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 comments. 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. 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.

In this project, we focused on political trolls on Twitter, those whose tweets aim to disrupt the flow of information in twitter's political sphere.

Data

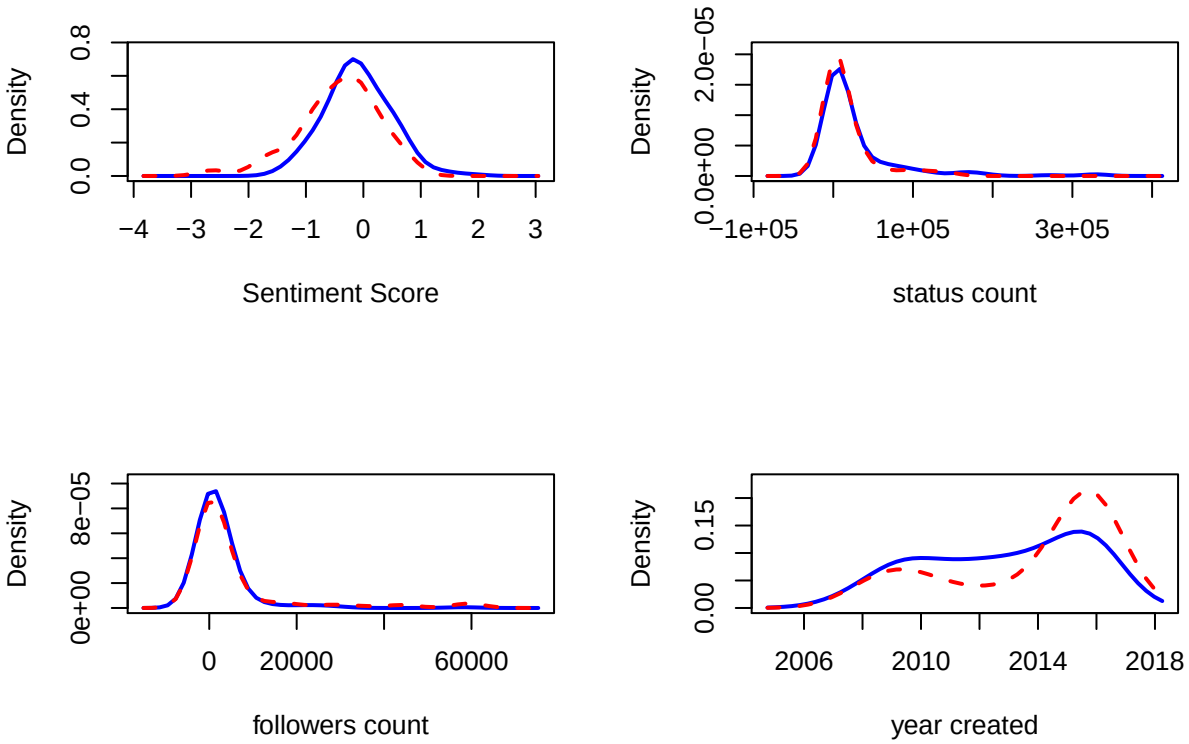
Trolls were defined as users with tweets which appeared to both:

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

Data was collected from twitter using both the Twitter API and web app. Initial data collection consisted of both searching twitter for political figures in tweet text (ie. clinton, trump). Manual user harvesting consisted of browsing political figures' accounts in an attempt to find users of interest. Once a list of usernames was collected, each user's troll status was determined by looking entirely at the previous 10 tweets and users were determined to be a troll by the presence of at least 1 troll tweet.

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). Additional lexicons which include the NRC Emotion Lexicon, which assigns 'scores' to each of two sentiments and eight emotions were used too. 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 was assessed.

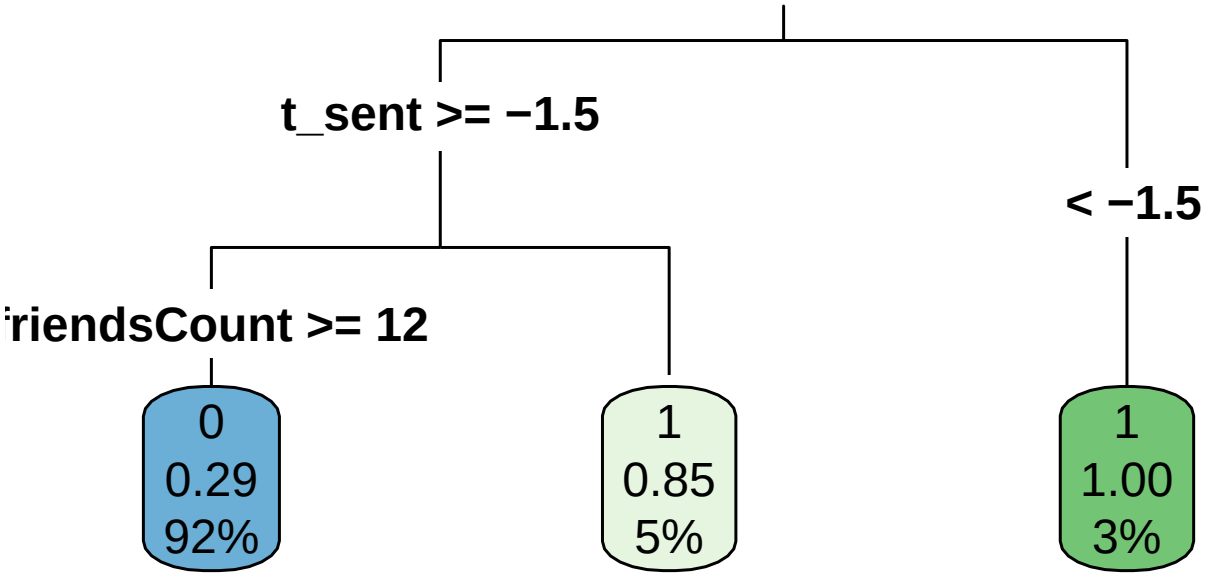


Several statistical and machine learning techniques were used in an attempt to predict whether a user was a troll or not including linear regression, cart, and svm.

Results

```
## Call:
## rpart(formula = troll ~ t_sent + created + followersCount + listedCount +
##     statusesCount + favoritesCount + friendsCount + lang + bdword +
##     user_w + user_a + user_m + angry + imagedefault, data = data,
##     method = "class")
##   n= 274
##
##           CP nsplit rel error    xerror       xstd
## 1 0.09782609      0 1.0000000 1.0000000 0.08497017
## 2 0.03985507      2 0.8043478 0.9782609 0.08450201
##
## Variable importance
##         t_sent   friendsCount  statusesCount followersCount favoritesCount
##             35             33             18             10              5
##
## Node number 1: 274 observations,    complexity param=0.09782609
##   predicted class=0  expected loss=0.3357664  P(node) =1
##     class counts:   182    92
##    probabilities: 0.664 0.336
##   left son=2 (265 obs) right son=3 (9 obs)
##   Primary splits:
##       t_sent         < -1.476667  to the right, improve=8.211431, (0 missing)
##       friendsCount   < 11.5       to the right, improve=8.021176, (0 missing)
##       created        < 1424626000 to the left,  improve=7.273412, (0 missing)
```

```
##       statusesCount  < 11.5       to the right, improve=6.338828, (0 missing)
##       followersCount < 1.5        to the right, improve=5.694031, (0 missing)
##
## Node number 2: 265 observations,    complexity param=0.09782609
##   predicted class=0  expected loss=0.3132075  P(node) =0.9671533
##     class counts:   182    83
##    probabilities: 0.687 0.313
##   left son=4 (252 obs) right son=5 (13 obs)
##   Primary splits:
##       friendsCount   < 11.5       to the right, improve=7.765789, (0 missing)
##       statusesCount  < 11.5       to the right, improve=6.782741, (0 missing)
##       created        < 1405575000 to the left,  improve=5.287246, (0 missing)
##       followersCount < 1.5        to the right, improve=5.206964, (0 missing)
##       favoritesCount < 0.5        to the right, improve=5.206964, (0 missing)
##   Surrogate splits:
##       statusesCount  < 11.5       to the right, agree=0.977, adj=0.538, (0 split)
##       followersCount < 0.5        to the right, agree=0.966, adj=0.308, (0 split)
##       favoritesCount < 0.5        to the right, agree=0.958, adj=0.154, (0 split)
##
## Node number 3: 9 observations
##   predicted class=1  expected loss=0  P(node) =0.03284672
##     class counts:     0     9
##    probabilities: 0.000 1.000
##
## Node number 4: 252 observations
##   predicted class=0  expected loss=0.2857143  P(node) =0.919708
##     class counts:   180    72
##    probabilities: 0.714 0.286
##
## Node number 5: 13 observations
##   predicted class=1  expected loss=0.1538462  P(node) =0.04744526
##     class counts:     2    11
##    probabilities: 0.154 0.846
```




```
## [1] 0.2105571
```

```
## [1] 0.2035869
```

Here we see that the pruned regression tree has a cv-prediction error rate of `r` using a simple logistic regression we found a prediction error of 0.31 and including more relevant variables 0.28. The regression tree was pruned according to 10-fold cross vadiated prediction error.

Discussion

Our best performing model did not classify well (results). However, this study did shed light onto a strange phenomina of different classes of trolls. From our observation, 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 whom 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 instantaniously while also seeming to be able to respond to certian 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. https://www.theguardian.com/technology/2016/oct/18/did-trolls-cost-twitter-35bn