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

Claire Hubacek

Abstract

> ”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 exhibit the magnitude of sexual assault victims. 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. This movement has grown so large that the hashtag #MeToo is now being used by people who are in support of the movement but not victims themselves victims, antagonists of the movement, and by victims communicating a personal experience. Without knowing the semantics of the tweet itself, the existing analysis is not as useful as it could be if the tweets had been categorized (specifically analyzing the tweets of victims experiences to find trends). This project involves the development of a program that can process these tweets and categorize them into one of the the categories described above (and further categorize the type of sexual harassment if possible). By filtering out tweets that are based upon discussion, researchers would be privy to more useful analysis of the data by working exclusively with tweets that express a personal experience with sexual harassment.

The minimum viable product version of this project is a program that can reliably and accurately categorize a tweet that contains #MeToo as being either a personal experience, in support of the movement, or against it. This project will primarily require knowledge natural language processing (NLP) techniques and tools (CoreNLP, NLTK, spacy, etc.). Additionally, it will require knowledge of machine learning algorithms and an analysis of a very large dataset.

The depth of the applicability of this project will depend upon survey results to evaluate how people might want to take advantage of such a tool in contexts beyond research. The program could potentially identify tweets on Twitter as they are made, and if the tweet is revealing a personal experience, have utility in that it could inform the appropriate organization (police, school, womens shelter, etc.) of the assault/harassment, connect the victim with another victim to talk, or even just archive the Tweet so that evidence of it exists even after it has been deleted.

Literature Review

Categorizing different types of sexual harassment is a problem that researchers began making significant contributions to in the 1980’s because it was in year [MISSING] that the United States first recognized sexual harassment as a crime in the workplace. Initially, the Title VII of the Civil Rights Act of 1964 made it illegal to discriminate based upon gender but it wasn’t until several women in the late 1970’s pushed to sue their employers for coercing them into sexual acts using Title VII provisions that sexual harassment became legally recognized. The Supreme Court ruled in favor of the victim in the 1986 case *Meritor Savings Bank v. Vinson*, which effectively established *quid pro quo* acts as a form of sexual harassment. Now, sexual harassment falls under the deomain of the the Equal Employment Opportunity Commission (EEOC), and the accepted definitions categorize sexual harassment as either being *quid pro quo* acts or behaviors that contribute to a ”Hostile Work Environment.”

Under U.S. law, *quid pro quo*, also referred to as sexual bribery, includes attempted and actual pursuits of a sexual nature against a person in a professional or academic environment, including coercion. The much more common category of sexual harassment is that of actions that contribute to a ”hostile work environment”. Smaller behaviors, such as offhand remarks, teasing or banter, and personal questions are not explicitly banned as acts of sexual harassment. However, if these acts increase with severity or frequency, they would then fall under the ”hostile work environment” category of sexual harassment and by law must be evaluated comprehensively. When considering behaviors that do not take place in a professional setting, a similar evaluation philosophy is observed until the behaviors can be considered with context as falling under harassment, stalking, cyberstalking, or sexual assault laws.

As the courts began to address more and more claims of alleged sexual harassment, a proper categorization became essential for the varying types of sexual harassment in order to appropriately assess the degree and severity of the acts. From both a legal and social research perspective, the lack of consistency in categorizations caused problems when trying to compare and use instances or research from one context as a guide in another. In 1992, James E. Gruber made a significant contribution to categorizing sexual harassment by setting three overarching categories and 3-

4 subtypes. 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 non-solicitory nature directed either to a woman (teasing, 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, 1992). Gruber originally wrote these categories based upon reviews of sexual harassment that only included female victims and male harassers.

Despite the one-sided gender limitations of Gruber's work in 1992, these mutually exclusive categories which create a flow in how certain behaviors can contribute to a hostile work environment have continued to be used in many more recent works. In a 2005 review of the past, present, and future directions of increasing 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 (Murrell and James, 2002). 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 (Ferguson et al., 2005).

In order to properly categorize a behavior as sexual harassment, the victim (including bystanders and witnesses) must be able to recognize it as such. The victim perception is integral not just for legal categorization but also to understand the degree and severity of harm on those who are exposed to the behavior. 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 (Schneider et al., 1997) 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 (MacMillan et al., 2000). 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. created a simpler categorization 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 in developing a more scalable questionnaire for surveying the frequency of sexual harassment (Fitzgerald et al., 1995). 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” (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).

In developing a program that can contextualize tweets that use the #MeToo hashtag as expressing an experience of sexual harassment or not, I have decided to use the categories of *patronizing*, *taunting*, and *predatory*. Twitter's character limit prohibits users from demonstrating an adequate context to classify a tweet accordingly, and the personal bias of the author of each tweet could result in miscategorization. Using patronizing, taunting, and predatory categories better accommodates the gray area created by each Twitter user's interpretation of the behavior (as they have not been reviewed by a human resources department, the police, etc. in many cases).

Woman, Action, and the Media (WAM!) created a team in 2015 to assist Twitter in evaluating harassment that took place within the website. While this tool does not look to analyze whether or not a tweet may be of a harassing nature, this report afforded insight into Twitter's system that could potentially have use in other contexts. 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. More importantly, 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). It is not unheard of for sexual harassment in the context of stalking and harassing to persist in impersonal media; while in the past this was phone calls, cards, or texts, modern day harassment includes social media messaging. This report was useful in showing that there might be utility in archiving tweets that use the #MeToo hashtag as soon as they can be identified.

References

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

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, 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.

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

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.

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

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.