Politi-troll

Identifying political trolls on Twitter

Matt Cole

Due: October 28, 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 3.5 billion in market capitalization as potential corporate acquirers have backed out from making buyout offers (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, unique 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 (if less than 10, as many as available), 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 facilitate the english language but likely will not contributed to the sentiment understanding of the text 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 (web app, tweetdeck), mobile tweets (phone apps, mobile websites), 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 73 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 number of followers and default picture status. These additional covariates were chosen

among common common metrics to identify undesirable users utilized by corporations (Morrison, 2014). We also assumed a quasibinomail family for this model as well.

Results

The simple logistic and full logistic models had raw error (misclassification) rates of 0.418 and 1 respectively with the full logistic model having a lower sensitivity and specificity compared with the full model (0.474 & 0 & 0.639 to 0 respectively) (Table 2).

Table 1: Logistic regression coefficients

	Estimate	Low	High
Intercept	0.987	0.522	1.866
Average Tweet Sentiment (AFINN)	0.475	0.279	0.807
Follwers (100)	1.008	1.001	1.016
Listed Count (100)	0.659	0.444	0.978
Mobile User (binary)	0.939	0.877	1.005
Default Image (binary)	3.037	1.095	8.418
Number of Friends (100)	1.000	0.986	1.014
Number of Badwords used	1.323	0.901	1.943
Web User (binary)	0.209	0.000	50.000

Table 1 Exponentiated logistic regression coefficients with 95% confidence interval from the full model.

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: 0.28 , 0.81), while the number of followers and friends increases the changes the odds of being a troll by 1.01 (95% CI: 1 , 1.02) 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: 0.88 , 1) and 0.21 (95% CI: 0 , 50). 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: 1.1 , 8.42) while for each unit increase in average 'bad word' per tweet odds of trolls increased by a factor of 1.32 (95% CI: 0.9 , 1.94)

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.

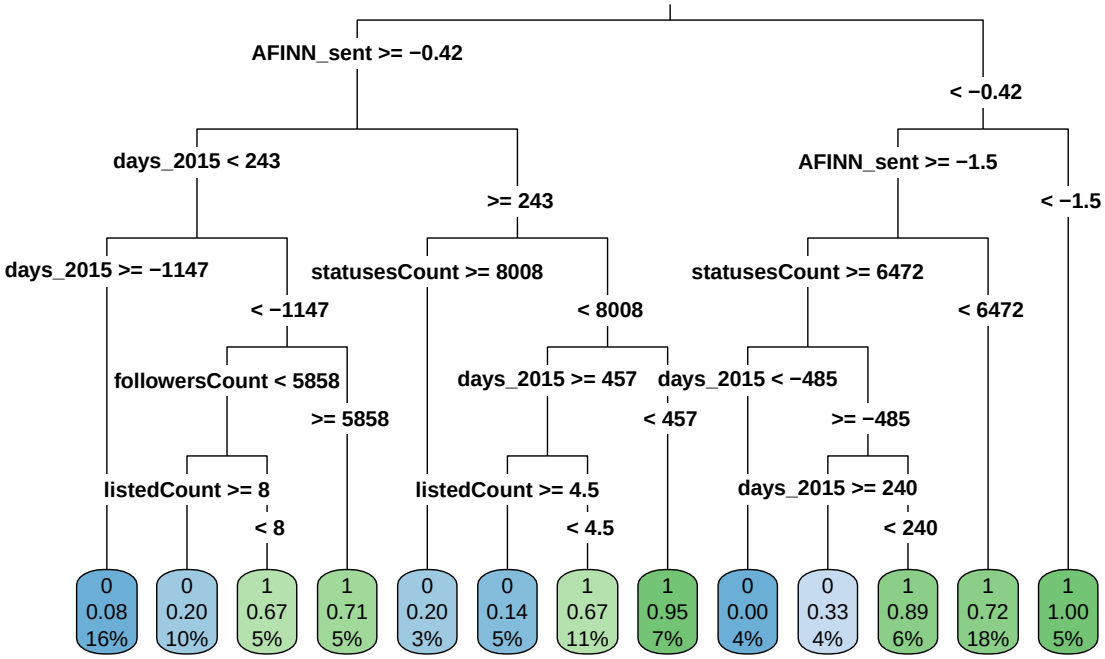

Figure 2 Pruned regression tree visualization. Nodes indicate prediction (0 or 1 for troll status), fraction of trolls in node (training set), and percentage of data contained in each node. Variables present in the pruned classification tree include: **AFINN_SENT** the average sentiment of a users tweets, **days_2015** - the number of days since 2015 that the account was created, **statusesCount** - the number of statuses made by a user at time of colelction, **listedCount** - the number of lists a user of is a part of.

Consider the great account, @mattcol3. @mattcol3 joined twitter in April 2011, has 1060 tweets, 212 followers, is not on any lists, and has an average sentiment score of 1.7 out of his last 10 tweets. Because the sentiment score is greater than or equal to -0.42 we would follow the classification tree branch to the left from the first split. Because the account was made in april 2011, about -1335 days after January 1st 2016, we would move down to the right. The next split concerns follower counts of less than 5858, because @mattcol3 fits that criteria we will move to the left to the last split. @mattcol3 is not on any lists so we slide to the right and reach our prediction: the classification tree estimates that @mattcol3 is a troll! We see that 5% of our training data fell into that terminal node and about 71% were trolls, you're in good company @mattcol3!

	Error Rate	Sensitivity	Specificity	AUC
Simple Logistic	0.42	0.47	0.64	0.64
Full Logistic	1.00	0.00	0.00	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 in the testing group.

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 matched our full logistic model with respect to specificity and AUC and was outperformed with respect to raw error rate in the testing group (Table 2).

We also expected the additional covariates in the full logistic model to increase its viability as a troll predictor (Table 2). This was evident as the error rate dropped 0.418 to 1, while sensitivity increase (from 0.474 to 0), as well as specificity (from 0.639 to 0), and AUC (from 0.642 to 0.744) in the two models (Table 2).

Surprisingly, many of the covaraites did not make the final model in the pruned classification tree model. These included the anger score, emoji score, number of individuals they are following 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.

Although not excellent, the results from our models suggest that there are attributes which may allow us to identify trolls and possibly silence them, shortening their troll reign.

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 (as opposed to double checking). This error could come from keystroke error or implicit political bias. Political bias could be difficult to identify and can occur when beliefs of what actually occurred alter whether a statement is inflammatory or extraneous. For instance, roumors circulating of sexual assaults commited by Bill Clinton may be considered extraneous by those whom view it as a falsehood while those whom believe it to have occured could consider talk of it to be propper.

Reproducibility

Because of the subjective nature of troll classifcation this project will not be completely reproducible. In addition to the classification issues, there are also potential additional difficulties inlucding username and tweet collection. Username colleciton was mostly conducted in an automated fashion and although API searches can be rerun at any time (using time constraints), Twitter users are notorious for tweet deletion, editing, and even username swapping. With this in mind, running the same commands will likely not replicate the response exactly. In addition, for the roughly 45% of the usernames were collected by browsing major presidential candidates for replies, and recording their usernames cannot be replicated exactly as users may delete accounts or replies at any time and was, in addition, not done in an automated fashion.

Generalization

It is plausible that these models could generalize to other spaces within twitter or even outside of this particular 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, for example, an anonymous online form would likely be less effective, as this user data would be missing. The models here would likely not be effective on 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.

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.

Special shout out to my MacBook's 5400 RPM hard drive for failing durring the last leg of this project.

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.

Morrison, K. (2014) Trolldor Helps Twitter Users Identify Trolls. SocialTimes

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.

Appendix / Supplement

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).

Exploratory Graphs

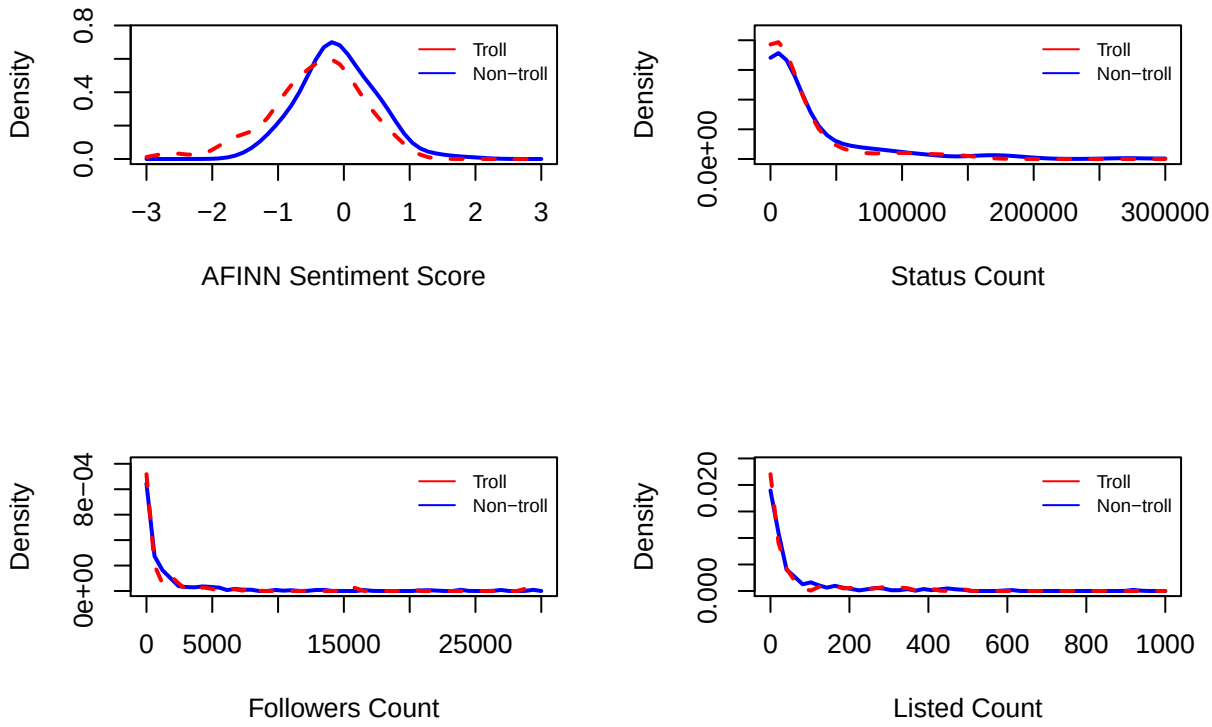

Figure A1 Smoothed density plots of common twitter metrics in both troll and non-troll groups. Trolls are denoted by the red line while non-trolls are denoted by the blue line. Distribution of AFINN sentiment score and year of account creation appear to potentially be different in the two groups while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

After models were fit, diagnostics were run to assess model fit. Interestingly, neither model showed a consistent residual over all predicted values. the simple logistic model showed a nearly monotonically increasing expected value of the residual as the predicted value increased while the 'full' logistic regression, possibly due to some extreeme predicted values, also showed non-zero expected residual values across the most common (-5,5) predicted logits.

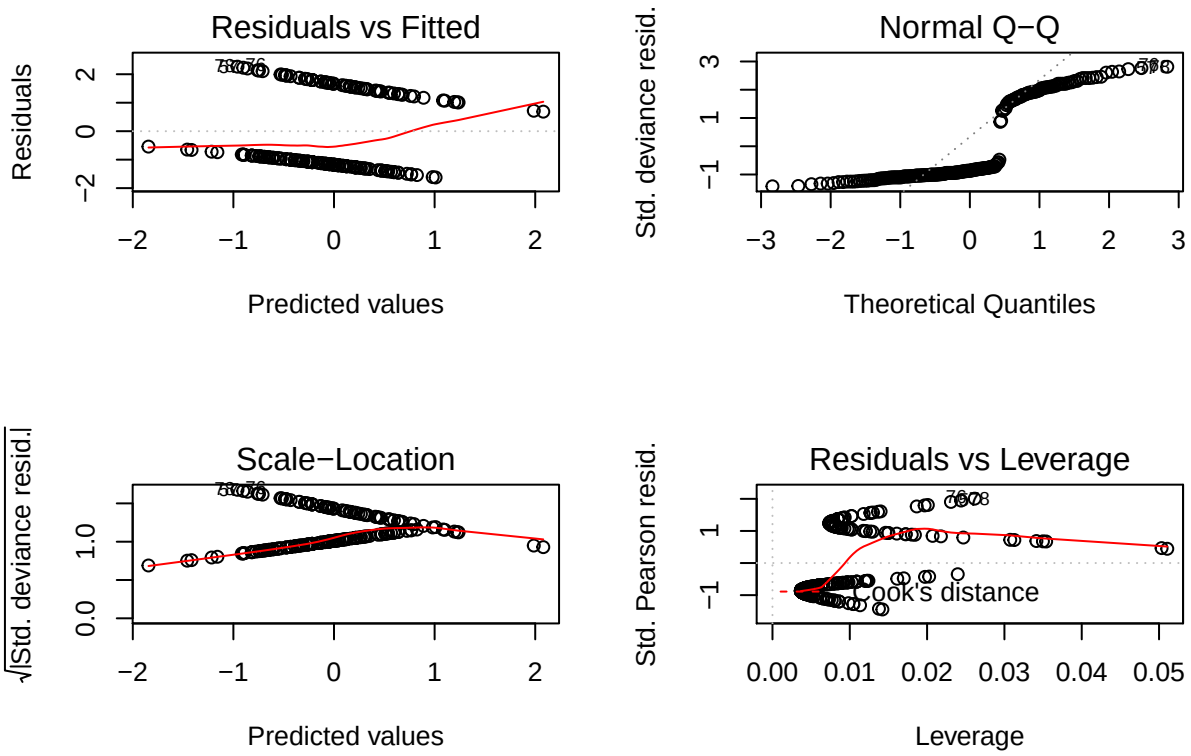

Figure A2 Diagnostic plots for the simple logistic regression.

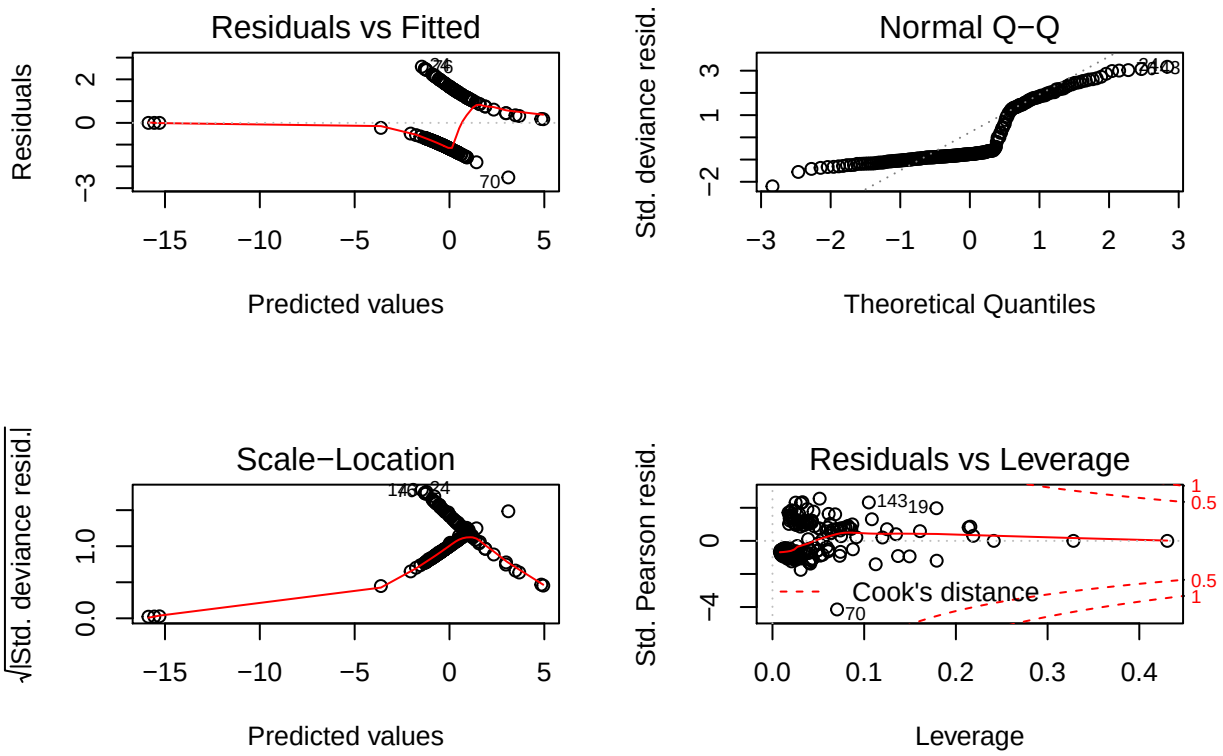

Figure A3 Diagnostic plots for the multiple logistic regression.

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

Other things learned

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'.