Naïve Bayes Classifying Twitter Bots

Steven Hansen

Abstract

Tweets from various sources are compared to determine if a simple Naive Bayes model can 'learn' to distinguish between a potential propaganda account from the Russian Internet Research Agency (RIRA) and a regular account.

Three models were constructed using three different non-RIRA sources:

- Random Set (from: https://www.kaggle.com/vikasg/russian-troll-tweets?select=tweets.csv)
- Corpus Set (from: nltk.corpus corpus labeled "twitter_samples")
- Randomized tweets scrapped from top ten twitter accounts (see code for details

Then the Twitter accounts of US Congress members were tested and compared using each model. The model trained with the top ten accounts was the most successful in distinguishing non-RIRA accounts from RIRA accounts – this is likely because the top ten account includes US presidents Barack Obama and Donald Trump.

To improve future models it is recommended to obtain a wider range of twitter accounts in training.

Motivation

The 2016 American Presidential election was fraught with controversy surrounding potential Russian influence. In the aftermath Twitter identified accounts associated with the Russian Internet Research Agency (RIRA). Twitter banned these accounts and released the tweets and associated meta data.

Machine learning is frequently used to identify bots/malicious accounts. I would like to try to understand if a simple model can distinguish between these RIRA accounts and non-RIRA accounts or if a more complex analysis would be required. Additionally I was curious if any members of congress would be identified as RIRA accounts.

Datasets

1. RIRA tweets were obtained from https://www.kaggle.com/vikasg/russian-troll-tweets?select=tweets.csv. Only the 'tweets.csv' file was used which was 56 MB

2. The set of tweets referred to as 'rand' in the data were obtained from https://data.world/data-society/twitter-user-data. This set was 59 MB

3. The set of tweets referred to as 'corpus' were retrieved from the nltk.corpus corpus labeled "twitter_samples"

4. Tweets from the US Congress and Top Ten Twitter accounts were scrapped from Twitter using code adapted from https://towardsdatascience.com/twitter-data-collection-tutorial-using-python-3267d7cfa93e

    a) Top ten twitter accounts based on number of followers according to Wikipedia: @barackObama, @justinbieber, @katyperry, @Rhianna, @taylorswift13, @Cristiano, @realDonaldTrump, @ladygaga, @TheEllenShow, @ArianaGrande

Data Preparation and Cleaning

The data sets were relatively easy to use and required little cleaning. Preparation included creating seperate files based on twitter handles and joining data sets. Attempts were made to join the Congress twitter set with data sets that listed congress by party affiliation; however, these attempts were not successful. If more time was allotted I would have like to completed this step.

Research Questions

Can a simple Naïve Bayes Classification model be used to classify Twitter accounts associated with RIRA?

How would members of congress be classified using these models?

Methods

Three models were trained using the RIRA set and three different non-RIRA sources. The models were then tested against tweets from United States Congress members.

Findings

<Feel free to replicate this slide to show multiple findings>

Present your findings. Include at least one visualization in your presentation (feel free to include more). The visualization should be honest, accessible, and elegant for a general audience.

You need not come to a definitive conclusion, but you need to say how your findings relate back to your research question.

Findings – Preliminary Testing of the Models

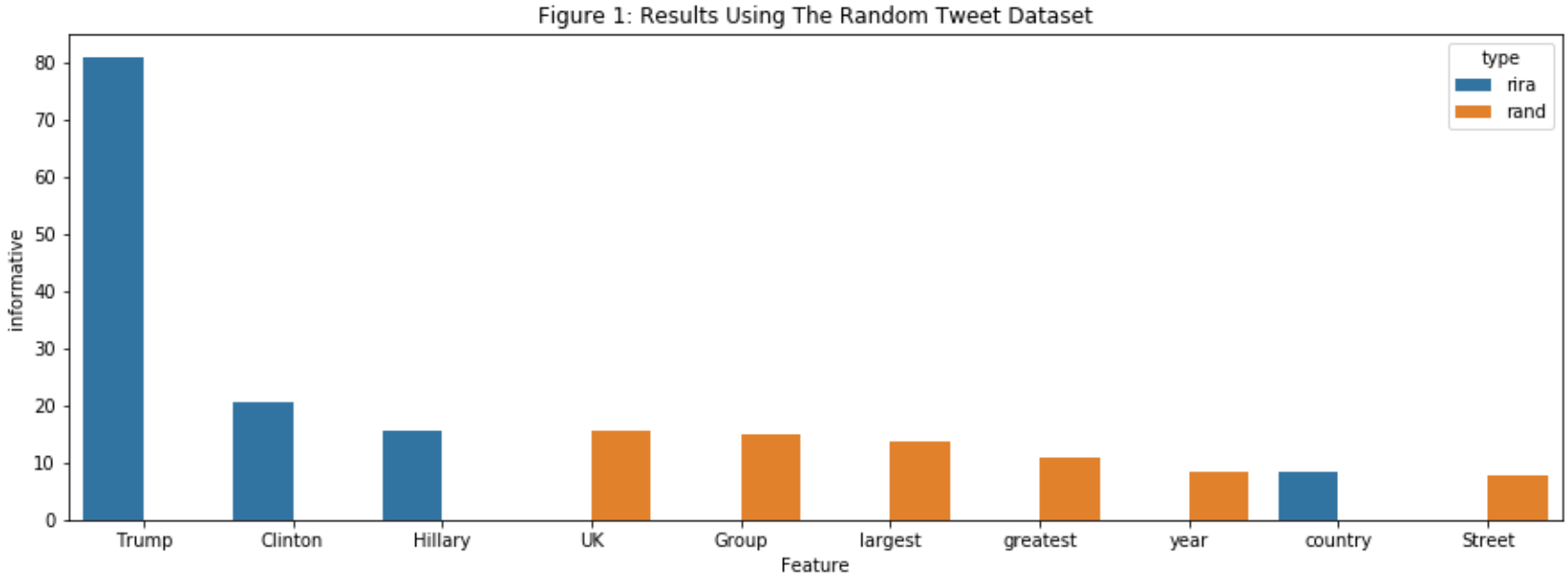
Model	Non-RIRA training Set	Training Accuracy
Rand	Data set from Reference 1	92.8%
Corpus	Data set from Reference 2	88.8%
Top Ten	Scrapped from Top Ten Twitter Accounts	86.4%

We can see that on the training set the Rand Model was the most accurate and the Top Ten model was the least accurate

Models	Classification of challenge tweets			
	"MAKE AMERICA GREAT AGAIN"	"Obama is the best"	"trump is terrible"	"This season of the Office was terrible"
Rand	rira	rira	rira	Non-rira
Corpus	rira	rira	rira	Non-rira
Top Ten	rira	Non-rira	rira	Non-rira

The Rand and Corpus models in general only delineate between political and non-political tweets. The Top Ten model performs a little better – likely because tweets from both Obama and Trump are in the Top Ten data set.

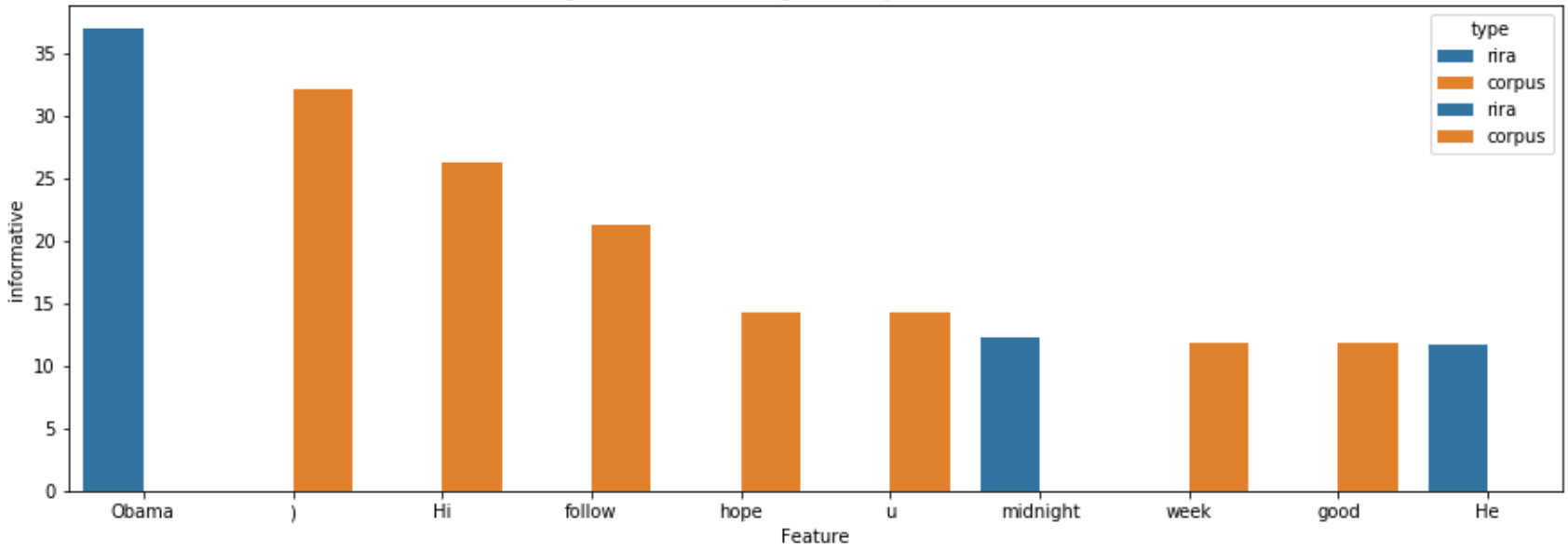
Findings – Rand Model Most Informative Features



Tweets referencing “Trump” were 80:1 more likely to be a RIRA account using the random dataset. This is probably be very few of the random tweets from this set reference Trump.

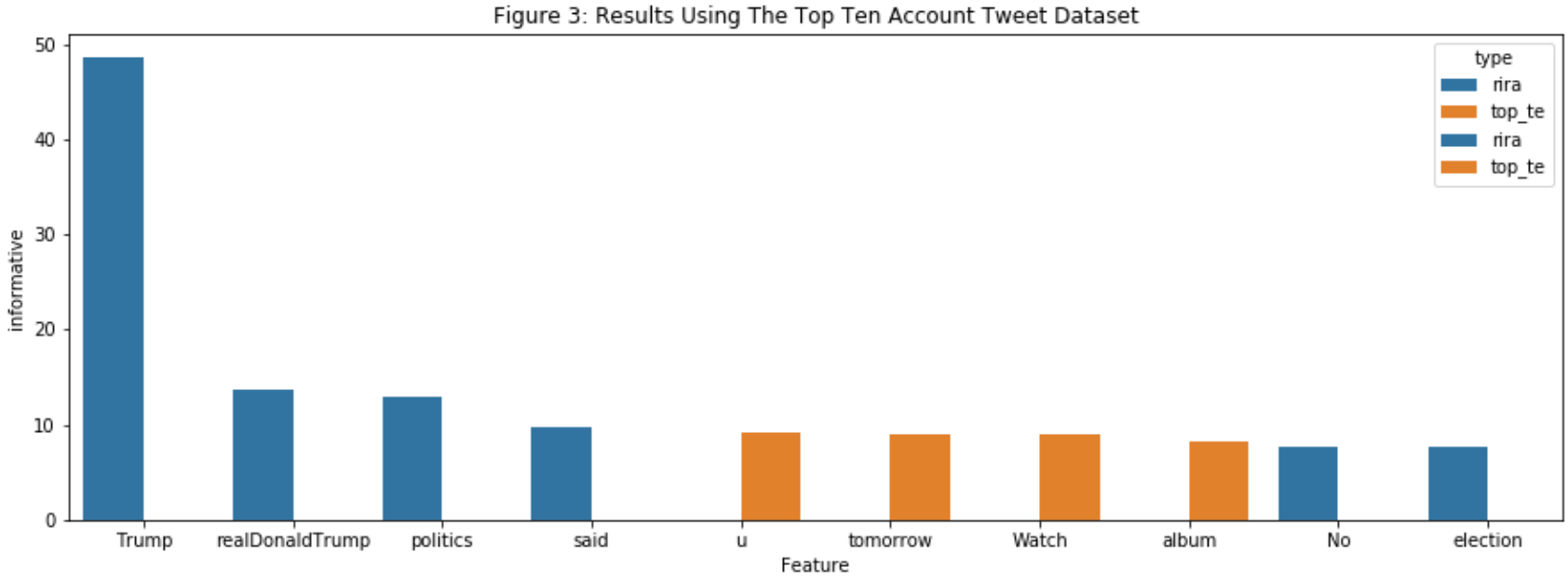
Findings – Corpus Model Most Informative Features

Figure 2: Results Using The Corpus Tweet Dataset



Using the tweets from the Corpus set, the ratios are a bit tighter with references to ”Obama” only occurring at a ratio of 35:1 and “Trump” does not show up as an informative feature.

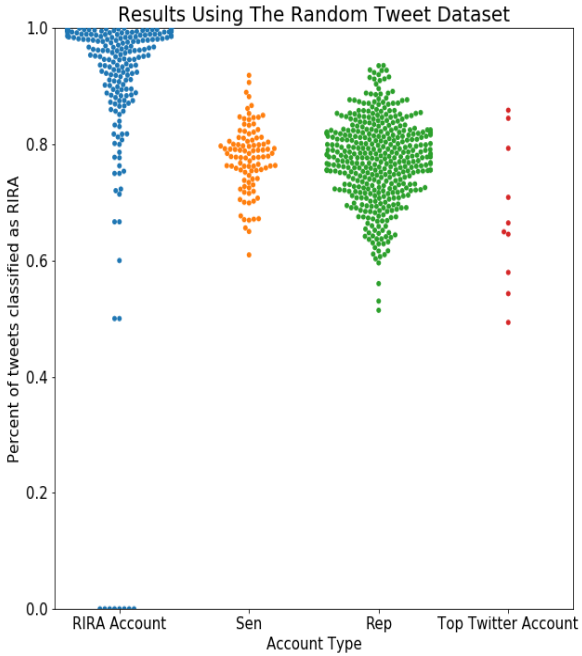
Findings – Top Ten Model Most Informative Features

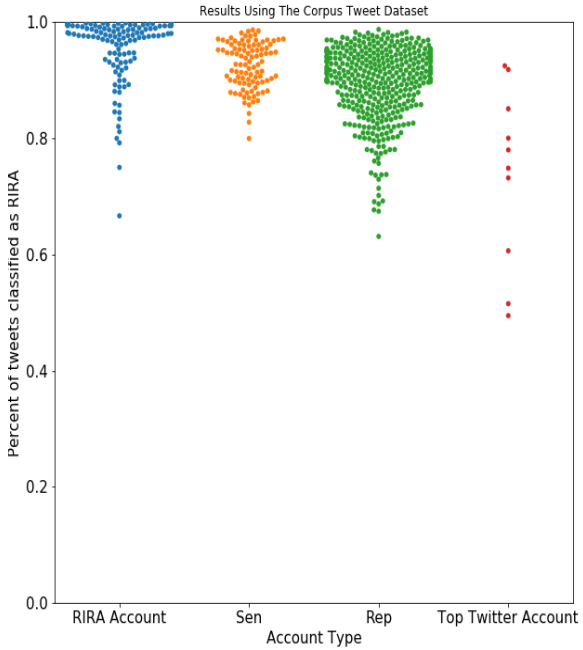


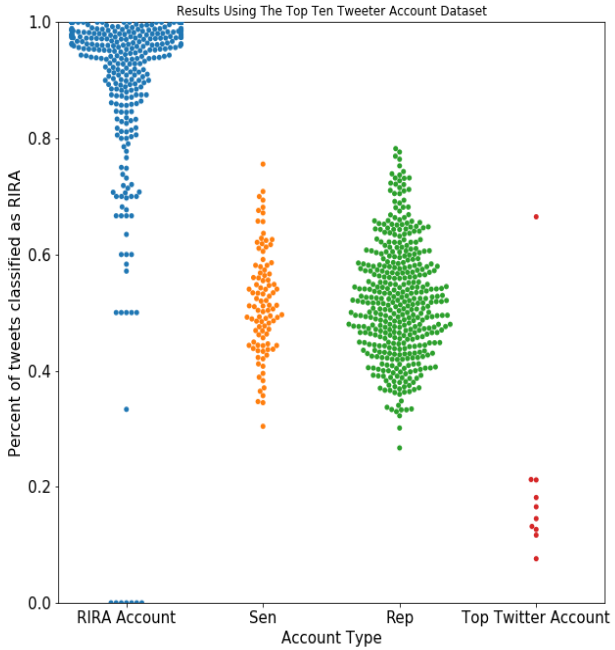
”Trump” appears again as an informative feature; however, the ratio is much smaller

Findings – Model classification of congress

Each dot represents a twitter account and for each account the last 3,000 tweets were analyzed.







Here we can see the improvement when training with the Top Ten Twitter Accounts

Limitations

- The models were trained on 2016 RIRA identified accounts; therefore, it is unlikely to have the same effectiveness in identifying RIRA accounts in 2020.
- The timing of the random and corpus twitter sets are not known and that may add an additional variable not accounted for. For example the RIRA accounts were all identified during the 2016 election, if the random tweets sets were collected significantly before or after they may not be a good representation
- The models were built using only the tweet text. The meta data available may improve the models.
- The models rely on Twitter's ability correctly to identify RIRA accounts

Conclusions

This was a great way to learn how to scrap twitter and build a simple model that could potentially be used to classify tweets; however, I have low confidence in these simple models to identify RIRA accounts. Better test data would be required as well as potentially incorporating more meta data associated with the accounts.

Acknowledgements

The code used to scrap twitter accounts was adapted from tps://towardsdatascience.com/twitter-data-collection-tutorial-using-python-3267d7cfa93e

See Reference section for other data set credits.

All other work was original.

References

1. RIRA tweets were obtained from https://www.kaggle.com/vikasg/russian-troll-tweets?select=tweets.csv. Only the 'tweets.csv' file was used which was 56 MB

2. The set of tweets referred to as 'rand' in the data were obtained from https://data.world/data-society/twitter-user-data. This set was 59 MB

3. The set of tweets reffered to as 'corpus' were retrieved from the nltk.corpus corpus labeled "twitter_samples"

4. Tweets from the US Congress and Top Ten Twitter accounts were scrapped from Twitter using code adapted from https://towardsdatascience.com/twitter-data-collection-tutorial-using-python-3267d7cfa93e

    a) Top ten twitter accounts based on number of followers according to Wikipedia: @barackObama, @justinbieber, @katyperry, @Rhianna, @taylorswift13, @Cristiano, @realDonaldTrump, @ladygaga, @TheEllenShow, @ArianaGrande