Methods of Using Real-time Social Media Technologies for Examining Feminism and Social Issues

# Abstract

*Objective:* Advances in social media analysis and “big data” permit the study of social issues generally, and feminism specifically, and how they are communicated and shaped via social networks. This pilot study uses Twitter as a means to gauge current views on modern feminism as well as exploring the feasibility of using Twitter as a medium to study how social views are formed and influenced.

*Methods:* In 2015, we collected tweets (N= 30,597) containing the word “feminism” or “feminist”. By hand-coding tweets as positive/neutral or negative/hostile to feminism, and then extrapolating the model over the whole corpus of tweets captured, we created both semi-supervised and unsupervised models of machine learning to analyze the data.

*Results:* We found that Twitter is more than 52% negative on feminism and that machine learning clusters of Twitter data mirror existing issues and sentiments. Further study is needed to determine the relationship between current events and perceptions of feminism in real-time.

*Conclusion:* Our research suggests that analysis of social media data constitutes a feasible methodology to study opinions of social issues and how they are formed.

# Introduction

Increasing awareness of social factors, like gender inequality, have been linked to many societal, health, and environmental factors. Views on gender equality and greater education on those issues leads to positive social outcomes. For example, research shows that education on gender equality leads to positive outcomes. Improved sexual education, including education on gender equality, has been shown to lead to more positive health outcomes (Haberland, 2015). Increased information concerning perceptions of gender inequality can improve work, health and environmental outcomes.

New technology provides us with novel ways to monitor public opinion and sentiment on a variety of issues. It allows us to study traditionally marginalized, victimized, or minority populations to gain a greater understanding of the factors that affect other populations (Krueger and Young, 2015). The hugely disparate number of topics discussed via social media can provide insight into the people who choose to discuss feminism and why. Social networking technologies have recently been used to study HIV prevention, and as a tool for recruitment and intervention (Young et al., 2014). The broad use of social networking technologies makes them an ideal tool for studying social behavior, social attitudes, and how those behaviors and attitudes they can be changed. Big Data is emerging as an increasingly valid way to gain insight into psychology, public health, etc (Young et al., 2014). However, little research has been conducted on the use of these technologies to study social factors, such as gender inequality and feminism, making it important to evaluate the feasibility of such an approach. Establishing this viability is an important first step in being able to enact positive social change, potentially leading to better outcomes.

Twitter, a popular social networking site relaying millions of up to 140 character real-time communications, can potentially be used to make inferences about current social beliefs and is a valuable tool to see how time and current events change or shape social beliefs. Twitter is a massive and growing social networking technology which allows participants to send short, public, real-time communications or “tweets”. With more than 300 million monthly active users and 500 million tweets sent per day, Twitter provides an unprecedented data source for research. Twitter provides public access to a raw form of this data through an Advanced Programing Interface (API). People who have opinions on gender equality might tweet their opinions to their followers, catalyzed by events in their personal lives or current events, stating or demonstrating their attitudes, or providing revelatory clues.

The study sought to investigate whether machine learning on social media, as a tool to understand views of feminism and how they are formed, can be applied to gain greater insight on a complex and nuanced issue. Specially, this study seeks to determine 1) whether sentiments on feminism can be successfully and accurately extracted from real-time social networking data, 2) the prevalence, content, and context of these sentiments, and 3) the feasibility of using feminism-related social media conversations in real time as a method of monitoring public perception of feminism and how current events change and shape people’s opinions thereof.

# Methods

We collected 30,597 tweets using the Twitter Advanced Programming Interface (API) between February 12 2015 and February 6, 2015. We used the Twitter “garden hose” method of collecting tweets, which provides a random sample of approximately 1% of all tweets. Tweets collected through the garden hose are available in real-time, the data are updated as tweets are sent through the service. A variety of metadata are available along with the tweet text (i.e. number of followers and time of tweet) but, for simplicity and space efficiency, only text and language information were captured for use in this study. Data were filtered to include only tweets in English and tweets that contained either of the words, “feminism” or “feminist.”

# Analysis

We applied two techniques to gain greater insight into the corpus of tweets collected: semi-supervised and unsupervised learning. Supervised learning uses only labeled data to find a function between input data and output data. Semi-supervised learning builds on that model but allows for unlabeled data. A small set of labeled data are used in semi-supervised machine learning to build a model to then accurately label the remaining unlabeled data. Unsupervised learning, addresses the task of finding hidden structure in unlabeled data and is useful for exploratory analysis of data, as well as finding clusters or themes within data.

## Semi-Supervised

We hand-coded a subsection of tweets (217 of each) as positive/neutral or negative. Hand-coding is preferable to using current preexisting sentiment analysis software because negative words present in a tweet does not necessarily indicate a negative sentiment about feminism. Using the codified tweets, we constructed a variety of machine learning models to extrapolate valence of sentiment across the remaining tweets to gain a single number or “stock market price” for feminism. We used the sklearn library of Python and chose Linear Regression and Linear SVC as models most likely to perform well with binary data. We also tried using Multinomial Naive Bayes and Bernoulli Naive Bayes but ultimately rejected those models for their lower levels of precision and accuracy.

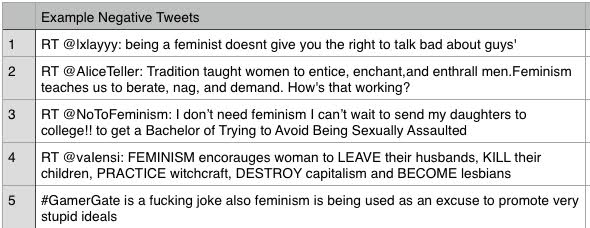
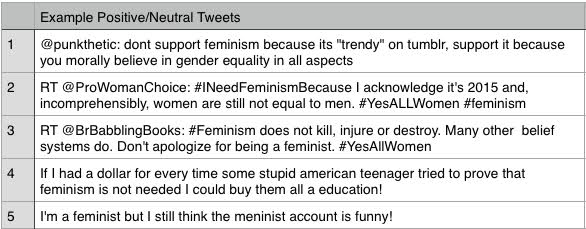
## Unsupervised

We ran a variety of machine learning clustering algorithms to find recurrent themes and patterns within the corpus of tweets. Using the gensim Python library, we found the most republished tweets as well as constructing a Latent Dirichlet allocation (LDA) model yielding 50 topics/clusters within the data. We also used the sklearn MiniBatchKMeans model to cluster data into 10 larger groups.

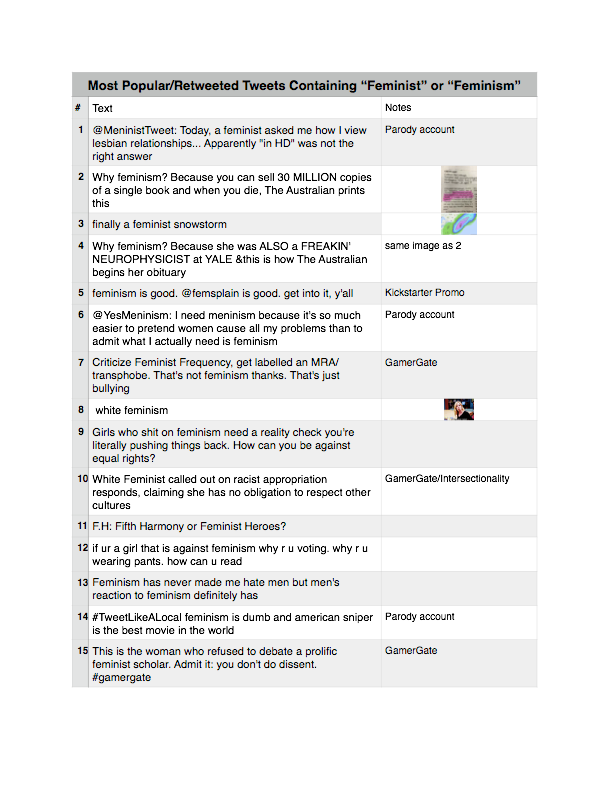
# Results



## Semi-Supervised

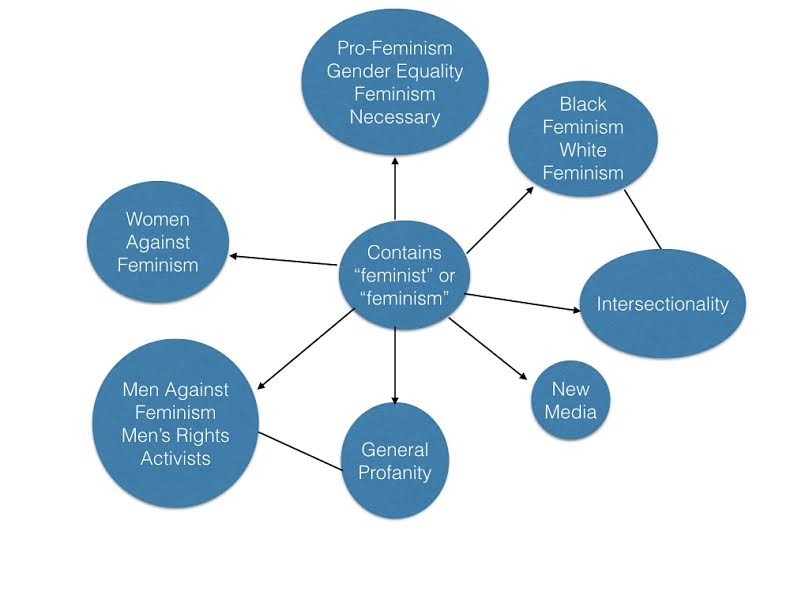
Twitter is between 52%, according to the Linear SVC model, and 54%, according to the Logistic Regression model, negatively valenced on feminism.

## Unsupervised

The unsupervised analysis yielded a variety of outputs. We collected the most retweeted tweets as a way to study popular, viral, or loudest trends.

We found that clusters within the data mirror various current trends in modern feminism. Topics of intersectionality (the study of intersections between forms or systems of oppression or discrimination), linguistic semantics, new media, the need for feminism, and Men’s Right’s Activism were the main recurring themes we found through the data set.

## Discussion

This study represents a novel approach to how social issues might be studied by using real-time social media data to extract, detect, and potentially monitor how attitudes about feminism and other social issues are formed and evolve. Our research supports the feasibility of using social media data to analyze, understand, and inform potential action or intervention regarding public opinion of social issues. Having the ability to use social media as a mirror for social issues provides valuable insights that can be otherwise impossible to attain from hard-to-reach or minority populations. Because of the growing amount of social media data and increasingly intelligent techniques for analyzing and understanding these data, researchers and progressive activists will soon be able to build upon these methods to more accurately monitor public opinion and how social media can shape or evolve these opinions.

This study was limited by a number of factors, primarily relating to the financial and technical limitations under which the study was done. Because of storage constraints, the script to capture tweets was run once each day for seven consecutive days. We ran the script at a different cycling 5 hour session each day in an attempt to correct for any time-zone or geographic bias in the data. Further research will ideally allow for a constantly running script to gather real-time data and compare it to historical data to see how current events or social changes impact the views of feminism as reflected via social networking. The hand-coding of the data provided other limitations including only one researcher being tasked with all of the hand-coding. Human coding allows for greater context and humor being identified and controlled for, as opposed to out-of-the-box sentiment analysis which cannot do this. Further research will include multiple researchers coding sections of the data for valence with control for any researcher bias. Additionally, because of time constraints, fewer tweets were coded than would be desirable. With a larger train and test set, the machine learning models will be more precise and more accurate. While social media presents an exciting new direction for social research, the demographic use of social media is not evenly distributed across the population, tending to skew both younger and predominantly male on Twitter in particular. Controlling for population bias will be an important improvement to future social media research.

# Conclusion

Results from this study suggest that it is feasible to use real-time social networking technologies to identify and monitor views about feminism and other social issues. The data collected mirror current clusters of thought within and about feminism. This study was designed to provide a call for future research to understand the potential of this approach for the social sciences, for social issues, and to impact positive change. Further research will build upon this feasibility to improve and refine methods of using real-time social media data for understanding and impacting social views.

# Citations

Haberland, Nicole A. "The Case for Addressing Gender and Power in Sexuality and HIV Education: A Comprehensive Review of Evaluation Studies." IPSRH International Perspectives on Sexual and Reproductive Health 41.01 (2015): 031-42.

Krueger, Evan A., and Sean D. Young. "Twitter: A Novel Tool for Studying the Health and Social Needs of Transgender Communities." JMIR Mental Health 2.2 (2015):

Young, Sean D., Caitlin Rivers, and Bryan Lewis. "Methods of Using Real-time Social Media Technologies for Detection and Remote Monitoring of HIV Outcomes." Preventive Medicine 63 (2014): 112-15.