Abstract

Twitter is a microblogging website where users read, write, and retweet millions of 140 character messages every day on a wide variety of topics from the inane to the weighty. This pilot study uses Twitter as a means to measure current views of modern feminism and how Twitter can be used to monitor how views on social issues are shaped or changed. By hand coding tweets as positive/neutral or negative/hostile to feminism and then extrapolating the model over the whole corpus of tweets captured, I found that Twitter is more than 52% negative on feminism and that LDA found clusters that mimic existing issues and veins of sentiment. Further study is needed to study the relationship between current events and perceptions of feminism in real-time.

Intro

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, show studies about how learning about bias reduces it. For example, research shoes that education on gender equality leads to positive outcomes in wage disparity , citation. Increased information about perceptions about gender inequality can improve work, health and environmental outcomes.

New tech provides us with new ways to monitor public opinion and sentiment on a variety of issues. It also allows us to study traditionally marginalized, victimized, or minority populations to gain a greater understanding of the factors that affect other populations. cite him and the study of trans people. Wider number of topics discussed via social media and/or gain insight to the people who choose to discuss feminism. Social networking technologies have recently been used to study HIV prevention, as well as as a tool for recruitment and intervention. Because people so often and so broadly make use of social networking technologies, it is an ideal tool for studying social behavior, social attitudes, and how they can be changed/shaped/impacted. Big data is emerging as a more and more robustly validated way to gain insight into psychology, public health, etc. However, little research conducted on using these technologies to study social factors, such as gender inequality and feminism, making it important to evaluated the feasibility of such an approach. Establishing this feasibility is an important first step in being able to enact positive social change for better outcomes.

Using Twitter, a popular social networking site relaying millions of 140 character real-time communications, to make inferences about current social beliefs is a valuable tool to be able to build on to be able 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 “tweets” communications. 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 Advances 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, showing or telling of their attitudes.

This study sought to invest whether social media as a tool to understand views of feminism and how they are formed and how machine learning 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 to feminism and how current events change and shape people’s opinions to these issues.

Do we need a paragraph as a refresher on current feminism/issues? like how 2014 was amazing and gave us the Hunger Games and Taylor Swift & Beyonce “coming out” as feminists, while 2015 had 50 Shades of Grey, GamerGate fall out, Hackeye slut shaming Black Widow, and Men’s Right’s Activists now boycotting Mad Max? That’s probably more a humanities paper.

Methods

We collected 30,000 (put in exact #) tweets using the free Twitter Advanced Programming Interface (API) between Date 1 and Date 2. 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 is 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 the word, “feminism” or “feminist.”

Analysis? Your paper has an Analysis section so I wasn’t sure if I oughta break it up.

We applied two techniques to gain greater insight into the corpus of tweets collected, both semi-supervised and unsupervised learning. In the context of Machine Learning, semi-supervised learning is a subsection of supervised learning techniques. Supervised learning uses only labeled data to find a function between an input data and output data and has existed within the scientific community since the Iris flower data set was introduced my Sir Ronald Fisher (<http://en.wikipedia.org/wiki/Iris_flower_data_set>) showing that the features of an iris, such as width of petals, can accurately distinguish one species from another. Semi-supervised learning builds on that model but allows for unlabeled data. A small set of labeled data is used in semi-supervised machine learning to build model to then accurately label the remaining unlabeled data. Unsupervised learning, in machine 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.

For the semi-supervised machine learn, 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 being present in a tweet does not necessarily indicate a negative sentiment about feminism ("Why do people use the word feminist like it's a bad thing that irritates me so much”). Using the codified tweets, we constructed a variety of machine learning models to extrapolate valence of sentiment across the 26000 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. For the unsupervised learning, we ran a variety of machine learning clustering algorithms to find recurrent themes and patterns within the 30,000 corpus of tweets. Using the gensim Python library, we found the most retweeted, republished by people who are not the original author, tweets and constructed and Latent Dirichlet allocation (LDA) model yielding 50 topics/clusters within the data, and used the sklearn MiniBatchKMeans model to cluster data into 10 larger groups.

Results

According to our semi-supervised analysis, Twitter is between 52%, according to the Linear SVC model, and 54%, according to the Logistic Regression model, negatively valenced on feminism.

FIGURE OF STOCK MARKET PRICE

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

FIGURE OF TOP TWEETS

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.

FIGURE OF CLUSTERS

Discussion

First new usage of tech to study feminism

Having ability to use social media as a mirror for social issues provides valuable insights that can be otherwise impossible to attain from populations.

Limited by a number of factors.

Handcoding:

* Pros: Some context and humor controlled for by human coding
* Cons: Time consuming so unable to code more of them, low context, uncaught irony, only me coding,
* Grade level of negative tweets consistently measured as at least a grade level below neutral/pro-feminist tweets
* Higher levels of profanity and misspelling in negative tweets

Improvements Required to Progress Beyond Pilot

* Constrained by small codified samples and the need to have both test and train sub-sets, accuracy could be improved by increasing number of codified tweets, ideally codified by multiple people to control for individual bias.
  + Sample size is 217 negative and 217 positive tweets to 30k uncoded tweets
* Goal: Have the script running constantly to display a stock market value of feminism on Twitter and how real time events affect it.
  + Be able to monitor twitter-verse for how events impact the public’s view on feminism, what causes backlashes, measure impact of Beyonce, etc