

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 professional versus casual environment discrepancy.

Behaviors that are subject to controversy, such as establishing the boundary between flirting and harassment, are likely to be placed here. These behaviors comprise the category called "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, level of attractiveness, etc.), and other minor behaviors that would legally contribute to a hostile work environment but cannot when reviewing peer-to-peer harassment.

3.3.2 Category B: 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 nature. Explicitly means that an unaffiliated, objective third party would also find the nature of the comment to be sexual. 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 complex to categorize because the harasser legally has the room to act within a reasonable degree of respectful, personal interest and flirting with the victim.

ID	Original Tweet Text	Relevant	Stance	Harassment Category
1	'. . . example . . . '	1	1	2
2	'. . . example . . . '	1	1	2
3	'. . . example . . . '	1	1	2
4	'. . . example . . . '	1	1	2
5	'. . . example . . . '	1	1	2

Table 3.4: Unwanted Sexual Attention Tweet Examples

Regarding instances where the harasser claims to be merely flirting in such an environment, the comments or questions made would fall under this category of unwanted sexual attention if the victim has communicated their lack of interest or if the comments or questions were egregiously sexual in nature. 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 with the exception of 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 forms of contact as harassment. This could be because the socially acceptable contact is coming from someone who has previously behaved inappropriately or another valid reason that could make the victim uncomfortable. To categorize these behaviors accurately, the standard of evaluation again is considering the opinion of what an unaffiliated, objective third party would interpret 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 be considered within his category of sexual categorical remarks, which is aggregated into patronizing behavior. However, within this thesis, 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 sexual assault, possibly because such an instance would be considered under different criminal boundaries. The likelihood of witnessing sexual assault is extremely small in general, but even more so with regards to professional environments. Because this thesis needs to consider victims who witnessed assault happening to a peer, bystander

harassment is classified as unwanted sexual attention. This 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, any explicitly sexual behavior that occurred within a workplace or academic environment (which falls within Title VII and Title XI regulations), any language or interaction that is beyond what is reasonably accepted as flirting, bystander harassment, and ambiguous forms of conventionally accepted physical contact.

3.3.3 Category C: 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 commonly accepted forms of contact addressed within the previous category. Predatory behaviors include attempted or successful rape, sexual assault and battery, and *quid pro quo* arrangements.

ID	Original Tweet Text	Relevant	Stance	Harassment Category
1	'...example ...'	1	1	3
2	'...example ...'	1	1	3
3	'...example ...'	1	1	3

Table 3.5: Predatory Tweet Examples

Some nuances regarding authority exist within the predatory behavior category. If the harasser is exhibiting behavior generally classified as unwanted sexual attention but stands in a position of authority over the victim (ex. boss or teacher), the behavior considered predatory due to the nature of the relationship between the harasser and victim. Any and all forms of sexual comments, remarks, questions, and contacts between an adult and a minor is considered predatory, even if the behavior would be considered patronizing or unwanted sexual attention if the two parties were of legal age.

3.3.4 Category D: 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." Anything that is clearly 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 within this category. Tweets that were classified as having a supportive stance but did not claim victimhood are left blank in this check.

ID	Original Tweet Text	Relevant	Stance	Harassment Category
1	'...example ...'	1	1	4
2	'...example ...'	1	1	4
3	'...example ...'	1	1	4

Table 3.6: Not Enough Context Tweet Examples

Assembling Gruber's 11 types of sexual harassment alongside 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 he or she experienced. As these users' individual accounts of their experiences have not necessarily been reviewed (such as by a human resources department, the police, etc.), this is the best approach to this problem. It also avoids claiming a degree of accuracy that cannot be determined with confidence. If there is not enough context to confidently classify a tweet within one of these three categories, it is labeled as not having enough context.

3.4 Categorization Rules

Some rules exist to accommodate the parameters of the algorithm. This list covers all categorization decisions that are not obvious or 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. Do 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 behavior.
7. If the tweet only says "#MeToo", give the author the benefit of the doubt and assume that they have indeed experienced a form of sexual harassment or assault but without enough context.
8. Tweets that *imply* a personal experience but do not claim one are categorized as having a supportive stance but left blank when determining the type of harassment. The only exceptions are tweets that only contain "#MeToo"
9. If the categorization is difficult to make, the decision should be made by giving the benefit of the doubt to the author of the tweet. This decision is based on the research suggesting that victims of sexual harassment experience the negative physical consequences (anxiety, insomnia, etc.) regardless of whether or not they can accurately identify a behavior as being sexual harassment or not.

CHAPTER 4

IMPLEMENTATION

4.1 User Requirements

The objective of this program is to classify tweets using the #MeToo hashtag, and the core is to output a categorization that is as accurate as possible. The program will attach three levels of classification to each tweet. In detail, the classification requirements are:

1) The first check is asking whether or not the tweet is relevant to the #MeToo movement and awarded either a 1 or 2 (tweets in languages other than English are eliminated from the dataset).

2) The second check determines whether or not the tweet is in support of the movement, against the movement, or questioning the movement and awarded a 1, 2, or 3 accordingly. If there is not enough information to make this determination, this variable is left blank. If the tweet is expressing an experience, at this stage it should be awarded a 1.

3) The third check attempts to categorize the specific type of harassment. If there is not enough context, or if the tweet is merely #MeToo without other context, it is categorized with a D. If the tweet is expressing an experience with sexist or patronizing behavior, it is marked A. If the tweet is expressing an experience with unwanted sexual attention, it is marked B. If the tweet is expressing an experience with predatory behavior, it is marked C.

There are more factors considered when categorizing a tweet. Only text within the tweet will be reviewed with the use of this program; no images, or links to text on other sites will be included. Tweets that do not contain enough context to properly categorize them will have the categorizations left blank accordingly and stop at whatever check they reached. Words contained within other hashtags will be included when evaluating context, but the words contained within username mentions will not be considered. Experiences that occurred when the victim was a minor will be categorized as predatory, and ambiguous descriptions using the #MeToo hashtag are categorized as not having enough context. Some samples are provided below:

will be a figure one day

4.2 Development, Design, and Deployment

This project is written with the Python programming language, using the NLTK, spaCy, and and pandas libraries for language processing and the scikit-learn library for classification algorithms. Jupityr will be used to handle the preprocessing and exporting of the dataset. Matplotlib has been used be used for basic visualizations using the completed dataset. Python was chosen for its flexibility from the high quantity of libraries available to assist in natural language processing and for supervised machine learning.

Dr. Hassan has already collected the dataset, which amasses to approximately half a million tweets of data. Currently, I have manually labeled 1,000 of the tweets according to the proper categorization with the intention to raise that number up to 10,000 manually labeled tweets for the purposes of supervised machine learning. To verify the integrity of my categorization definitions, multiple volunteers of both genders are categorizing the dataset alongside me. The project is hosted on my personal GitHub account under the username “claireballoon”.

While the program can categorize a local dataset simply by running it in Python3 and outputting a file with proper categorization, further utility is desired from this project. The immediate step after completion of the categorization is to create a simple online interface for researchers to use to categorize their data without having to install Python and categorize them locally. Accomplish our core goals necessitates only the installation of Python3 and the appropriate libraries for those wishing to categorize their own datasets of tweets. As Jupityr can export to PDF or LaTeX,

no further environment specifications are needed for researchers looking to use the results of this categorization in their data at this time.

4.3 Architecture

The overall architecture is very simple. It requires the raw data, the properly formatted dataset, the natural language processing and machine-learning algorithms, the formatted categorization results, and the visualizations that can be made from those results. The process flows linearly.

//todo - figure

4.4 User Interfaces

At this time, the implementation of the project only requires use of Python, the command line, and a text file with the dataset of tweets formatted properly (ID of the tweet and the text). The program will then output a new file with the categorizations and possibly some visualizations. In addition to creating visualizations with the user's data, the master dataset would have utility in being posted publicly so that others researching sexual harassment can use the categorized data in their work. The easy, next step is to create an online interface that doesn't require the user to categorize their dataset locally.

4.5 Testing and Integration

The manual labeling of tweets will be used to verify whether or not the program has successfully categorized the tweets. By manually categorizing a large enough dataset (goal set for 10,000), it can be determined with a high degree of accuracy how successful the algorithm and project will be.

CHAPTER 5

RESULTS

none because I'm a failure

5.1 Accuracy

//todo

5.2 Self Evaluation

//todo

CHAPTER 6

SUPPORTING MATERIALS

Research and Application

6.1 Survey

blah blah

6.2 Findings

blah blah

CHAPTER 7

CONCLUSION

None becuase I suck