Tracking #metoo on Twitter to Predict Engagement

Ana Tarano (atarano)
Dana Murphy (dkm0713)

Summary:

#metoo was created as a way for women and men to share their experiences of sexual harassment and assault[1]. We analyzed the language in 3,750 tweets in order to predict levels of engagement with that tweet. Using Naïve Bayes and SVM algorithms, we are able to predict a #metoo tweet's engagement with 90% accuracy.

Data:

Our data came from Talkwalker, an archive of tweets with information including tweet content and number of retweets[2]. We collected 3,750 english #metoo tweets from 11/12/2017 to 12/10/2017 filtered by 3 categories: engagement, potential reach (i.e. number of followers), and recentness.

We processed the data as a matrix of word occurrences, with each row as a tweet and each column as a word in our lexicon. We calculated engagement as whether a tweet passed a certain threshold of retweets.

Features:

Our feature set consisted of the number of words occurrences in a given tweet. To generate our lexicon, we collected words from a subset of 750 #metoo tweets, removed noise, applied Porter stemming[3], and removed stop words, creating a feature set of about 5,000 stems, reducing the original lexicon by over 25%.

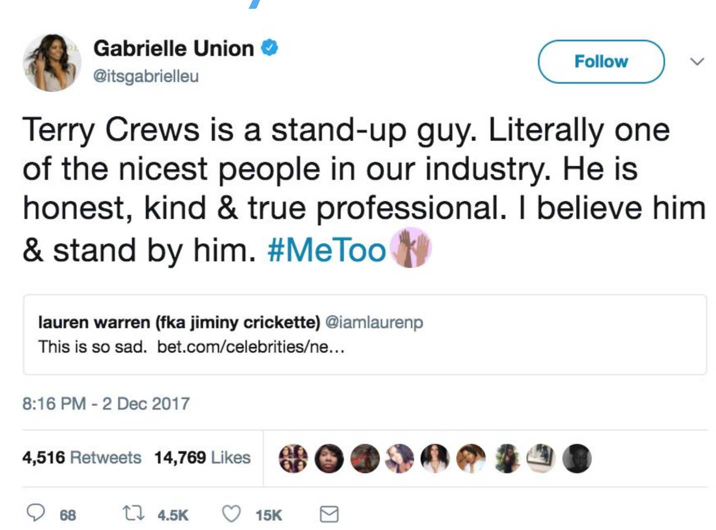
The features included directed accounts (i.e. @Alyssa_Milano) and hashtags (i.e. #himthough).

Results:

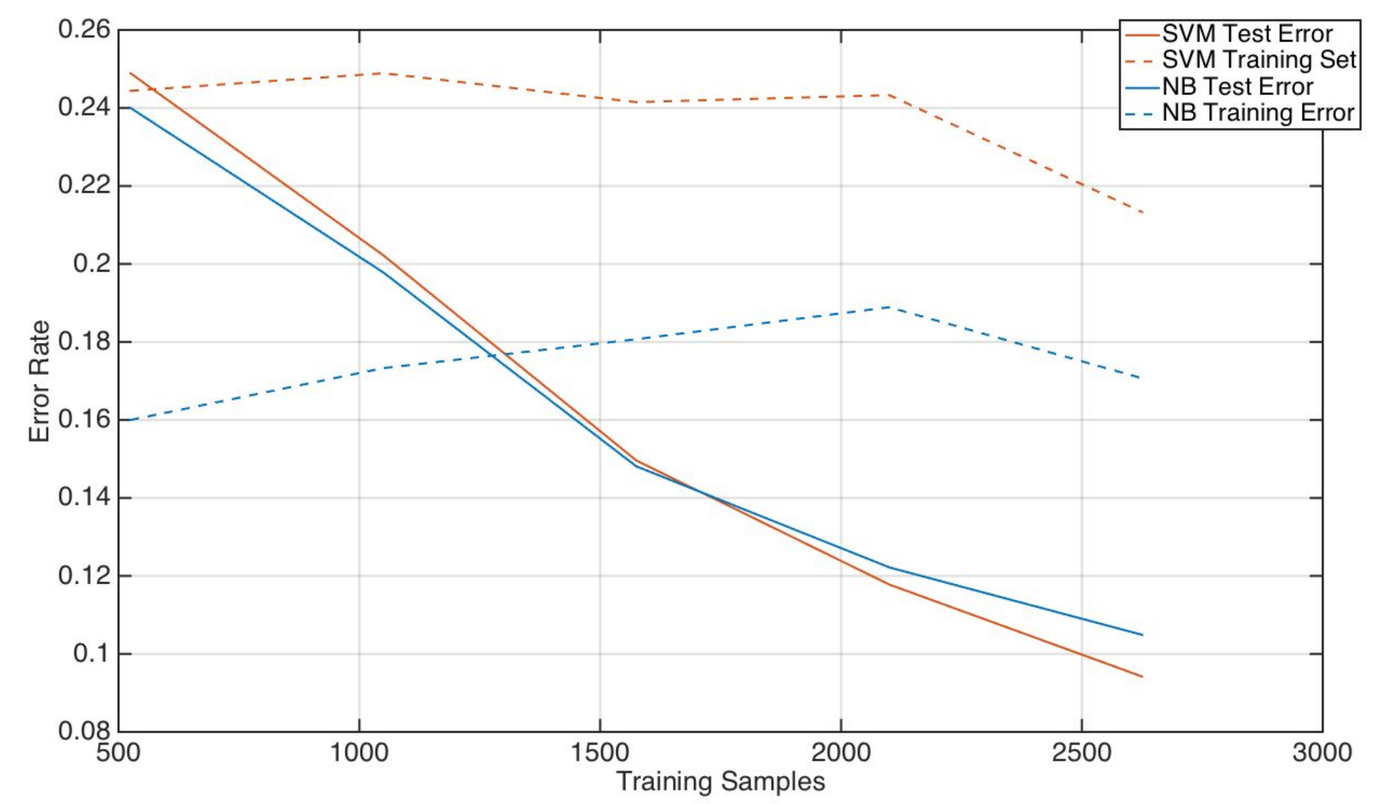
Accuracy:

	Training	Test
Naïve Bayes	92.11%	89.78%
SVM Model	88.99%	90.58%

Correctly Predicted Tweet



Error Quantification:



Our five most indicative tokens were sen, risen, profess, and restrict, and @youtube. One example of a correct prediction is also shown, which uses one of the most indicative stems: "profess" for "professional."

Discussion:

The most surprising aspects of our project stemmed from how the movement itself changed and the relative importance of various factors in our project. For example, even over the span about a month, the conversation surrounding #metoo shifted several times, leading to considerable differences in the content of our data sets. We suspect this at least partially contributed to the lower accuracy. This may have also contributed to a considerable level of variance within our code- even different permutations created notable differences in accuracy. The high variance is likely also influenced by the fact that training requires more examples and the data comes from 3 different categories.

Models:

For our models, we used a multinomial Naïve Bayes algorithm and an SVM. Our classification is based on whether the number of retweets passes a threshold of retweets, calculated as the average number of retweets from random sample of the training data.

NB performs significantly better than SVM with regards to the training error, so we used NB's training error as our baseline.

Future Work:

In the future we would like examine how this analysis of engagement could be applied to other social media movements besides #metoo and acquire more examples to resolve the high variance issue.

We would also like to use ICA to better determine which of our features are most relevant.

Citations:

1. http://www.cnn.com/2017/10/17/us/me-too-tarana-burke-origin-trnd/index.html
2. https://www.talkwalker.com
3. people.cs.georgetown.edu/~nazli/classes/ir-Slides/Preprocessing-13.pdf