Analysis of the Structure and Predictive Capabilities of Russian Tweet Trolling

Michelle Yuen, Josh Gardner, Michael Gao, Avinash Boppana

Motivation

- The impact of "fake news", on social media, particularly those disseminated by Russian troll tweeters, has become increasingly important
- We want to use both supervised and unsupervised learning techniques to analyze the structure of the data of troll tweets, answering questions such as the most significant content and characteristics of high-volume accounts, what types of content troll tweets tend to have, and differences between left and right-wing troll accounts

Background

- Dataset was curated and published by FiveThirtyEight in partnership with Clemson University professors Darren Linvill and Patrick Warren
- Previous work has already been done on common topics associated with Russian troll tweets and being able to identify troll accounts
  - Group at University of Michigan used keywords in tweets to distinguish between troll and non-troll accounts with 99% account
  - University of Canberra study used LDA to identify 26 most common topics of tweets
  - Clemson University compiled a list of 2,973,371 tweets associated with the Russian Internet Research Agency sent by 2848 Twitter accounts. This important database contains key metrics beyond the tweet itself, such as the political orientation of the tweets and the number of followers, retweets, and followed accounts

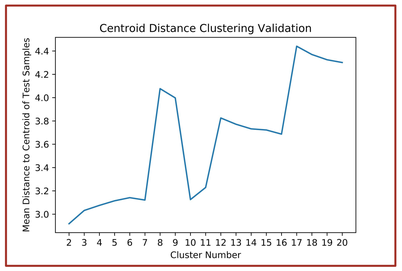
Methodology

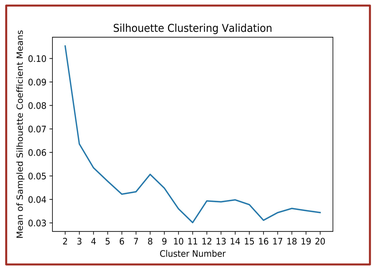
1. Unsupervised Learning: Use K-means clustering and LDA to identify which keywords in each tweet were most commonly associated, gives classification of tweets

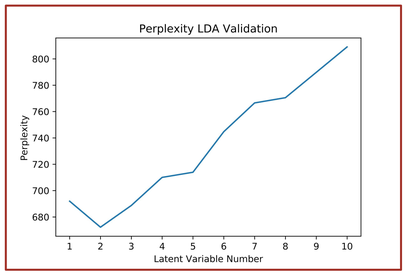
2. Supervised Learning using Categorical Learning: Used AdaBoost, Naive Bayes, and Random Forest to successfully distinguish left and right trolls (used cross-validation to tune hyperparameters for Random Forest) assessed against a statistical baseline of 50:50 random guess

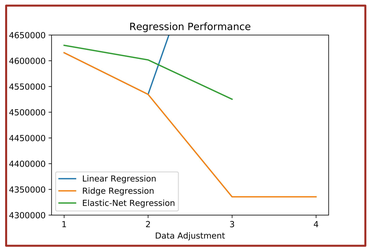
3. Supervised Learning using Continuous Models: Used Linear and Ridge regression to build models to predict the number of followers of accounts using parameters such as country of origin, as well as the contents of the tweet, and their associations as determined by LDA in Part 1

Results: Model Performance

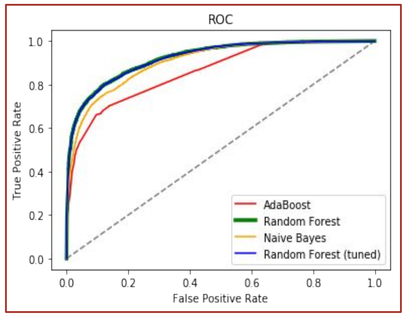








Data Adjustment
1: Numerical + One-Hot-Encoded Features
2: Latent Variable Memberships + 1
3: Bag-of-Words + 2 + 1
4: Parameter Tuning (Alpha = 1)

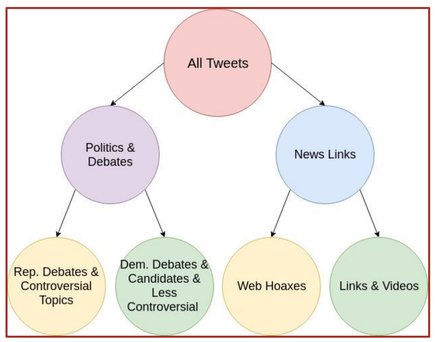


Model Performance on Fearmonger Tweets
Of 9671 tweets by fearmongers, the Random Forest model classified 4996 of them as Right Trolls (0.51659)

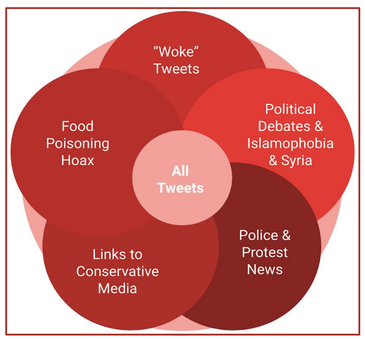
	Accuracy	Precision	Recall	Specificity
Baseline	0.5073	0.5028	0.5028	0.5055
AdaBoost	0.7792	0.8381	0.6971	0.8653
Naive Bayes	0.8126	0.8177	0.8038	0.8208
Random Forest	0.8328	0.8444	0.8261	0.8478
Random Forest (tuned, n=95)	0.8393	0.8532	0.8310	0.8572

Results: Structure of Data

Cluster Tree



Latent Tweets



Results: Important Features

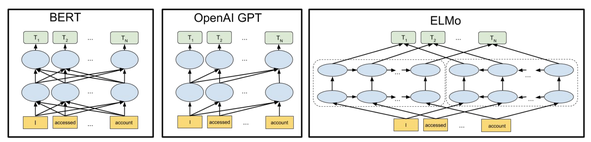
All Troll Tweets							
Cluster 1				Cluster 2			
Cluster 1-1		Cluster 1-2		Cluster 2-1		Cluster 2-2	
Feature	Factor	Feature	Factor	Feature	Factor	Feature	Factor
islamkil	0.005455	demndeb	0.903822	rt	0.039047	co	0.696609
blacklivesmatt	0. 004242	demdeb	0.571992	kochfarm	0.023584	http	0.691616
rt	0.004030	carson	0.088262	turkey	0.014553	trump	0.130711
gopdeb	0.003977	hillari	0.063608	blacklivesmatt	0.011572	'	0.084598
vegasgopdeb	0.003228	pari	0.058663	usda	0.010011	"	0.061992
love	0.003111	ben	0.054151	peopl	0.008768	‘‘	0.061448

Heavily Weighted Regression Features	
Feature	Weight
altright	2706.698687
amber	2624.121964
berkeley	2610.878235
daca	-2125.574106
donlemon	2737.293083
irma	-2262.143364
manchest	2218.450844
nytim	2879.226464
sandra	-3393.529541
steeler	-2009.780956

Binary Classification Most Important Features	
AdaBoost	Random Forest
trump	rt
https	https
rt	blacklivesmatter
vegas	black
vegasgopdebate	trump
foxnews	blacktwitter
cop	hillary
liberal	staywoke
antifa	breaking
blacktwitter	police

Future Work

The bag of words model could be improved by using bigrams or n-grams. However for a dataset of this size, this could be computationally difficult and require special hardware. Nevertheless, it would likely yield more revealing analyses. Additionally, we could employ POS-tagging and word vectors, as well. Furthermore, there are sophisticated NLP tools which avoid many of the problems of bag-of-words or n-grams. Google Research's Bidirectional Encoder Representations from Transformers (BERT) has achieved bleeding edge result on various sentiment analysis and classification tasks and could be used to analyze troll tweets.



Citations

Badawy, Adam, Emilio Ferrara, and Kristina Lerman. "Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign." *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018.

Jensen, Michael. "RUSSIAN TROLLS AND FAKE NEWS: INFORMATION OR IDENTITY LOGICS?." *Journal of International Affairs* 71.1.5 (2018): 115-124.

Boatwright, Brandon C., Darren L. Linvill, and Patrick L. Warren. "Troll factories: The internet research agency and state-sponsored agenda building." *Resource Centre on Media Freedom in Europe* (2018) .

Roeder, Oliver. "Why We're Sharing 3 Million Russian Troll Tweets." *FiveThirtyEight*, FiveThirtyEight, 31 July 2018, fivethirtyeight.com/features/why-were-sharing-3-million-russian-troll-tweets/.

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).