Tracking #metoo on Twitter to Predict Engagement in the Movement

Ana Tarano (atarano) and Dana Murphy (dkm0713)

Abstract:

In the past few months, the social movement #metoo has garnered incredible social reach and impact. The goal of our project was to better understand and predict what types of tweet receive particularly high attention and engagement. In doing so, we can provide insight into the potential reach future social media movements by understanding what content is likely to reach the most people. We examined 3,750 tweets within the #metoo movement; by comparing the word occurrences within the content of the tweets, we were able to predict whether a tweet would be retweeted above a mean threshold with 90% accuracy.

Introduction:

#metoo is a Twitter social movement created as a way for women and men to share their experiences of sexual assault and harassment, in an effort to provide solidarity to survivors as well as demonstrate how prevalent and underreported sexual harassment is in society. While #metoo was initiated in 2006 by Tarana Burke, it has resurged recently due to high-profile Hollywood stars, such as Alyssa Milano, coming forward with the harassment and assault they faced in their industry (Santiago, 2017). The #metoo movement has since unprecedented social reach and impact. Statehouses in Illinois, California, Oregon, and Rhode Island have moved to pass additional sexual assault laws in response to the movement (Tareen, 2017), and Time Magazine recently named the individuals who started the #metoo movement as 'Person of the Year 2017' (Zacharek, 2017).

The goal of our project was to analyze and predict what aspects of given tweets in the #metoo movement led to their remarkable level of engagement. Namely, we examined the content of given tweets in the #metoo movement to predict whether other users would engage with the tweet through retweeting. The input to our algorithm was the number of word occurrences in the content of a given tweet. We then use multinomial Naïve Bayes and an SVM to predict whether or not a given tweet passed a threshold of retweets.

We did not find other machine learning research focused around a specific social movement such as #metoo. However, there is a precedent in research analyzing twitter retweets as well as using Naïve Bayes and Support Vector Machines for text classification and analyzing Twitter data.

One paper which also focused on retweet analysis was Suh et al. (2010), which used Principal Component Analysis to compare which features of a tweet were given the strongest weights and were thus more relevant to a tweet becoming retweeted. While the paper was very thorough in its analysis, it ultimately did not use this data to predict whether tweets would be retweeted. Another paper, Petrovic et al. sought to apply prediction to twitter retweeting using a passive-aggressive algorithm, which they then compared to human accuracy (2011). While their potential feature set was quite extensive, including several social and tweet features, their choice of the PA algorithm was less thoughtful, as they believed that "the choice of the algorithm is not crucial and any approach that is feature-based would be appropriate" (2011). Ultimately their algorithm predicted a tweet's reach with an accuracy better than chance, but worse than human prediction.

In terms of approach, Irani et al. used text classification algorithms, such as Naïve Bayes, to predict 'trend-stuffing', or users spamming trending topics on Twitter with unrelated content in order to gain more visibility (2010). One particularly clever approach of theirs was using Information Gain to determine strongest features, reducing the number of features by almost 98% while keeping the error rate essentially the same for Naïve Bayes. While we would like to implement a similar approach to feature reduction given more time, at our current scale it was not as relevant. Another paper which used Naïve Bayes and SVM to process twitter data was Go et al., which sought to predict positive and negative sentiment in tweets (2009). Rather than manually marking tweets as negative or positive, the algorithm used smiling and frowning emoticons as indicators of sentiment. While this is an interesting solution to the potential bottleneck of manual classification, it greatly limits the dataset to a small subset of tweets. Another paper, Joachims, discusses the potential gains of text classification using SVMs over algorithms such as Naive Bayes (1998). At the time, this paper referred to state-of-the-art research. However, we ultimately did not find such drastic differences in our own implementations of Naive Bayes and SVM.

Dataset and features:

Our data was gathered from Talkwalker, an archive of tweets which provides information including the tweet's content, the number of retweets, and the author's demographic information ("Talkwalker..."). We filtered the set of English

tweets based on the hashtag #metoo (as well as common variants such as #meToo and #MeToo). We then collected the resulting tweets through categories of engagement, potential reach, and recentness. Each category provided 250 tweets, resulting in 750 tweets per sample. We repeated this processes five times between 11/12/2017 and 12/10/2017, collecting a total of 3,750 tweets. We randomly distributed these tweets into a 70-30 split for our training and test data. The final result was a training set of 2,625 tweets and a test set of 1,125 tweets.

In order to run the data into our Naïve Bayes and SVM algorithms, we had to first process the data as a matrix of word occurrences. We created a matrix corresponding to our tweet example, with dimensions of the number of tweets we were processing and the number of tokens in our vocabulary. Each entry in the matrix corresponded to the number of a given tokens (corresponding to the column) found in the content of a given tweet (corresponding to the row). The resulting engagement of each tweet was stored in an array. Engagement was calculated as whether or not a tweet passed a certain threshold of retweets. This threshold was calculated as the mean of 250 random number of retweets from our training set.

Our feature set was the number of occurrences of tokens in our lexicon. For our initial test we used the lexicon provided in the second homework. For later iterations, we developed our own lexicon from the word content of 750 #metoo tweets from our training set. After removing capitalization, we then applied Porter stemming to tokenize these words (e.g. store words like 'abusing', 'abusive', and 'abuser' under the single token 'abus') (Goharian, 2013) and removed commonly occurring stop words (e.g. and, or, to) from the lexicon. Our final feature set contained about 5,000 tokens, and included directed accounts (@Alyssa_Milano) and hashtags (#himthough).

Methods:

For our first algorithm, we used Naïve Bayes, with specifically the multinomial event model used for text classification. Naïve Bayes takes in discrete parameters x_i's (in our case, the number of occurrences of words in our lexicon) to predict a given a Bernoulli value y (in our case, whether a tweet passes a certain threshold of retweets). Naïve Bayes makes the assumption that all of our parameters are conditionally independent of each other given y. While this is not always necessarily the case, it often works well in practice and greatly simplifies the number of parameters we have, allowing us to calculate $p(x_1, ..., x_n|y)$ as $\prod_{i=1}^{n} p(x_i|y)$. For multinomial Naïve Bayes, we assume that parameters have a multinomial distribution. The overall probability of a message is still $p(y)\prod_{i=1}^{n} p(x_i|y)$, but now $x_i|y$ is assumed to be a multinomial distribution, rather than a Bernoulli distribution. This allows us to calculate the joint likelihood of the data as

$$L(\Phi_y, \Phi_{k|y=0}, \Phi_{k|y=1}) = \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}) = \prod_{i=1}^{m}(\prod_{j=1}^{n_i} p(x_j{}^{(i)}|y; \Phi_{k|y=0}, \Phi_{k|y=1}))p(y^{(i)}; \Phi_y)$$

Maximizing this equation with respect to our joint likelihood parameters, and applying Laplace smoothing to prevent divide-by-0 errors for unseen input, we get

$$\Phi_{k|y=1} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j{}^{(i)}=k \wedge y^{(i)}=1\}+1}{\sum_{i=1}^{m} 1\{y^{(i)}=1\}n_i+|V|},\ \Phi_{k|y=0} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} 1\{x_j{}^{(i)}=k \wedge y^{(i)}=0\}+1}{\sum_{i=1}^{m} 1\{y^{(i)}=0\}n_i+|V|},\ \Phi_y = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=1\}}{m}$$

From here we can use the parameters and Bayes Theorem to calculate $p(y=1|x)$ and $p(y=0|x)$, applying the y giving the higher probability as the predicted result.

For our second algorithm, we applied a support vector machine. Our implementation again used the starter code from the second homework, including the SVM implementation files provided, and uses a gaussian kernel (RBF).

Broadly speaking, the goal of SVM is to find the hyperplane which best maximizes the margins, or the distance of the points closest to the hyperplane. The lines that pass through these points are called support vectors. We can write the equation for the margin $\gamma^{(i)} = (w^T x + b)$, and solve for $max_{\gamma, w, b}(\gamma\ s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq \gamma,\ i = 1, .., m,\ ||w|| = 1\ )$. We can rewrite this equation as $max_{\gamma, w, b}(\frac{1}{2}||w||^2\ \ s.t.\ \ y^{(i)}(w^T x^{(i)} + b) \geq \gamma,\ i = 1, .., m,)$ which is an optimization problem that can be effectively solved. We can combine this concept with kernels, which exploits the ability to write equations in terms of inner products between input feature vectors to calculate high dimensional features in very little time. This allows SVMs to function in high-dimensional spaces as well. Finally, in order to account for when data is not linearly separable, we allow for the margins to be less than one, with a corresponding cost e_i. Our final equation is

$$max_{\gamma, w, b}(\tfrac{1}{2}||w||^2 + C\sum_{i=1}^{m} e_i\ \ s.t.\ \ y^{(i)}(w^T x^{(i)} + b) \geq \gamma,\ i = 1, .., m, e_i \geq 0)$$

Results:

We conducted three main experiments: (1) feature set selection, (2) definition of engagement evaluation, and (3) model performance. For the first experiment, we compared the accuracy between two different feature sets. Our baseline feature set was the list of words used in problem set 2 while the second feature set were stems derived from the content of 750 randomly selected tweets from the training set. On the second experiment, we compared accuracy between defining engagement as the number of retweets and as the sum of retweets and likes normalized by number of followers. The third experiment consisted evaluating which model, SVM or Naïve Bayes, led to the highest accuracy of the training set predictions. We performed a simple cross-validation where the data was randomly separated between a training set and a cross-validation set with a ratio of 70:30, as we had greater than 100 examples with which to train the algorithms.

Accuracy was the primary metric used for evaluating the success of our classification algorithms. We used accuracy to select which feature set and engagement definition would best predict whether a tweet was engaging. Accuracy was the preferred definition of performance because prediction errors of any particular class did not outweigh the other. Accuracy was more important than precision, sensitivity, or specificity because we wanted to maximize the the number of correct classifications since error in classification does not have broader negative consequences.

Engagement Selection Experiment

To determine which engagement definition led to the highest accuracy, we used the Naïve Bayes model and the old dictionary—based on the features from problem set 2. We compared two definitions of engagement: (1) based on number of retweets and (2) based on the sum of likes and retweets normalized by the user's followers. This experiment concluded that a classification based on number of retweets is more accurate than one based on number of followers and likes additionally. Comparing confusion matrices table 1 and 2, corresponding to definition (1) and (2), respectively, leads to the selection of the definition of engagement based on number of retweets solely.

Total = 1125	Labeled Engaging	Labeled Non-Engaging
Predicted Engaging	1003	16
Predicted Non-Engaging	92	14

Table 1 – Confusion matrix for 1125 examples from the cross-validation set using the lexicon from problem set 2 as the features and Naïve Bayes as the model where engagement was defined by the number of retweets. The results are that accuracy is 90.4%, precision is 98.4%, positive recall is 91.6%, and negative recall is 46.7%.

Total = 1125	Labeled Engaging	Labeled Non-Engaging
Predicted Engaging	968	18
Predicted Non-Engaging	114	25

Table 2 – Confusion matrix for 1125 examples from the cross-validation set using the lexicon from problem set 2 as the features and Naïve Bayes as the model where engagement was defined by the number of retweets plus number of likes divided by number of followers. The results are that accuracy is 88.3%, precision is 98.2%, positive recall is 89.5%, and negative recall is 58.1%.

Although the difference in accuracy is not large, we chose to continue with the definition of engagement based solely on retweets because it is slightly more accurate and is better representation of how far the conversation has spread, independently of number of followers. Furthermore, the definition of engagement based on retweets also had higher precision and positive recall. However, the second definition of engagement was better at classifying non-engaging tweets correctly. The number of followers must help better classify the non-engaging tweets because if the potential reach of a tweet is small, then it must not be as engaging as tweets from accounts with large followings.

Feature Selection Experiment

We used the Naïve Bayes model and the definition of engagement based on solely retweets to decide which feature set led to the most accurate results.

Total = 1125	Labeled Engaging	Labeled Non-Engaging

Predicted Engaging	996	23
Predicted Non-Engaging	95	11

Table 3 – Confusion matrix for 1125 examples from the cross-validation set using content from tweets and token normalization as the features and Naïve Bayes as the model where the engagement was defined by number of retweets. The results are that accuracy is 89.5%, precision is 97.7%, positive recall is 91.3 %, and negative recall is 10.4%.

There is less than 1% difference in accuracy between using the feature set based on problem set 2 (table 1) and the set based on actual #metoo tweet content (table 3). Because of this very small difference in accuracy, we decided to use the feature set developed by token normalization of 750 random tweets from the training set. This choice was made because the features would then include directed accounts and other hashtags used in the movement. By using these additional features, we hope to determine whether the hashtags and @ mentions would be indicative tokens for a successful tweet.

Model Selection Experiment

The following experiments compare accuracy between Naïve Bayes and SVM using the definition of engagement based on retweets and the lexicon made using token normalization from 750 random tweets.

The Naïve Bayes implementation used does not require specifications of hyperparameters. On the other hand, the Gaussian kernel used in the SVM implementation uses a free parameter, specified in the code as tau, of 8. Using a larger free parameter did not increase accuracy. On the other hand, when the free parameter was decreased, the accuracy decreased (i.e. when tau = 0.1, accuracy went down to 75%). Therefore, we specified tau of 8. Similarly, other parameters were not changed because it did not improve accuracy.

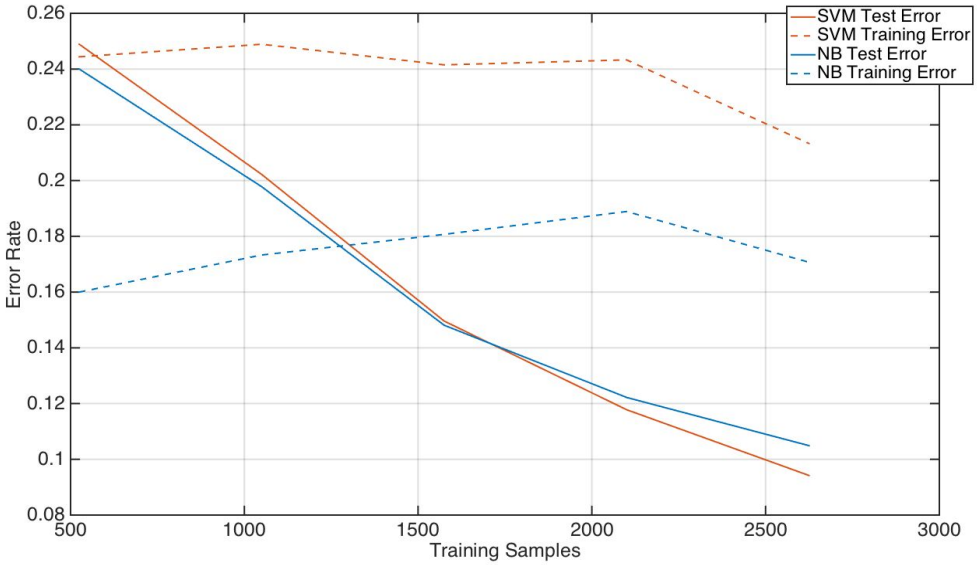

Figure 1 - The error rates of the training set and test set for both SVM and Naïve Bayes. Both the training and test errors are calculated using 30% of the data. For example, if the training set is composed of 1050 examples, the training set error is calculated on only 30% of the training data. Additionally, if the training set is composed of 1050 examples, the test error is calculated on 450 tweets from randomly permuted test data, corresponding to a ratio of 70:30.

Figure 1 shows that Naïve Bayes performs better than SVM when the error is calculated on training set using a 70:30 ratio. However, when comparing test error, both models perform similarly without significant difference in accuracy (1 - error rate). Table 4 shows accuracy for both models using the full training set to calculate well-classified tweets. When looking at the entire training set error, Naïve Bayes continues to perform slightly better than SVM. However, Naïve Bayes is less than 1% less accurate than SVM when using the full test cross-validation set. Therefore, we use Naïve Bayes as our model because it has less error in predicting the training set since the change in accuracy.

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

Table 4 – Accuracy percentages for SVM and Naïve Bayes using the largest size of training samples (2,625 tweets). The test error corresponds to 1,125 tweets that were left-out to test this final accuracy.

Moreover, Naïve Bayes assumes the features are independent while SVM does not. Even though some features should be dependent, i.e. “harvey” and “weinstein”, Naïve Bayes performs with similar accuracy as SVM. The similarity in accuracy suggests that the dependencies within words are distributed evenly within engagement classification.

Using Naïve Bayes for Engagement Classification Results

As seen in table 4, the accuracy of the tweet classification is about 90% for the Naïve Bayes implementation using the stemmed token from original #metoo content and engagement defined by number of retweets.

Using Naïve Bayes, we found that the six most indicative tokens for a successful tweet included sen, risen, profess, restrict, @youtube, and nice, in decreasing order of indication. Figure 2 shows an example of a correctly classified tweet, which uses two of the most indicative tokens: “profess” and “nice”.

Figure 2 - Sample of correctly classified tweet. It was retweeted over 4,000 times and was a reply to another tweet.

Even though the conversation changed dramatically during the month the data was collected, words that elicited conversations about the government (i.e. words that have “sen” as stems, such as senator and senate) and the workplace (i.e. words that have “profess” as stems, such as professional and professor) were the most engaging in the movement. Moreover, discussions praising and supporting the people who broke the silence also brought significant engagement, as evidenced by the indicative token “risen,” which implies a positive sentiment towards the victims. Moreover, when a video from youtube was shared on twitter, it also led to a high probability that it would be an engaging tweet (“via @youtube” appears inside tweet when shared from youtube.com). Tokens “nice” and “restrict” were engaging within the #metoo topic because they could be used for both supporting or opposing a victim or an attacker.

Conclusion:

Our project sought to predict which tweets within the #metoo movement would go on to be engaged with by other users. We collected over three thousand tweets, processing the content of each by the number of word occurrences. Though we experimented with a lexicon provided and our own lexicon customized to the dataset, we ultimately found this made little difference in practice. We analyzed the results using Naïve Bayes and SVM, training the algorithm to predict whether a given tweet would be retweeted more than 450 times. Ultimately both algorithms performed very similarly, with SVM outperforming Naïve Bayes just slightly at 90.58% accuracy on test data. We also found that discussions involving the government, the workplace, and the women who came forward brought the most engagement in the movement.

Given more time, we would want to expand our project in three ways: exploring shifts within the movement, trying additional algorithmic approaches to the data, and applying our algorithms to other social movements. Even within the one-month span of the data we gathered, the conversation within the #metoo tag shifted several times. If we were to continue to gather data over the span of the movement, we could better understand the broader patterns of it while minimizing the variance of individual conversations. Similarly, we would likely extend our project to look at features beyond the content of the tweet, such as author demographic information, perhaps using techniques such as ICA to better determine which features are most relevant and classify engagement based on other #metoo subtopics, such as #himthough. Finally, we would want to better understand how the success of these #metoo tweets could be more generally applied to other social media movements. Our analysis could help bolster the conversations and increase the engagement and impact of future social media conversations.

Contributions:

Ana Tarano- Gathered Data, Implemented Naive Bayes and SVM, Generated Lexicon, Error Analysis

Dana Murphy- Pre-processed Data, Found Research Papers, Coded Engagement, Designed Poster

Sources:

Go, Bhayani, et al. “Twitter Sentiment Classification using Distant Supervision.” *Stanford Engineering*, 2009, www-cs.stanford.edu/people/alecmgo/papers/TwitterDistantSupervision09.pdf

Goharian, Nazli. “Pre-Processing.” *Georgetown College*, Georgetown University, 2013, people.cs.georgetown.edu/~nazli/classes/ir-Slides/Preprocessing-13.pdf

Irani, Webb, et al. “Study of Trend-Stuffing on Twitter through Text Classification.” *Semantic Scholar*, 2010, pdfs.semanticscholar.org/1c3b/e9108479950f8b20400857ab66e2175a7a4f.pdf

Joachims, Thorsten. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features.” *Proceedings of the European Conference on Machine Learning (ECML)*, 1998, www.cs.cornell.edu/people/tj/publications/joachims_98a.pdf

Petrovic, Osborne, et al. “RT to Win! Predicting Message Propagation in Twitter.” *Association for the Advancement of Artificial Intelligence*, 2011, www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2754/3209

Santiago, Cassandra, and Doug Criss. “An activist, a little girl and the heartbreaking origin of 'Me too'.” *CNN*, Cable News Network, 17 Oct. 2017, www.cnn.com/2017/10/17/us/me-too-tarana-burke-origin-trnd/index.html.

Suh, Bongwon, et al. “Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network.” *2010 IEEE Second International Conference on Social Computing*, 2010, doi:10.1109/socialcom.2010.33.

“Talkwalker User Manual.” *Talkwalker*, Talkwalker, 2017, www.talkwalker.com/user-manual/talkwalker

Tareen, Sophia. “Latest Front in Weinstein Scandal: Statehouses Say 'Me Too'.” *usnews*, U.S. News, 24 Oct. 2017, www.usnews.com/news/best-states/illinois/articles/2017-10-24/open-letter-alleges-sexual-harassment-in-illinois-politics.

Zacharek, Dockterman, et al. “The Silence Breakers.” *Time*, TIME magazine, 2017, time.com/time-person-of-the-year-2017-silence-breakers/