Technical Report of Twitter and Feminism Pilot Study

1. Description of problem and hypothesis:
   Mostly an exploratory/descriptive study but with the hypothesis that Twitter seems more negative on feminism than positive. To explore the details and demographics of the people who choose to discuss feminism on Twitter.
2. Detailed description your data set.
   30597 tweets were captured by using the Tweepy StreamListener (named TwitterBotFinal.ipynb in my github), 217 positive and 217 negative tweets were hand coded by me. I chose a random page of the mongoDB online interface (hosted through compose.io) and coded the most clear examples on that page. I would've liked to code more for the test train set but the free trial ended and coding them by hand through the localhost mongodb command line created a time problem.
3. How did you decide what features to use in your analysis?
   I decided to use simple valence, 0 is neutral or positive, 1 is negative or hate speech. In the unsupervised learning, I also did a grade level analysis and used different cluster models.
4. What challenges did you face in terms of obtaining and organizing the data? What did you learn from the initial exploration phase?
   Obtaining the data presented challenges because of the shortness of tweets and the short lived nature of their relevance. In a next phase, I would also capture retweet data and only capture unique tweets, but weight them by their retweet number, instead of having so many identical tweets. Coding the tweets presented some problems in that there isn't a lot of context for humor. Are people making fun of feminism because they are against it or because they thought of a funny joke to share with their feminist friends? Mostly people's intentions were clear but there is insufficient context to be sure for each one.
5. Describe what kinds of statistical methods you used, and perhaps others you considered but did not use, and how you decided what to use.
   - Semi-Supervised Learning : Trained on codified subsample and used Multinomial Naive Bayes, Logistic Regression, Linear SVC and Bernoulli Naive Bayes in the sklearn library on the data. Logistic and Linear SVC yielded the more accurate results. Results could be improved with more codified training data.(TwitterSemiSupervised.ipynb)
   - Un-Supervised Learning: Used gensim library to find most common tweets and construct a LDA model yielding 50 topics/clusterings within the data to get at smaller currents and clusters within the data (TwitterGensim.ipynb). Used sklearn MiniBatchKMeans to cluster data into 10 main groups to pick up on larger trends in the data. (TwitterMiniBatchKMeans.ipynb)Used textstat to determine grade level of average negative or positive tweet.(TwitterGradeLevel.ipynb)
   - I considered using the NLTK python library but rejected it as the code yielded no interesting or informative data. (TweetAnalysisNLTK.ipynb)
6. What business applications do your findings have?
   My findings are relevant for research into social issues, psychology and sociology. With a broader implementation and more resources to code tweets, a viable picture of internet feminism could be attained for use in research, academia, marketing or other applications.
7. Describe the implementation plan in detail from the ingesting of data to how end-users would access it.

This project was inspired, in part, by http://www.nohomophobes.com/, run by Institute for Sexual Minority Studies and Services, University of Alberta. With a transfer from localhost server to a remote MongoDB server and running the Tweepy streamer constantly, there could be live updating dashboard like NoHomophobes providing insight into what affects Twitter's view of feminism. That information could then be used to analysis the place of feminism in America and predict the public's response to various events. With a more robustly trained Logistic Regression or Linear SVC, a stock market price for feminism could be created to show what percentage approval feminism is at and how recent events has impacted those numbers. This information could be used by NGOs and other groups for direct usage, like how well #HeforShe is perceived by the public, or broader uses like research into the demographics of modern feminism or predictions on the creation of fourth wave feminism.