Politi-troll

Identifying political trolls on Twitter

Matt Cole

October 25, 2016

Introduction

Internet trolls are considered a menace in nearly all online communities as they strive to generate emotional responses and cause disagreements between users (Stein, 2016). Motivations behind these users are unclear, but some social scientists suspect that a strange phenomenon known as the "disinhibition effect" may be to blame (Suler, 2005). Masked behind the apparent anonymity provided by many forms, sites, and the internet in general, social reservations that facilitate normal, face-to-face conversations can disappear, resulting in sometimes wild and rude behavior (Suler, 2005). This 'troll behavior' can stretch from simply posting the same status repeatedly to annoy and clog streams of information to violent threats and many 'things' in between. The effect of these trolls on Twitter in particular, has been linked to many popular users leaving the platform entirely (Kim, 2016; Stein, 2016). The damage trolls may cause can move beyond individuals however, with Twitter's troll problem considered to be partially responsible for the social networking site's recent loss of 35 billion in market capitalization as potential corporate acquirers have backed out from making buyout offers while many popular users have left the platform because of trolling (Hern, 2016; Kim, 2016). While some trolls have made threats or otherwise broken laws, many are simply an annoyance to other users as well as the site in general, and when unchecked can result in a loss of userbase. Because trolls are a nuisance, it would be beneficial to censor, block, or ban such users before they could damage user experience. In this project, we focused on identifying political trolls on Twitter, an area considered to be particularly saturated with such users.

Data

Twitter data collection began by identifying and collecting usernames, also known as screen names, of accounts that have taken part in the political dialogue on Twitter. These usernames were collected from Twitter both through automated means via the Twitter API and manually through the web app (https://www.twitter.com/) and was limited to English tweets (Gentry, 2012). The initial data collection consisted of searching Twitter, via the twitteR API wrapper for users with tweets containing one of the two major political party presidential candidate's Twitter username within the tweet text (@HillaryClinton or @RealDonaldTrump). This collection occurred on October 1st, 2016 at 4:14:19 PM EST. Twitter API mechanics resulted in a mix of most recent tweets (up to the second), as well as 'hot' tweets that were somewhat recent and popular in terms of favorites and retweets. From these tweets, the usernames of the accounts publishing or retweeting the tweets were extracted and recorded. In total, of the 142 captured tweets, 142 usernames were obtained through this method. From October 7th through October 14th, because of a lack of trolls in the automatically sampled usernames, manual harvesting was conducted, consisting of browsing political figures tweets and responses, from the twitter web app (accessable at https://www.twitter.com/) while logged into our personal twitter account (@mattcol3), in an attempt to collect users with a higher proportion of trolls for modeling purposes (Appendix: Manual Harvest). During this collection period, 132 additional usernames were added for a total of 274.

Once the list of 274 users who have participated in the twitter political discourse was created, each user's 10 most recent tweets, as well as relevant tweet data (favorites, location, retweets, etc.) and data associated with the account (friends, description, etc.), were compiled. Using only the text from each user's 10 most recent tweets, including retweets and replies, users were classified as a troll or non-troll using the following scheme:

Trolls were defined as users with at least one of the previous ten tweets appearing to be both:

- inflammatory: intended to arouse angry or violent feelings
- extraneous: irrelevant or unrelated to the subject being dealt with

Ten tweets were chosen in an attempt to capture enough of each user's tweeting tendencies and classifying users as trolls with at least one inflammatory and extraneous tweet allows us to capture more subtle trolls. In total, all 274 users were classified as troll/non-troll after examining roughly 2,700 individual tweets.

Methods

Utilizing each user's previous 10 tweets, we were able to construct average tweet sentiment using the AFINN lexicon. This lexicon comprises of English words each with an integer valued sentiment ranging from 5 (very positive) to -5 (very negative) scored manually determined by Finn Årup Nielsen from 2009-2011. Stop words, or words that likely will not contributed to the sentiment understanding of the sentence such as: as, the, is, at, which, or on, were removed prior to text analysis.

Sentiment scores for each tweet were then calculated by summing together the values of each word. Then, a user sentiment score was generated by averaging the tweet AFINN sentiment scores over all recorded tweets for each user. Additionally, the NRC lexicon, which assigns scores to each of two sentiments (positive and negative) as well as eight emotions was used used to generate an 'angry' score comprising of anger, fear, and disgust subscores as well as a 'happy' score comprising of joy and trust emotion subscores. As emojis have become increasingly popular in today's web-based communication, an emoji dictionary was scrapped from http://www.unicode.org and emoji usage as well as sentiments were determined and averaged across tweets. When an emoji was identified, the keywords corresponding to the emoji were run through the AFINN lexicon to determine positivity or negativity associated with the emoji within the tweet. For instance (eye | face | grin | smile). Tweet source (how tweets were sent) was collected and divided into three categories, browser-based tweets, mobile tweets, and other tweet systems which included automated tweeting and unknown tweet sources.

After exploring the data (Appendix: EDA) we focused on three models to predict troll status: a simple logistic regression using only one covariate, a 'full' logistic regression model with many covariates as well as a classification tree. In order to assess model fit, 20% of the data was randomly partitioned to a designated 'testing' data set comprising of 55 observations, of which 19 are trolls. The remaining 219 users (including 19 trolls) were used to construct all models which were tested by predicting the troll status of the test set.

Classification and regression trees (CART) are a group of supervised machine learning methods which construct decision tree models utilizing observed data. Unlike other popular methods for classification or regression, CART methods do not produce global models, where a single predictive formula is employed to make decisions or predictions over all observations (James et al., 2013). Instead, CART methods break down the feature space containing the observations into small subspaces (called recursive partitioning) until the subspaces can be represented by relatively simple models. Benefits of CART methods include its high level of interpretability, automatic variable selection, and ease of visualization.

A classification tree was grown using the rpart package in R (Therneau, et al., 2010). Because of the unbalanced data set, we opted to weight the observations as such: non-troll observations were given a weight of 1 while troll observations were weighted as the ratio of non-trolls:trolls in the training set, 2. The classification tree's nodes were pruned back by reducing the number of nodes to minimize the 10-fold cross-validated error rate (Figure 2).

Logistic regression models the probability of a user being a troll for each observation i (p_i) by treating the logit function, $log(\frac{p_i}{1-p_i})$ as a linear function of the covariates. In this study, a major limitation may be logistic regression's inability to capture non-linear trends associated with the covariates. However, this limitation can be partially overcome by the addition of splines or interaction terms. First, we examined a simple logistic regression only consisting of AFINN sentiment, to better understand the analysis. Because we do not assume that our observations are independent of one another, we utilized a quasibinomial family to allow the dispersion parameter to be greater than one. Another logistic model was fit, utilizing additional

variables such as _____ and _____ until _____. We also assumed a quasibinomail family for this model as well.

$logit(p_i) = \beta_0 + \beta_1 AFFIN$

Results

Both the simple logistic and full logistic models had error rates of 0.418 and 0.273 respectively with the simple logistic model having a slightly higher sensitivity but lower specificity compared with the full model.

In our simple logistic regression we found a prediction error of 0.58 in the testing group. In addition, we see that without controlling for any other covariates, for each 1 unit decrease in average AFINN sentiment score, expected odds of being a troll increased by a factor of 0.8561949.

Table 1: Logistic regression coefficients

	Estimate	Low	High
Intercept	0.9871	0.5757	1.692300e+00
Average Tweet Sentiment (AFINN)	0.4747	0.3027	7.443000e-01
Follwers (100)	1.0083	1.0022	1.014400e+00
Listed Count (100)	0.6589	0.4717	9.205000e-01
Mobile User (binary)	0.9386	0.8860	9.943000e-01
Default Image (binary)	3.0366	1.2807	7.199700e+00
Number of Friends (100)	0.9999	0.9882	1.011800e+00
Number of Badwords used	1.3233	0.9560	1.831600e+00
Web User (binary)	0.2087	0.0000	1.649304e+70

Table 1 exponentiated Logistic regression coefficients.

Here we see that for each one unit increase in AFINN sentiment score, the odds of a user decreases by a factor of 0.47 (95% CI: 0.3 , 0.74), while the number of followers and friends increases the changes the odds of being a troll by 1.01 (95% CI: 1 , 1.01) and 1 (95% CI: 0.99 , 1.01) respectivly. Users whom access Twitter via mobile devices and those using through desktop applications were both associated with a decrease in likelihood of being a troll, with mobile users and desktop users associated with a change in odds of trollhood by 0.94 (95% CI: 0.89 , 0.99) and 0.21 (95% CI: 0 , 1.6493038×10^{70}). Presence of a default image (sometimes refered to as the beginer's egg) is associated with an increase in odds of being a troll of 3.04 (95% CI: 1.28 , 7.2) while for each unit increase in average 'bad word' per tweet odds of trolls increased by a factor of 1.32 (95% CI: 0.96 , 1.83)

For our classification tree, We found a CV prediction error rate of 0.345, sensitivity of 0.526, specificity of 0.722, and a AUC of 0.736 in the testing group (figure 1, table 2). This showed a considerably higher degree of accuracy compared with the logistic regression models that were produced.

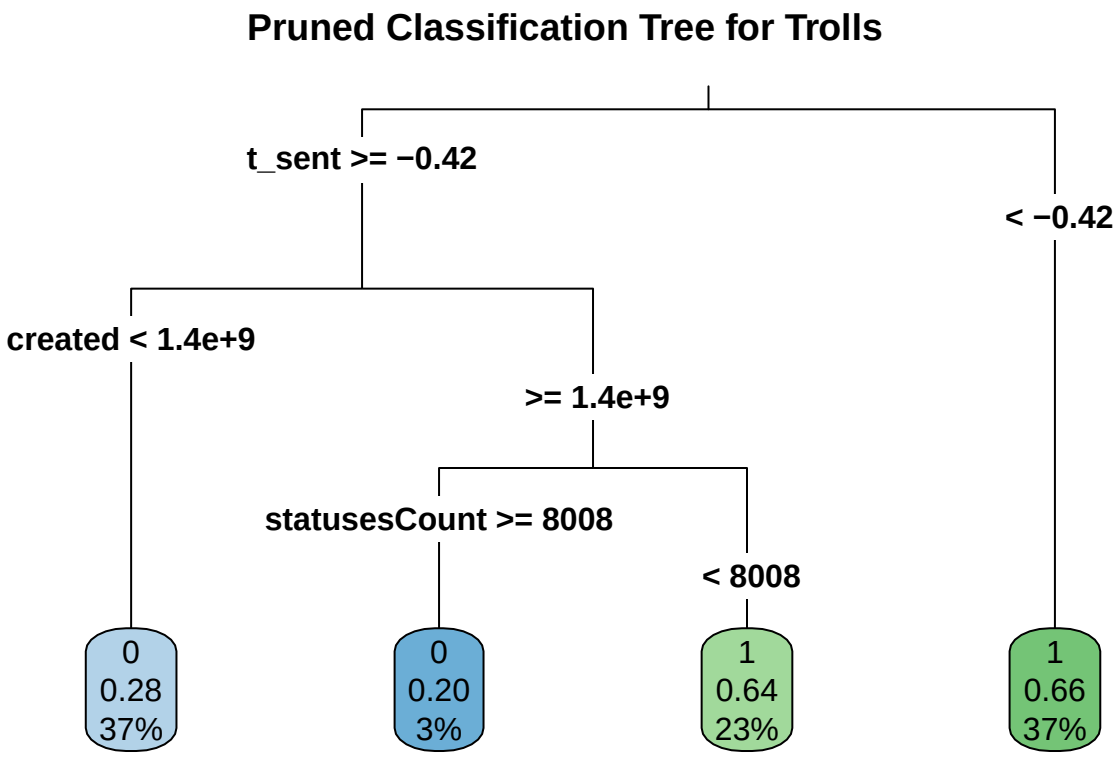

Figure 2 Pruned regression tree visualization. Nodes indicate, prediction (0 or 1 for troll status), mean troll value, and percentage of data contained in each node.

	Error Rate	Sensitivity	Specificity	AUC
Simple Logistic	0.42	0.47	0.64	0.64
Full Logistic	0.27	0.74	0.72	0.74
Regression Tree	0.35	0.53	0.72	0.74

Table 2 raw error rate, Sensitivity, Specificity, and "AUC" comparison of all three models.

Discussion

Results

It was expected that the regression tree would be able to outperform the logistic regression as the relationship between covariates and troll status seem opaque at best and would likely be best modeled with a non-parametric method (table 2). As we saw, the classification tree was in fact able to outperform the logistic regression models with respect to error rate, sensitivity and AUC however, the specificity results were abysmal with a value of 0.55, suggesting that our model was barely better than flipping a coin for classification of non-trolls, and much worse in this regard than either of the logistic regression models (table 2).

logistic regression 0.418 to 0.273, while increasing the sensitivity (from 0.474 to 0.737), specificity (from 0.639 to 0.722), and the AUC (from 0.642 to 0.744).

This study did shed light on a strange phenomena of different classes of trolls. From our anecdotal experience, there seemed to be three types or classes of twitter trolls.

- 'personal accounts' comprising of students, professionals, etc. whom occasionally say mean/emotionally charged things on twitter.
- bots: usually newly created, low tweet/follower/friend count which seem to spew politically charged propaganda.
- troll/fan pages: these accounts seem to be semi-autonomous as they respond to political figures (@therealdonaldtrmp, @hillaryclinton) almost instantaneously while also being able to respond to certain accounts in a thought out manner. These accounts have many tweets, were created long ago and have many followers. In fact, Fortune, POLITICO and the New York Magazine all have run stories of these 'bot armies'.

Surprisingly, many of the covaraites did not make the final model in either the logistic or classification tree case. These included the anger score, bot users, emoji score, etc. This could be because of a lack of relationship between these variables and troll status, or it could be due to the lack of data to identify such nuanced features.

Limitations

The main limitation of this study was a lack of available data, particularly trolls. Although there were 274 observations there were only 92 trolls in our dataset, more would likely allow us to identify additional relationships and interactions between variables and troll status. In addition, there is the possibility for human error as a single person classified all tweets. This error could come from keystroke error or implicit political bias. Political bias could be difficult to identify and overcome where beliefs of what actually occurred could alter whether a statement is either inflammatory or extraneous. A good example of this are statements such as "Bill Clinton is a rapist", which were classified as both inflammatory and extraneous, but to an individual whom believes that the statement is true, would likely not view it as extraneous as it is simply stating the truth.

Reproducibility

Because of the subjective nature of troll classifcation this proje g will nto be completely reproducible. There are several reproducibility issues ti i n this project inlucding username and tweet collection as well as troll classificaiton. Username colleciton was mostly conducted in an automated fashion, however, although API searches can be rerun at any time (using time constraints), Twitter is notorious for tweet deletion, editing, and even usernames changing. With this in mind, running the same commands will likely not replicate the response. In addition, for the roughly 45% of the usernames were collected by browsing major presidential candidates for replies, and recording their usernames will be subject to the same sorts of issues as the API, etc. **_ADDMOREHERE**

Generalization

It is plausible that these models could generalize to other spaces within twitter or even outside of the social media site where trolling is a serious issue. However, the model relies heavily on non-text covariates such as friends, dates, and followers and applying this model to, an anonymous online form would likely be less effective, as this user data would be missing. the models here may not generalize to sites where the long posts are common as tweets are capped at 140 characters, however potentially averaging sentiments or bad words (ie. average sentiment per 100 characters or 10 words) may possibly be effective.

Reproducibility

- with this project the tradeoff between reproducibility and both generalizability as well as relevance became apparent very quickly. “Did each user have 10 total tweets? How did you define inflammatory or extraneous? A bag of words? Your judgment?” -John

Future Work

In the future, it would be beneficial to further optimize tuning parameters associated with these models and inspect how other algorithms perform, particularly convolutional neural networks as well as support vector machines which may be able to identify additional hard to see relationships in the data resulting in better prediction accuracy. In addition, future work should include identifying word n-grams which may allow us to gain additional insight beyond lexicons. n-grams are popular in computational linguistics as two words in sequence contain additional information about the meaning of the sentence compared to lexicon analysis of the same two words. Future would could focus on classify individual tweets as troll / non-troll tweets or even non-binary degrees of ‘troll’, for instance, classifying three types of trolls mentioned earlier in the discussion. It would also be interesting to run a latent class analysis on the variables In our study we examined two types of models, Logistic regression as well as regression trees which were chosen for interpretability.

Thank you

Thanks to Finn Årup Nielsen manually labeled the words in the AFINN lexicon from 2009-2011, which was used in this analysis.

Thank you to http://unicode.org/emoji/charts/full-emoji-list.html for letting me borrow the emoji database without asking.

References:

Hern, Alex. “Did Trolls Cost Twitter $3.5bn and Its Sale?” The Guardian. Guardian News and Media, 18 Oct. 2016. Web. 21 Oct. 2016.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 6). New York: springer.

Kim, E. (2016, October 17). Twitter trolls were part of the reason why Salesforce walked away from a deal. Business Insider

Martin, J. H., & Jurafsky, D. (2000). Speech and language processing. International Edition, 710.

Stein, J. (2016, August 18). How Trolls Are Ruining the Internet. TIME

Suler, J. (2005). The online disinhibition effect. International Journal of Applied Psychoanalytic Studies, 2(2), 184-188.

Packages

Bowman, A. W., & Azzalini, A. (2013). R package sm: nonparametric smoothing methods (version 2.2-5).

Gentry, J. (2012). twitteR: R based Twitter client. R package version 0.99, 19.

Robinson, D. & Silge, J.(2016). tidytext: Text Mining using ‘dplyr’, ‘ggplot2’, and Other Tidy Tools R package version 0.1.1

Therneau, T. M., Atkinson, B., & Ripley, B. (2010). rpart: Recursive Partitioning. R package version 3.1–42.

Wickham, H., & Francois, R. (2015). dplyr: A grammar of data manipulation. R package version 0.4, 1, 20.
Wickham, H., Hester, J., & Francois, R. (2016). R package readr: Read Tabular Data. version 1.0.0.

Exploratory Graphs

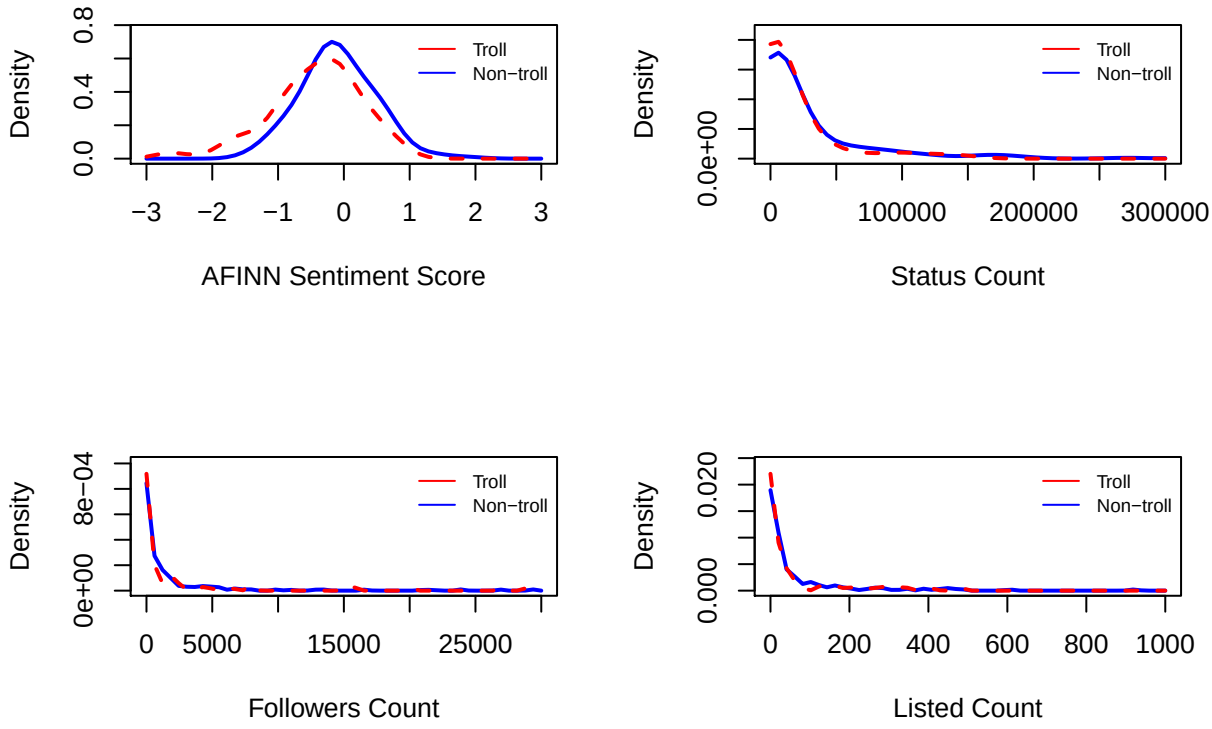

Figure A1 Smoothed density plots of common twitter metrics in both troll and non-troll groups. Trolls are denoted by the red line while non-trolls are denoted by the blue line. Distribution of AFINN sentiment score and year of account creation appear to potentially be different in the two groups while there is no noticeable difference in follower or status count.

EDA

Exploratory data analysis revealed that no single factor showed good separation, between trolls and non trolls (appendix, EDA). Several covariates expected to have a strong relationship with troll status such as number of followers did not appear to show such in the 2 dimensional plots (appendix, EDA). However, others such as AFINN sentiment score did show a possible relationship, although weak at best (appendix, EDA).

After models were fit, they were assessed

plot(logistic_fit1)

class(log2)

Manual Harvest

After viewing the automatically harvested tweets, it had become clear that the dataset would be too imbalanced to conduct a meaningful analysis (about 2% of collected users were defined as trolls). It was apparent that there was a need to collect users with a higher likelihood of being a troll than the general 'political' space of twitter. From anctidotal experience, it has been observed that the replies to major political

figures have a higher prevelance of trolls than 2%, and it was decided that recording usernames of those responding to major candidates