# **Classification of Tweets Related to Gun Control or Gun Rights**

**Introduction**

In light of the Feb 14th, 2018 Parkland Florida school shooting and the energizing force it appears to be in the gun debate, I am interested in which side is winning the “Gun Rights” vs “Gun Control” argument in the court of public opinion. I am more interested in determining which side has momentum and is gaining strength than in which side is stronger today. The gun control debate will not be won or lost in the short term. Instead, momentum over time will prove to be more important than current strength.

I hypothesize that both gun control advocates and gun rights advocates are energized and mobilizing to grow their base. To test these hypotheses, I look at how the group performs relative to itself by removing initial group size from the dynamics and giving a view of momentum. Then we compare the momentum across groups to see if one or the other has an advantage long-term.

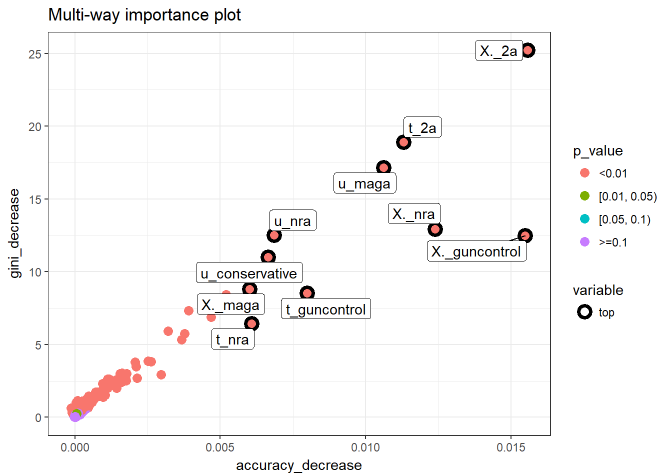
**Collection**

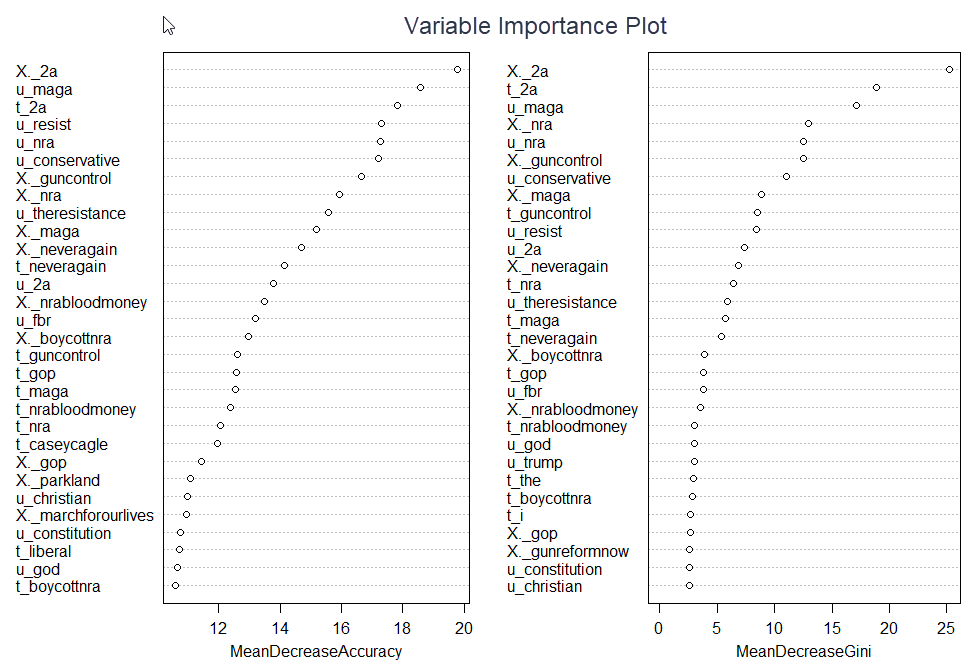
The study analyzed over 400k original tweets by 67k unique users for #NRA and #GunControl across two weekends (Feb 24-25, Mar 3-4) and screened to identify tweets by bots. No tweet bot activity was identified in the pool of original tweets.

The machine learning algorithm RandomForest in R categorized tweets as “Gun Control”, “Gun Rights”, or “Neither”. A manual classification of 1,900 tweets created a golden set to train the algorithm.

**Machine Learning**

Python extracted each word of the user description, hashtag, and text in the golden set and then allowed the random forest algorithm to determine which words helped in the classification of a tweet. The golden set has 21,000 unique combinations of words and section of a tweet. That list was reduced to 3,500 impacting the algorithms ability to classify a tweet.

The Multi-way Importance Plot and the Variable Importance Plot illustrates the most important variables in the categorization algorithm. The leading X.\_ indicates the variable originated in a hashtag, the leading t\_ indicates the variable originated in a text, and the leading u\_ indicates the variable originated in a user description.

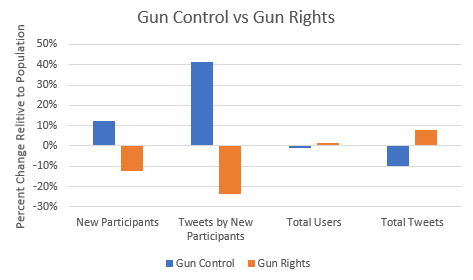


The random forest model consisted of 3,500 variables, each with two factors, achieved an error rate of 14.8%. Meaning that on average 85 out of 100 tweets from #NRA or #GunControl will be correctly identified as in support of Gun Control, Gun Rights, or Neither. Given that there are misspellings, slang, sarcasm, and symbols used in tweets, that is very good. Looking at the misidentified tweets in the golden set the algorithm struggles to interpret the tone of a tweet.

After the model was built, the raw tweets could be analyzed to determine which of the 3,500 variables applied to each tweet. Due to the resource intensive process, 8 large files were broken into 123 small files, and processed across 3 computers to reduce processing time. Once the tweets were prepared for evaluation by the random forest algorithm, R code was written to pull the files one by one, apply the algorithm to the tweets, and then combine all the small files into a new file for each day of the week (2/24, 2/25, 3/3, 3/4).

**Conclusion**

Week one was a higher tweet weekend for the gun debate vs. week two, however I was still able to determine which of the two sides fared better as far as attracting new members and engagement level of the members. Resulting in the following findings:

e

From this small sample, it appears that the Gun Control side of the debate is better at getting new people to join the ranks and tweet, but the Gun Rights side is better about keeping all of their people engaged and generating more original tweets on average than the gun control advocates.

The study would benefit from the following:

* Improving the algorithms ability to pick up on ridicule and sarcasm.
  + Further research would be required to identify potential algorithms for discerning tone.
* Pulling tweets every day for months and then looking at month to month trends to better gage momentum.
  + Requiring constant training of the model, due to future events creating new variables of significance.

**Summary of rejected hypotheses:**

X-squared = 72.585, df = 1, p-value < 0.00000000000000022

The evidence supports rejecting Ho at the .01 significance level. There is a greater percentage of new users joining the conversation for Gun Control.

-----------------------------------------------------------------------------------------------------

X-squared = 1292.4, df = 1, p-value < 0.00000000000000022

The evidence supports rejecting Ho at the .01 significance level. There is a greater percentage of tweets in support of Gun Control that comes from new users vs percentage of tweets from all new users.

-----------------------------------------------------------------------------------------------------

X-squared = 4.2146, df = 1, p-value = 0.02004

The evidence supports rejecting Ho at the .05 significance level. There is a greater percentage of users supporting Gun Rights.

-----------------------------------------------------------------------------------------------------

X-squared = 612.14, df = 1, p-value < 0.00000000000000022

The evidence supports rejecting Ho at the .01 significance level. There is a greater percentage of original tweets supporting Gun Rights.