Troll Project

Matt Cole

October 7, 2016

Introduction

Internet trolls are considered a menace in nearly all online communities as they create fruitless arguments in an attempt to generate emotional reactions from individuals. Motivations behind these users are unclear, but some social scientists have postulated that a strange phenomenon known as the "disinhibition effect" may be to blame. 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. This 'troll behavior' can stretch from simply posting the same status repeatedly to annoy and clog streams of information to violent threats and many, many 'things' in between. The effect of these trolls on Twitter, in particular, has been linked to many popular users leaving the platform entirely. The damage trolls may cause moves further than individuals, with Twitter's troll problem partially responsible its recent loss of 35 billion in market capitalization according to some analysts (Hern, 2016).

In this project, we focused on identifying political trolls on Twitter, and area considered to be particularly saturated with Trolls.

Data

Usernames were collected from Twitter using both the Twitter API and web app, and data collection was focused on english users. Initial data collection consisted of searching twitter, via twitter API for users with tweets containing one of the two major political party candidates for president's twitter handles in the tweet text (@hillaryclinton or @therealdonaldtrump) on October 1st 2016 at 4:14:19 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. In total, 143 usernames were collected through this method. Later, from October 7th through October 14th, because of an apparent lack of trolls in the API generated users, manual user harvesting was conducted. This manual harvest consisted of browsing political figures tweets, in an attempt to find users responding whom are of interest. Once a list of usernames was collected, user's 10 most recent tweets as well as relevant tweet data (favorites, location, retweets, etc.) were collected as well as data associated with the account as a whole (friends, description, etc.).

Trolls were defined as users with at least one of the 10 collected tweets which appeared to both:

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

10 tweets were chosen in an attempt to capture enough of each user's tweeting tendencies and classifying users as trolls with at least one troll tweet allows us to capture more subtle trolls.

Methods

Using data from the last ten tweets (inclusive of retweets), we were able to assess average tweet sentiment using the AFINN lexicon which comprises of English words with valence measures between -5 and 5 inclusive with an integer between minus five (negative) and plus five (positive). Additionaly, the NRC lexicon, which assigns 'scores' to each of two sentiments and eight emotions were used to generate an 'angry' score comprising of anger, fear, and disgust as well as a 'happy' score comprising of joy and trust emotions. As emojis have become increasingly popular in today's web-based communication, an emoji dictionary was scrapped from

unicode.org and emoji usage as well as sentiments were assessed across tweets. Tweet source was collected and divided into three categories, browser-based tweets, mobile tweets, and other tweet systems which included automated and unknown tweet sources.

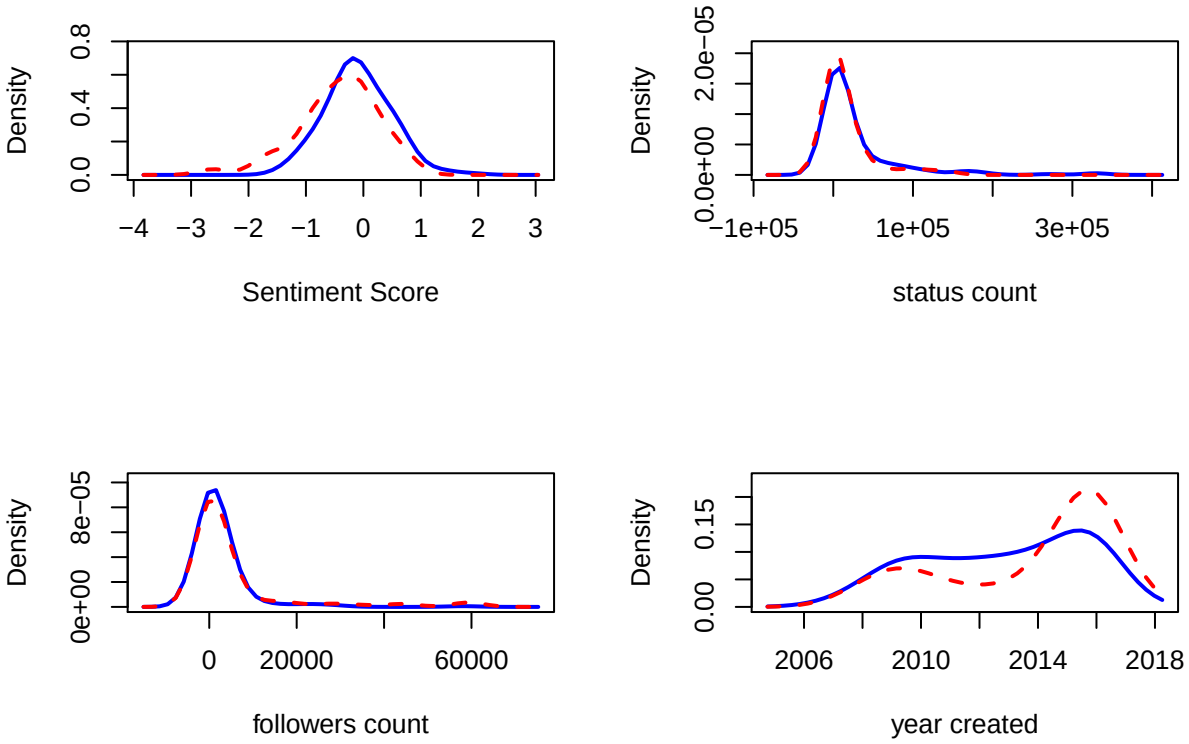
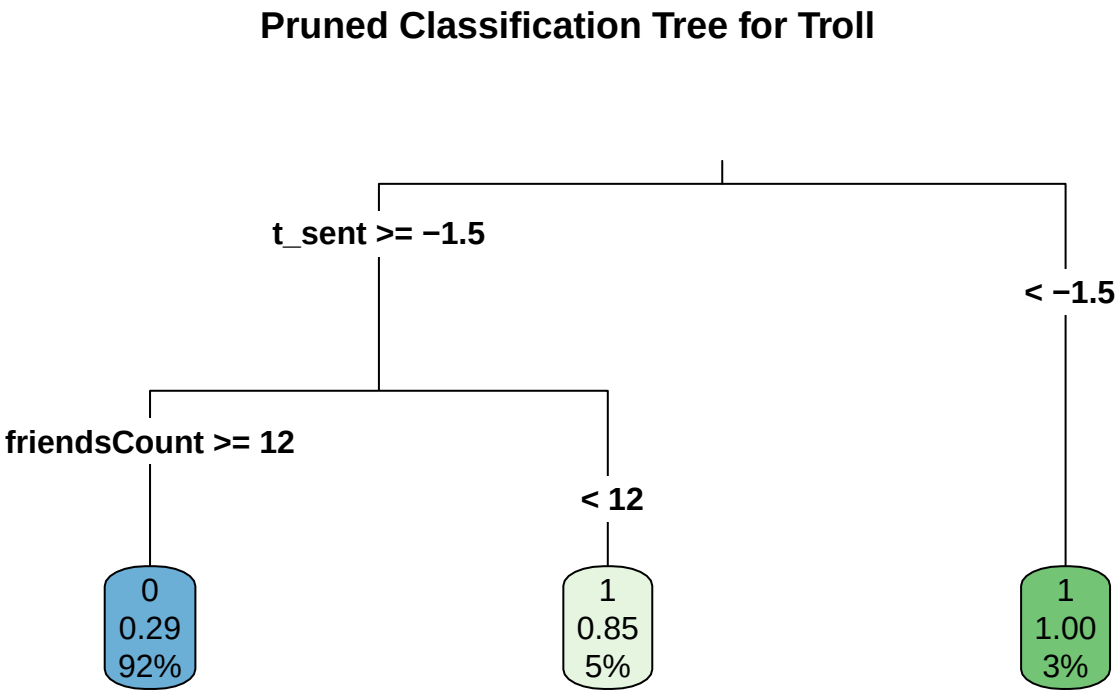


Figure 1 Density plots of common twitter metric in both troll and non-troll groups. Trolls are denoted by the red line while non-trolls are denoted by the blue line.

Once all data was preprocessed, we focused on two models a logistic regression and classification trees. 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, regression trees are not 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 are able to be represented by relatively simple models that do not necessarily utilize all covariates presented in feature space. Benefits of CART methods include its high level of interpretability and simple models, automatic variable selection, and ease of visualization. The classification tree was grown using many features and pruned to minimize the cv-10 error rate. A major issue of classification trees is the potential for overfitting, with the cv-pruning to eliminate this issue. Logistic regression models the probability of a user being a troll for each observation i (p_i) my 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.

Results




```
## [1] 0.1894773
```

Figure 2 Pruned regression tree visualization

Table 1: A table produced by printr.

	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	-0.69954	0.28972	-2.41457	0.01575
Average Tweet Sentiment (AFFIN)	-0.79279	0.26018	-3.04708	0.00231
Follwers	0.00011	0.00003	3.36250	0.00077
Listed Count	-0.00343	0.00227	-1.51046	0.13093
Number of Favorites	-0.00002	0.00001	-2.19782	0.02796
Badwords count	0.41887	0.16408	2.55277	0.01069
Mobile User	-0.06772	0.03205	-2.11313	0.03459
Default Image	0.97239	0.43808	2.21966	0.02644

Table 1 Logistic regression coefficients exp(confint(logistic_fit2))

Here we see that our regression tree, pruned to minimize the 10-fold cv-error measure has a cv-prediction error rate of 0.27. Using a simple logistic regression only consisting of AFFIN sentimant analysis we found a prediction error of 0.31. A 'full' logistic model created from backwards, stepwise elimination of least significant coefficients untill all were significant at $\alpha = 0.1$ had cv error rate of 0.24.

Discussion

It was expected that the regression tree would be able to outperform the logistic regression as the relationship between covariates and being a troll seemed opaque at best and best modeled with the non-parametric method.

However, as we saw, a simple logistic regression had a similar classification error rate as the regression tree and a model including additional important variables was able to outperform the regression tree. The main issue of this study was a lack of available data, particularly trolls. Although there were 273 observations, more would likely allow us to identify additional relationships and interactions between variables and troll status. 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 together in sequence contain additional information about the meaning of the sentence they are contained in compared to lexicon analysis of the same two words. In fact, word n-grams have been utilized to detect text meaning (Martin, et. al, 2000) in various settings.

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

- 'regular' users of twitter (students, professionals, personal accounts) whom occasionally say mean/emotionally charged things on twitter
- bots: usually newly created, low tweet/follower/friend count who seem to spew politically charged propaganda
- troll/fan pages: these accounts seem to be semi-autonomous as they respond to policial figures (@therealdonaldtrmp, @hillaryclinton) almost instantaneously while also seeming to be able to respond to certain accounts in a thought out manner. These accounts unlike the previous two have many tweets, were created long ago and have many followers.

Thank you

Finn Årup Nielsen manually labeled the words in the AFINN from 2009-2011.

Thank you to http://unicode.org/emoji/charts/full-emoji-list.html for letting me borrow the list 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.

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

Packages

Hadley Wickham and Romain Francois (2016). dplyr: A Grammar of Data Manipulation. R package version 0.5.0. https://CRAN.R-project.org/package=dplyr

Terry Therneau, Beth Atkinson and Brian Ripley (2015). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-10. https://CRAN.R-project.org/package=rpart

Hadley Wickham, Jim Hester and Romain Francois (2016). readr: Read Tabular Data. R package version 1.0.0. https://CRAN.R-project.org/package=readr

Bowman, A. W. and Azzalini, A. (2014). R package 'sm': nonparametric smoothing methods (version 2.2-5.4) URL http://www.stats.gla.ac.uk/~adrian/sm, http://azzalini.stat.unipd.it/Book_sm

Stephen Milborrow (2016). rpart.plot: Plot 'rpart' Models: An Enhanced Version of 'plot.rpart'. R package version 2.1.0. https://CRAN.R-project.org/package=rpart.plot

Jeff Gentry (2015). twitteR: R Based Twitter Client. R package version 1.1.9. https://CRAN.R-project.org/package=twitteR