Team Troll Trappers: Laura Pintos, Ramiro Cadavid, and Anna Jacobson

W207 | Applied Machine Learning | Spring 2019

The Askeladden Algorithm

Background

In February 2019, as part of special counsel Robert Mueller’s investigation of the Russian government's efforts to interfere in the 2016 presidential election, the United States Department of Justice charged 13 Russian nationals with illegally meddling in American political processes.

The defendants worked for a well-funded “troll factory” called the Internet Research Agency (IRA), which reportedly had 400 employees, or “trolls”, working 12-hour shifts from a nondescript business center in St. Petersburg.

The IRA ran a sophisticated, coordinated campaign to spread disinformation and sow discord into American politics via social media, often Twitter.

Troll Tweets

Twitter has identified and suspended thousands of these malicious accounts, deleting millions of the trolls' tweets from public view on the platform. While other news outlets have published samples, it has been difficult to understand the full scale and scope of the IRA's efforts, as well as the details of its strategy and tactics. According to Alina Polyakova, a foreign policy fellow at the Brookings Institution,

"Wiping the content doesn't wipe out the damage caused, and it prevents us from learning about how to be better prepared for such attacks in the future."

To address this problem, and "in line with our principles of transparency and to improve public understanding of alleged foreign influence campaigns," Twitter has now made publicly available archives of Tweets and media that it believes resulted from potentially state-backed information operations.

Why We Care

1

INTENT

The IRA showed a clear intent to influence the 2016 US election, exhibiting a strong and consistent preference for Donald Trump and negative content about a wide range of other Republican candidates, as well as Hillary Clinton.

2

IMPACT

There were approximately 109 Twitter accounts masquerading as news organizations, including U.S. local news outlets. The 44 US accounts had amassed 660,335 followers between them, with an average of 15,000 followers each. Clinton needed only 53,650 swing votes to have won in 2016.

3

EFFECTIVENESS

Troll tweets employed voter suppression tactics that included malicious misdirection, candidate support redirection, and voter turnout suppression.

4

POTENTIAL

The trolls' threat remains; there is evidence of continued interference operations across social media platforms.

Project Objective

Let's help Twitter catch some trolls!

or

Can we build a machine learning algorithm that can identify a troll from its tweets?

The Data

01 | Troll News

Twitter IRA Trolls Dataset

- 3,077 unique Internet Research Agency (IRA) accounts
- Approximately 8M unique tweets
- Approximately 3M unique English language tweets
- May 2009 to June 2018

“Fake News” IRA Trolls Dataset

- 2016 to 2018
- 33 unique user accounts with “fake news” screen names such as TodayNYCity, ChicagoDailyNew, and KansasDailyNews
- 296,949 unique English language tweets

02 | Real News

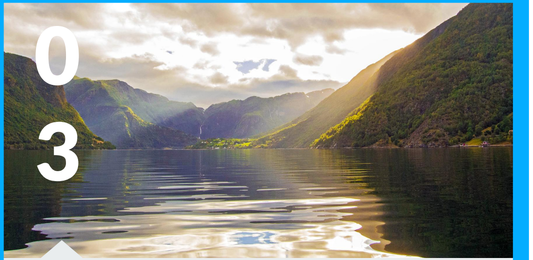
Harvard Dataverse News Dataset

- Tweet IDs of 39,695,156 tweets
- Approximately 4,500 news outlet user accounts
- August 4, 2016 and July 20, 2018

"Real News" Dataset

- 49 unique user accounts from a variety of news outlets, including The Hill, Politico, Fox News, CNN, The Economist, and MSNBC
- 153,188 unique tweets

03 | All the News



"Fake News" IRA Trolls Dataset

153,188 unique tweets (randomly selected from 296,949)
2016 to 2018
33 unique user accounts

"Real News" Dataset

153,188 unique tweets
2016 to 2018
49 unique user accounts

"All the News" Dataset

306,376 unique tweets
2016 to 2018
82 unique user accounts

The Models

Preliminary Models

Default Parameters

Model	Bernoulli Naive Bayes	Logistic Regression			Ensemble *(Logistic Regression, Linear SVC, Bernoulli NB, Ridge, Passive Aggressive)*	Random Forest	Doc2Vec		
		Unigram	Bigram	Trigram			DBOW	DMM	DBOW+ DMM
Count Vectorizer	86.08	88.23	89.52	89.35	89.81	86.81	82.33	72.66	83.38
TFIDF Vectorizer	86.08	87.89	87.83	87.77	-	-			

Optimizations

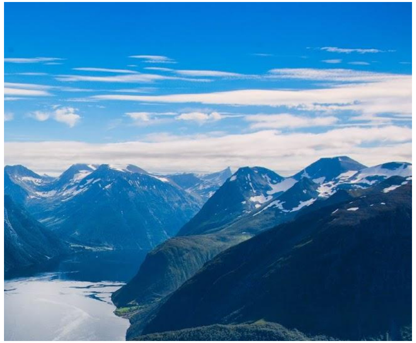
01 | C Value

02 | Stop Word Removal

03 | Custom Preprocessing

04 | min_df

05 | max_features



The Troll Tactics

Camouflage

Many troll tweets seem to be innocuous, focused on seemingly uncontroversial topics. 8 of the 12 "trolliest" tweets (shown below) are about sports.

Local sports digest: San Jose State third, Cal fourth after first round of Bay Area Intercollegiate #sports

#Idlib | Renewed clashes between Suqour al-Sham Brigade (Ahrar al-Sham) and Jabhat Fatah al-Sham near Ehsim town https://t.co/hX6OiW8alk

Bay Area sports fans sought for Super Bowl 50 Halftime Show #news

Kyrie Irving continues to progress for Cleveland Cavs: Cleveland Cavaliers point guard Kyri... https://t.co/i7pWHybj11 #Cleveland #sports

San Jose Sharks beat Edmonton Oilers 2-1 in shootout, Todd McLellan returns to San Jose #news

#Hama #Idlib | Russian airstrikes on Tahrir al-Sham and Liwa al-Aqsa positions in Jisr al-Shughur and al-Ghab Plain https://t.co/AQbH94B1cV

Missouri business groups divided on LGBT-discrimination ban #business #news

Cleveland ends 52-year drought without championship, Cavs beat Warriors to win NBA Finals https://t.co/h5K3KoyxKM #news

Shumpert is Cleveland Cavs' 'glue guy': Shooting guard\xa0Iman Shumpert is the Cleveland Caval... https://t.co/qCZz8aGcie #Cleveland #sports

Biggest Winner for February 10, 2016: The "Biggest Winner" in local sports for Wednesday Fe... https://t.co/Xx6G71UAcd #Cleveland #sports

Local sports digest: Santa Clara women's basketball coach Payne departs #sports

Footage of the confrontations between #IS forces & #SAA west of #Palmyra #Syria 21/03/2016 https://t.co/hV1ALzPxYH https://t.co/0lrLQTdU0u

Confusion

Troll tweets often sound like real news tweets. Of the 20 most confused tweets, 80% are troll tweets misclassified as real news.

Predicted Label: REAL
True Label: TROLL

It is a national disgrace that at 21.8 percent, the U.S. has the highest childhood poverty rate of any major country https://t.co/tSJUQO7PZa

Predicted Label: REAL
True Label: TROLL

The racist organizer of the #UniteTheRight rally in #Charlottesville was chased by protesters during a press conference. https://t.co/qY6lIlQ1Oy

Predicted Label: REAL
True Label: TROLL

.@EricTrump: "The fact that we have a presidential candidate who could be under indictment while in the White House is unthinkable." https://t.co/dSd4SWRDp1

Predicted Label: TROLL
True Label: REAL

5 things for Monday May 1st: 1. Donald Trump 2. Weather 3. Turkey 4. Politics 5. Milwaukee jail death... https://t.co/RIUfzZ4OvK

Predicted Label: TROLL
True Label: REAL

Nigerian-born, Brooklyn-based artist Laolu Sebanjo wants his art to start conversations around politics, religion,... https://t.co/iyvBhIR59o

Predicted Label: TROLL
True Label: REAL

Local medical caregivers to speak at Health Partners Gala - https://t.co/Ihgvj9YFyH https://t.co/5swzofpsVc https://t.co/i0NnvTqoEY

Variation

The top fifty most predictive features of a troll tweet are seemingly varied. However, upon closer examination, patterns emerge.

politics	NewYork	Trump2016	tech	Aleppo
Observations AtTheMarathon	business	local	LiberalLogic	SDF
StLouis	todayinsyria	RT Pamela_Moore13	Damascus	Roberts
showbiz	RT TEN_GOP	Pamela_Moore13	S A	Cardinals
sports	TEN_GOP	Cleveland	Phoenix	NeverHillary
SanJose	FeelTheBern	IS	DeirEzzor	Art
news	SAA	MakeAmericaGreatAgain	WakeUp America	Oakland
The Latest	money	Akron	PHX	RT if
health	hockey	KC	Saints	FSA
entertainment	Montini	IslamicState	Wichita	Cincy

Variation | Generic Words

politics	NewYork	Trump2016	**tech**	Aleppo
Observations AtTheMarathon	**business**	**local**	LiberalLogic	SDF
StLouis	todayinsyria	RT Pamela_Moore13	Damascus	Roberts
showbiz	RT TEN_GOP	Pamela_Moore13	S A	Cardinals
sports	TEN_GOP	Cleveland	Phoenix	NeverHillary
SanJose	FeelTheBern	IS	DeirEzzor	**art**
news	SAA	MakeAmerica GreatAgain	WakeUp America	Oakland
The Latest	**money**	Akron	PHX	RT if
health	**hockey**	KC	Saints	FSA
entertainment	Montini	IslamicState	Wichita	Cincy

Variation | Political Slogans

politics
Observations AtTheMarathon
StLouis
showbiz
sports
SanJose
news
The Latest
health
entertainment

NewYork
business
todayinsyria
RT TEN_GOP
TEN_GOP
FeelTheBern
SAA
money
hockey
Montini

Trump2016
local
RT Pamela_Moore13
Pamela_Moore13
Cleveland
IS
MakeAmerica GreatAgain
Akron
KC
IslamicState

tech
LiberalLogic
Damascus
S A
Phoenix
DeirEzzor
WakeUp America
PHX
Saints
Wichita

Aleppo
SDF
Roberts
Cardinals
NeverHillary
art
Oakland
RT if
FSA
Cincy

Variation | Syria

politics
Observations
AtTheMarathon
StLouis
showbiz
sports
SanJose
news
The Latest
health
entertainment

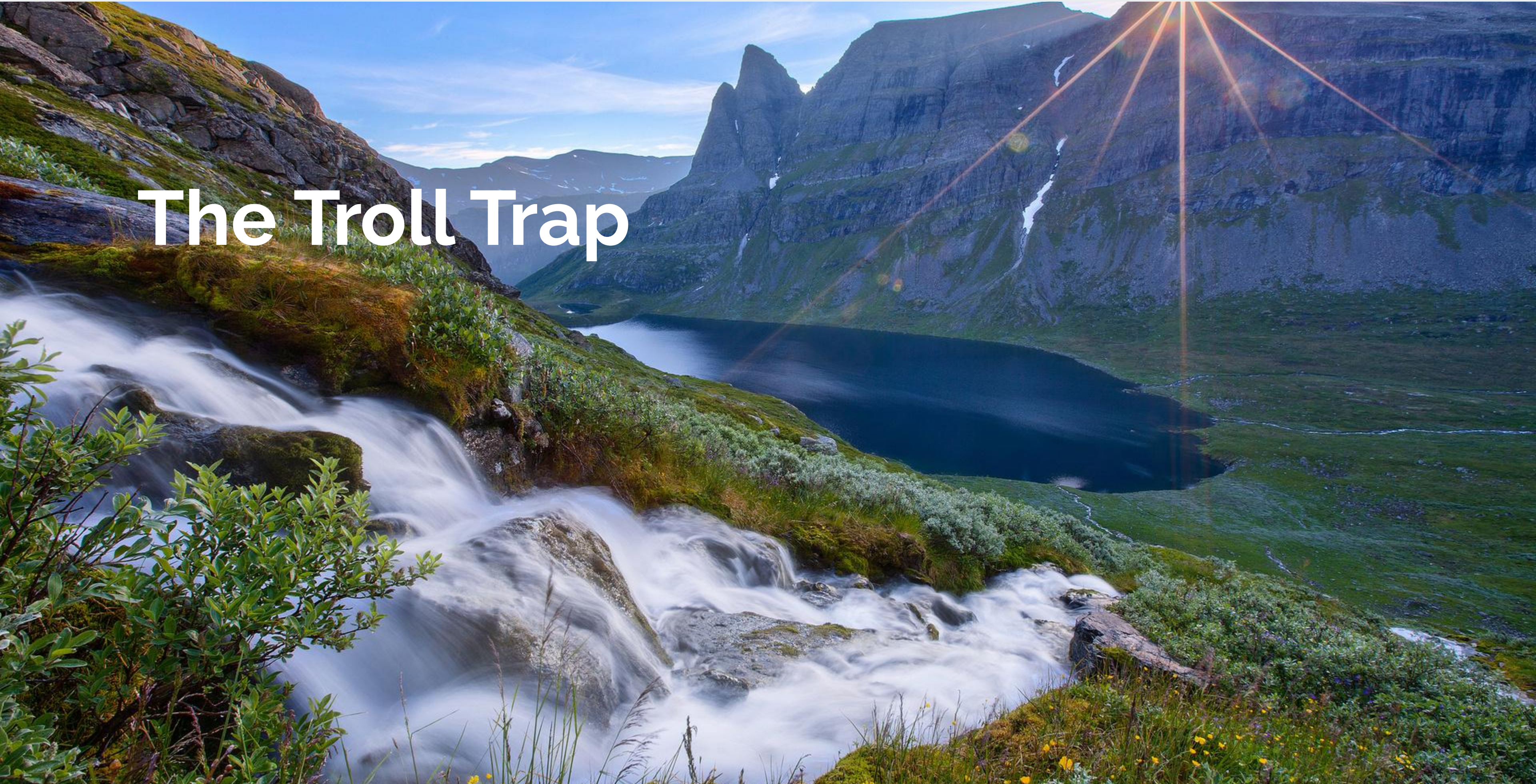
NewYork
business
todayinsyria
RT TEN_GOP
TEN_GOP
FeelTheBern
SAA
money
hockey
Montini

Trump2016
local
RT
Pamela_Moore13
Pamela_Moore13
Cleveland
IS
MakeAmerica
GreatAgain
Akron
KC
IslamicState

tech
LiberalLogic
Damascus
S A
Phoenix
DeirEzzor
WakeUp
America
PHX
Saints
Wichita

Aleppo
SDF
Roberts
Cardinals
NeverHillary
art
Oakland
RT if
FSA
Cincy

The Troll Trap

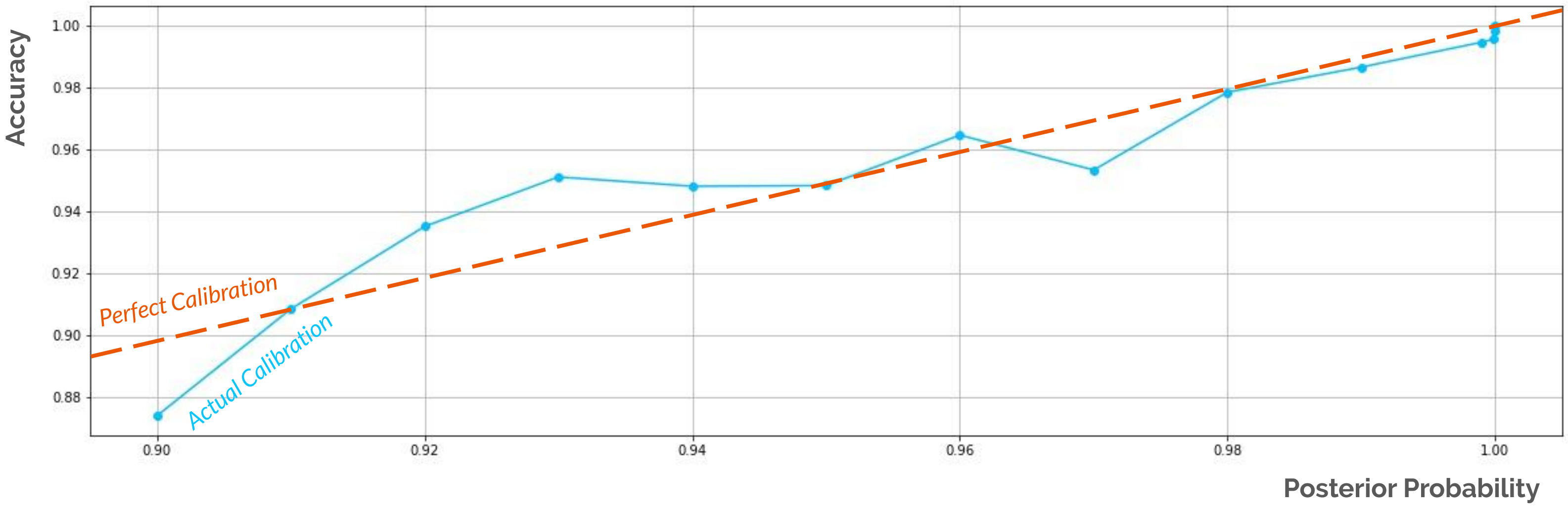


The Final Model

Count Vectorizer	ngram_ range= (1, 2)	token_pattern= r'\b\w+\b'		preprocessor= empty_preprocessor		Accuracy on Test Data 90.32%
Logistic Regression	penalty = 'l2'	C=0.17	solver = 'saga'	multi_class = 'multinomial'	class_weight= 'balanced'	

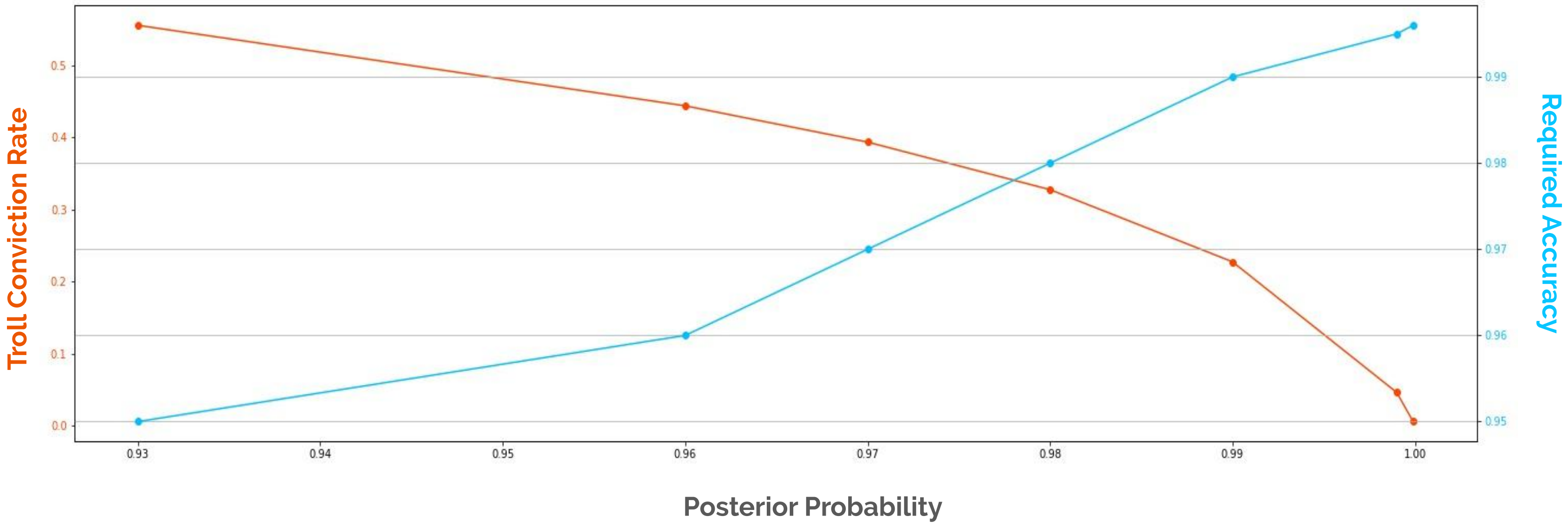
Calibration

Visualizing calibration shows that our algorithm is well-calibrated. Generally, however, we can observe that a higher confidence is associated with higher accuracy.



Confidence Thresholds

In order to ensure that real news isn't mistakenly classified as trollish, we must establish a required level of accuracy to "convict" a troll. Using calibration, we can then set a confidence threshold for conviction using the posterior probability.



Implementation

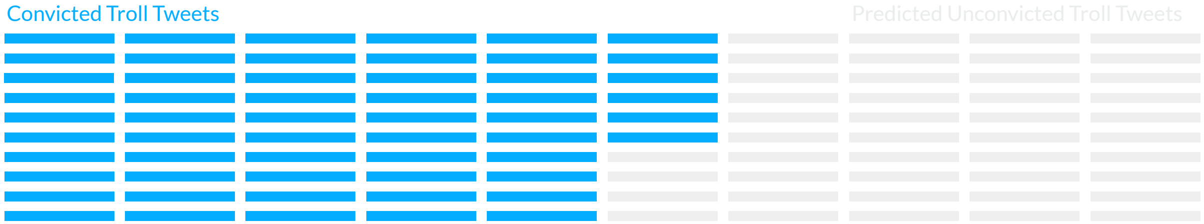
In practice, as the required level of accuracy increases, the confidence threshold also increases, and the number of troll “convictions” decreases.

Required Accuracy

95%

Confidence Threshold

0.93



Required Accuracy

99%

Confidence Threshold

0.99

Lessons Learned

Lesson 01

Data quality is as important as algorithm quality.

Lesson 02

Utility is as important as complexity.

Lesson 03

Process is as important as processing.

Suggestions for Future Work

ML Engineering

- Is there additional preprocessing and/or parameter tuning that would increase the model's accuracy?
- Would additional data improve our model? Since we have almost 500,000 features, it would be good to have at least as many examples in our training set.
- Can our algorithm pick out a troll from a random selection of tweets?
- Would our algorithm work for identifying other trolls (e.g. the Venezuelan trolls)?

Implementation

- How could the algorithm be implemented by Twitter? What other work is required to make the algorithm useful in practice?
- Does the algorithm transcend the specifics of the political context of the training data?
- How could the algorithm be 'tricked' or 'gamed'?

Suggestions for Future Work

Fairness

- Does the algorithm have the potential to falsely identify legitimate users as trolls?
- How could it be misused?

Legality

- What are the legal issues involved with suspending accounts using machine learning?
- Does curtailing troll activity necessarily infringe on free speech? Would this be a bigger concern if the trolls were American citizens?

Research

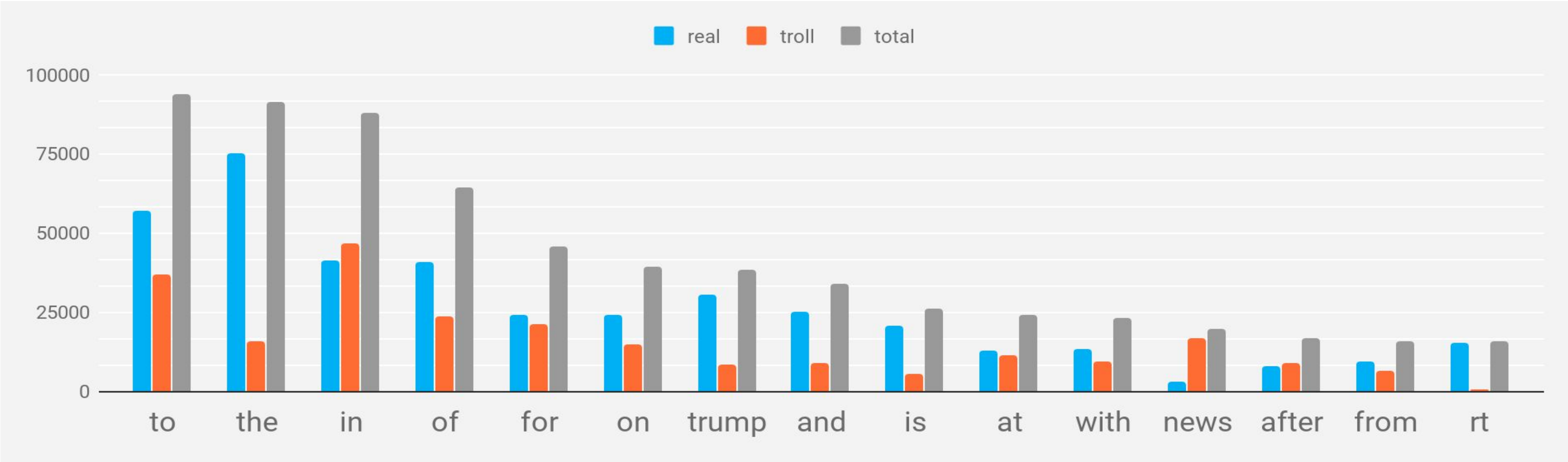
- What was the practical impact of these trolls' tweets? Did they actually influence people to change their votes in 2016?
- Who was directing the IRA? How was it funded?

Questions?

Appendix

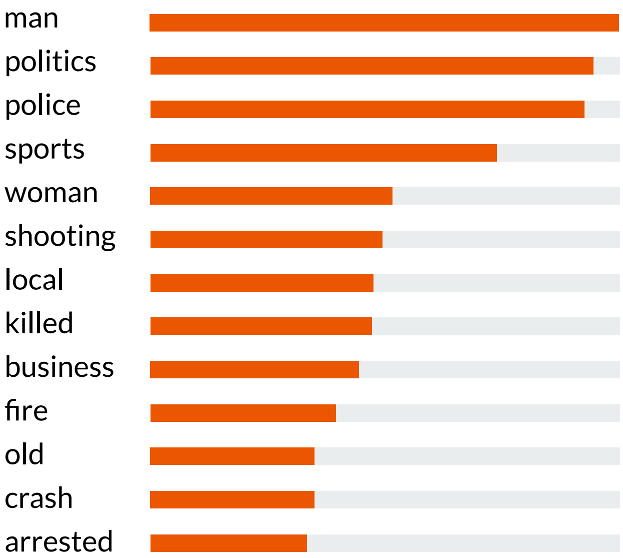
Most Common Features

Using CountVectorizer for feature extraction yields 470,051 features. After "https" and "co", many of the overall most common features are standard stop words. Interestingly, the feature "trump" occurs much more frequently in real news tweets than troll tweets, while the feature "news" occurs much more frequently in troll tweets than in real news tweets.



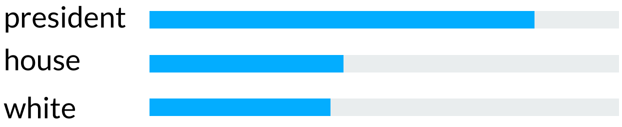
Troll Features

Non-Stop Words



Real News Features

Non-Stop Words



Variation | Troll Handles

- politics
- Observations AtTheMarathon
- StLouis
- showbiz
- sports
- SanJose
- news
- The Latest
- health
- entertainment

- NewYork
- business
- todayinsyria
- **RT TEN_GOP**
- **TEN_GOP**
- FeelTheBern
- SAA
- money
- hockey
- Montini

- Trump2016
- local
- **RT Pamela_Moore13**
- **Pamela_Moore13**
- Cleveland
- IS
- MakeAmerica GreatAgain
- Akron
- KC
- IslamicState

- tech
- LiberalLogic
- Damascus
- S A
- Phoenix
- DeirEzzor
- WakeUp America
- PHX
- Saints
- Wichita

- Aleppo
- SDF
- Roberts
- Cardinals
- NeverHillary
- art
- Oakland
- RT if
- FSA
- Cincy