Supplemental Information for "Asymmetric Flooding as a Tool for Foreign Influence on Social Media"

Table of Contents

A Data

Our data comes from three sources:

1. Twitter's own release of Russian troll *tweets* (available here: https://transparency.twitter.com/en/information-operations.html),

2. Linvill and Warren (2020)'s hand *labels of accounts* (available here: https://github.com/fivethirtyeight/russian-troll-tweets),

3. Hand coded *labels of tweets* we collected through Amazon Mechanical Turk and Figure Eight (data will be made available in replication materials).

A.1 Twitter's Official Tweet and Account Information Data Set

Twitter's Elections Integrity Initiative released public data sets in 2017 and 2018 containing the propaganda efforts led by the Internet Research Agency, a so-called "troll factory" reportedly linked to the Russian government (Bertrand, 2017). The data sets contain more than 9 million tweets sent by over 3,600 accounts affiliated with the Internet Research Agency (IRA), a Kremlin-based Russian troll farm. These accounts represent the efforts of human-controlled Russian operators, or "trolls", as opposed to computer-controlled accounts (or "bots"). Out of these accounts, Twitter originally established that 2752 were operated by the IRA (United States Senate Committee, 2017). In January 2018, this list was expanded to include 3814 IRA-linked accounts (Twitter, 2018b), of which over 3,600 are still identified by Twitter as IRA-linked accounts. We use the combined data released in 2018 and available here under "Internet Research Agency" (enter email at bottom to access): https://transparency.twitter.com/en/information-operations.html.

The released data contains all tweets and metadata, as well as profile information, for all IRA-linked accounts. Tweets and metadata contain the full text of a tweet, including hashtags, user mentions, and links (which are part of the tweet text), as well as time posted and whether the tweet was a "retweet." Note that some retweets originate from other trolls, and some tweets that are not retweets are also not original texts – they can originate from other sources without a reference to those sources.

The data also contains numerous posted images and videos (274 gigabytes). The images are not directly studied in our analysis, but accompanying tweet texts are included in our text analyses. For example, if a user posts a photo/link to photo and comments on the photo/link to photo, then the comment is included in our analysis.

Note the data set released by Twitter does not include 'liked' tweets – i.e. tweets posted by other accounts that the trolls then clicked on to "like" without actually retweeting.

Many of the 9 million tweets were posted after the election or were not in English. In section A.4 below, we describe the removal of post-election tweets and tweets not in English.

A.2 Linvill and Warren Account Categories

Linvill and Warren (2020) use unrestricted open coding to classify accounts into categories: Right Troll, Left Troll, News Feed, Hashtag Gamer, Fearmonger, Commercial, Unknown, and Non-English. Right Trolls posted right-leaning, populist, and nativist messages as well as about Trump; Left Trolls tweeted support of the left, socially liberal values, and Black Lives Matter; News Feed accounts mimicked local news stations and served as news aggregators; Hashtag Gamer accounts posted hashtag games to promote various hashtags; and Fearmonger accounts promoted a specific instance of fake news, related to salmonella-contaminated turkeys, near the 2015 Thanksgiving holiday.

We use four of the categories (Right Troll, Left Troll, News Feed, Hashtag Gamer) because the remaining accounts were largely inactive in 2016 or did not tweet in English (see Figure A10 for numbers of tweets, and Figure A1 for results including all English language categories active at all in 2016). Throughout our analysis, we refer to Right Troll accounts as conservative accounts, and Left Troll accounts as liberal accounts.

A.3 Our Hand-Coded Tweet Categories

To validate the text scaling method we used to identify categories of apolitical language, we hired Amazon Turk users to code a random sample of 450 left and 450 right account (see A.2) tweets posted between June 2016 and the 2016 election. This hand coding process is described in detail in section C.1. We then use supervised machine learning (see C.2) to label the remainder of the tweets in Twitter's IRA data (described above), including news and "hashtag gamer" tweets. Note, however, that inferences based on the supervised labels did not substantively differ from the hand labeled data alone – we are simply able to study more tweets and accounts over a longer time period. We compare these results in section C below.

A.4 Data Merge and Processing Details

A.4.1 Merging Twitter Official Release to Account Categories

In all of our analyses, we use the data set released by Twitter itself. Twitter's data included the complete histories of the troll accounts, while researcher collected data (both our own and the publicly available Linvill and Warren data) would typically be limited by Twitter's API constraints.[14]

The data in Twitter's release was partially anonymized – the user IDs of accounts with fewer than 5,000 followers were replaced with hashed versions of the user IDs. This prevents us from linking the Linvill and Warren account categories based on user ID alone.

However, Twitter's data set was not anonymized on any other identifiers, including tweet

[14]The Linvill Warren data contains just over 1 million tweets from before the 2016 election and from accounts using English. We add around 1 million tweets to their corpus using this merge procedure.

IDs. Because the Tweet IDs are unique, and unique to a user, linking on tweet IDs allowed us assign user categories to all users appearing in the Twitter data, as long as at least one of their tweets appeared in the Linvill and Warren data. We illustrate this merger below (Table A1).

Twitter's Data			LW's Data		
user ID (hashed)	tweet ID		user ID	tweet ID	acct. category
X	1		a	1	left troll
X	2				
Y	3		b	3	hashtag gamer
Y	4				

Merged Analysis Data		
user ID	tweet ID	acct. category
X	1	left troll
X	2	left troll
Y	3	hashtag gamer
Y	4	hashtag gamer

Table A1: *Example Twitter - Linvill Warren data merger.* Data were merged using the tweet ID column.

A.4.2 Pre-Election, English-Speaking Data

We analyze tweets in Twitter's official data set that were posted or retweeted by the troll accounts before the election on November 8, 2016. We also remove non-English accounts, for example those using the Russian alphabet.[15] In the main text, we focus our text analysis on troll messaging during the general election, and so present analyses based on tweets posted after January 1, 2016 – further analyses are included in the appendix. For hand labeled data,

[15]Non-English accounts: 1) account language set to language other than English, 2) labeled non-English by Linvill Warren, 3) account description with UTF-8 characters in integer range 1000 to 1999 (R code applied to user profile description: any(utf8ToInt(x) %in% 1000:1999)). We also included accounts as English accounts if they were *not* labeled non-English by Linvill Warren.

we studied tweets posted after the end of the Republican presidential primary (starting our analysis in June 2016), but we also present longer time series based on our hand coded labels in Section D.3.

Table A2 shows the reductions in sample size after removing post-election and non-English accounts, as well as tweets without any content after text processing (see Section A.4.5) and a match to the Linvill Warren account labels. Most of the data is removed by the pre-election English language restrictions.

Twitter's official data set	N tweets	N accounts with tweets
All IRA Tweets	9,041,308	3,667
+ Pre-Election	7,053,777	3,235
+ Any Content after Text Processing	6,332,480	3,234
+ English	2,657,397	1,905
+ Linvill Warren accounts	2,282,142	1,135

Table A2: *Sample Sizes*. This table shows the numbers of tweets and accounts remaining after subsetting the Twitter data set to pre-election tweets, English language tweets and accounts, processing with defaults in R package "stm", and overlap with the Linvill Warren hand labels. The largest reduction in sample size came from the removal of non-English tweets. Note that a small number of these accounts were no longer in Twitter's official data set as of March 2020 (3,613 accounts).

A.4.3 Data Subsets and Training Sets

Because prior work has established that Russian trolls promoted Republicans over Democrats, and so might have had different messaging goals for their artificial Republican-leaning versus Democratic-leaning communities, we analyze two sets of tweets: 1) all pre-election tweets, and 2) pre-election tweets within liberal and conservative clusters.

In text scaling model *training*, we further hold out a) "news aggregators" because they posted large volumes of spam-like and repetitive content,[16] and b) for models that we inter-

[16] Tweets from the "news aggregators" are then scored using models trained on the less repetitive data from other accounts.

pret directly (rather than through hand labels), content posted prior to 2016, since we were primarily concerned with messaging around the 2016 election. Table A3 shows the training sets for each analysis in this paper.

Section D.3 contains a full timeline of tweet categories based on our hand coded categories and text scaling from 2015 through 2016 (content prior to 2015 was very sparse, as shown in Figure A10). Other work has documented recruitment strategies used by trolls and imitation of local news outlets, as well as their campaigns before and after 2016 presidential election (Tucker et al. 2018). However, see Section D.3 for the analysis of tweets posted in 2015, many of which tweets appeared to concern Ukraine.

Analyses Using Text Scaling Output as Outcome	
Scores applied to:	Scores trained on:
All tweets pre-election in 2016	All tweets pre-election in 2016, excluding news aggregators
Left troll tweets pre-election in 2016	Left troll tweets pre-election in 2016
(in SI) Right troll tweets pre-election in 2016	Right troll tweets pre-election in 2016
Analyses Using Hand Labels as Outcome	
Scores/embeddings applied to:	Scores/embeddings trained on:
All tweets pre-election	All tweets pre-election, excluding news aggregators

Table A3: *Text Scaling Training Sets.* When using text scaling as the outcome (i.e. when we interpret the dimensions themselves in 2016), we train the text scaling in 2016 and remove news aggregators (these accounts posted repetitive and spam-like content). When we do not need to interpret dimensions directly– and instead use the text scaling as a word embedding method to assist in the supervised model labeling – we only remove news aggregators from training.

The liberal and conservative clusters were classified by prior work (Linvill and Warren 2020) and we validated those labels using community detection on troll user mentions (see appendix). The Mueller Report suggests that Russian operators created politically neutral accounts to gain credibility, and cooperated with each other in teams to amplify messages and appear authentic. Prior studies have also documented high levels of clustering among the IRA accounts (Dawson and Innes 2019; Stewart, Arif and Starbird 2018; Howard et al. 2018).

A.4.4 Unit of Analysis

We do not know the number of operators behind the accounts, and the IRA accounts likely functioned within a coordinated unit. As we show in the community detection section below, the accounts are perhaps best considered as clusters of accounts rather than independent accounts, given that they were highly interconnected (and likely coordinated, especially within cluster).

We nonetheless consider within troll account changes in Section D.2 where we center each account at its dimension or category mean. These analyses show that the shifts from politics to entertainment on the left occurred within accounts.

A.4.5 Text Processing

Analyzing the tweet text requires some pre-processing of the text when converting text into a document-term matrix.[17] For this, we used the default text processing settings in the R

[17] This matrix contained one row for each tweet, and one column for each word. An entry for a word was 1 if present in a given document and 0 otherwise.

package "stm,"[18][19] but did not 'stem' words so that tables were easier to read and because much of the platform-specific language in tweets cannot be easily stemmed. In keeping with those defaults, we also did not remove hashtags (which improve searchability on Twitter and are often used to link content to an ongoing conversation on the site), user mentions (i.e. the accounts promoted in the tweet), or web links.

[18] https://cran.r-project.org/web/packages/stm/index.html

[19] Default used: convert to lowercase, remove stopwords, remove numbers, remove punctuation, words 3 or more letters only.

B Text Scaling

We use text scaling to identify the kinds of language that would fit the description of "flooding" previously used by authoritarian regimes. In China, for example, users partly posted positive comments about Chinese history (King, Pan and Roberts 2017). We do not expect Russian trolls to discuss Chinese history to flood American social media, and so we need some way to determine what they might have promoted instead. Once we analyze the text using scaling, we then validate those analyses using hand coded categories. In this, we chose relatively ordinary-sounding categories for coding, and we did not ask human coders to evaluate whether a tweet was distracting. Prior work argues that crowd-sourced tasks must be provided in clear and simple terms, even if broader goals are more abstract (Benoit et al. 2016).

We do not use topic models here because automated selection of the number of topics in these models typically leads to a very large number of topics (e.g. 100). Each of these topics might explain only a small amount of text in a given data set, and we might need to combine a large number of topics to estimate a broad "distraction" category. Although it is possible for researchers to specify very small numbers of topics in model fitting, topic models are not commonly used to estimate only a handful of categories.

As a robustness check, we train a GloVe word embedding model (Pennington, Socher and Manning 2014) on the same IRA Twitter corpus and show that these word embeddings lead to somewhat poorly calibrated supervised models for our hand labels.[20] Despite their relatively poor calibration (they accurately predict *individual* labels but under-fit the hand labeled *proportions* over time), however, the word embedding based results are not substantively

[20] As noted below, calibration, and the accurate estimation of proportions as opposed to the correct categorization of individual documents, is widely considered the most important metric when evaluating supervised models based on hand labels in social science (Hopkins and King 2010; Card and North 2018), even though measures of precision and recall matter for model efficiency.

different from the text scaling based ones (see Section C.2.6).

Ultimately, each of our text analyses, including the supervised models of hand categories, create a dictionary in which each word is assigned a score (e.g. the probability a word is "political") and each document is the average of its words' scores.

B.1 Text Scaling: Explanation of Method

We use a form of principal component analysis, called pivoted text scaling (Hobbs 2019), for text scaling. The method applies singular value decomposition to a standardized and truncated word co-occurrence matrix, and, as in PCA, its right singular values are then used to score words and documents. This approach is closely related to latent semantic analysis (Deerwester et al. 1990) and its many derivatives commonly used in automated text analysis today.

Document-term matrix

	a	b	c
Doc 1	1	1	0
Doc 2	1	0	1

Word co-occurrence matrix

	a	b	c
a	2	1	1
b	1	1	0
c	1	0	1

Standardized
word co-occurrence matrix

	a	b	c
a	0.7	0.5	0.5
b	0.7	0.7	0
c	0.7	0	0.7

Truncated and standardized
word co-occurrence matrix

	a	b	c
a	0.7	0.5	0.5
b	0	0	0
c	0	0	0

Table A4: *Example Matrix Transformation in Pivoted Text Scaling.* The text scaling used in this paper estimates word vectors using singular value decomposition on a truncated and standardized word co-occurrence matrix (example in bottom-right). Much like principal components, these word scores are the standardized word co-occurrence matrix (representing all words without truncation, bottom-left) multiplied by the right singular vectors of the truncated and standardized co-occurrences (using only the common words' rows in estimation). Document scores are averages of their words' scores.

In this, the co-occurrence matrix is the cross product of the document-term matrix. In the document-term matrix, documents are rows and columns are words (with entries the number of a given word in a document). The co-occurrence matrix is symmetric and has words as both rows and columns. As noted above, we do not remove any text from the tweets in this analysis, other than the "stopwords" (e.g. "it", "the") removed by default in the widely text analysis package "stm". Hashtags are words.

Analyzing the word co-occurrences rather than the documents themselves, as we do here, is a standard approach when analyzing short text, such as those found on social media. As examples of this approach for topic models, see the biterm topic model (Yan et al. 2013) and the word network topic model (Zuo, Zhao and Xu 2015).

Standardization controls for word frequency. In this method, the matrix is standardized by taking the square root of each count and then dividing each row by its Euclidean norm. Standardization is almost always used in one way or another in text scaling models. Without this standardization, PCA on a word co-occurrence matrix tends to produce a 1st dimension for the most common word, a 2nd dimension for the 2nd most common word and so on. With the standardization, there is a single dimension (called dimension 0) for word frequency and document length, while subsequent dimensions estimate word polarization.

Truncation, in turn, helps ensure that the top dimensions of the output still capture variation in commonly used language. Following Hobbs (2019), this specifically analyzes the co-occurrences for words that appear more often than their accompanying words. The accompanying words' co-occurrences contain noisy (i.e. they still contain few words even after aggregation) and duplicated information, since they are rare and already analyzed when they appear in the co-occurrences of the more common words.

Truncating such a matrix is closely related to a technique called sparse principal component analysis (Zou, Hastie and Tibshirani 2006). Sparse PCA estimates a small number of loadings that explain a large amount of variance in a matrix. In text, the sparse loadings maximizing explained variance are very often the most common words (Zhang and Ghaoui 2011). Pivoted text scaling makes this sparsification explicit by truncating the word co-occurrence matrix at the point where words on average appear less often than their accompanying words. This sparsification is applied to the rows of the word co-occurrences, and so its right singular values (closely related to PCA loadings, which we will use interchangeably in this context[21]) still estimate the locations of *all* words so that we can score documents that do not use common words.

The top output dimensions of PCA on this matrix are then vectors that explain the greatest variance in the standardized and truncated word co-occurrence matrix. As with principal components, each word is assigned a vector of numbers based on the right singular vectors applied to the standardized word co-occurrence matrix. Lack of truncation in this step scores all words using the principal components of common words.

These vectors represent words' locations on latent dimensions (semantic vectors).[22] Similar to document scoring using 'word embeddings' (Mikolov et al. 2013; Pennington, Socher and Manning 2014), documents are then scored using the average of their words' scores. Also like standard word embeddings, we use this PCA output (the top 10 dimensions) as input to later supervised models. As a robustness check, we compare those models to ones trained on GloVe word embeddings (Pennington, Socher and Manning 2014).

[21] In downstream analyses, document scores are scaled to standard deviations and their scales are not interpreted as the amount of variance explained in the document-term matrix. We use hand labels to both validate our categories and create more human-interpretable scales.

[22] The principal components for the words' "documents" (left singular vectors) are the same as the loadings for common words, but zero for rare words. These vectors are used only for identifying keywords.

B.2 Text Scaling: Output and Interpretation

As a reminder, we use text scaling to identify the kinds of language that would fit the description of "flooding" previously used by authoritarian regimes. With the top dimensions of the PCA output, we then identify two theoretically relevant latent variables to analyze and validate with crowd-sourced hand coding:

1. A partisan dimension, which for example separates the Linvill and Warren conservative accounts from liberal accounts, and
2. A social de-mobilization dimension, in which trolls post American entertainment content, such as tweets about popular music.

These latent dimensions can be constructed using addition and subtraction of the top two principal components of the overall analysis (all pre-election, English tweets in 2016, excluding news troll spam) and the left troll analysis respectively. Although scaling in political science is often used to identify a top partisan dimension, top dimensions of unsupervised scaling output do not necessarily capture variables of interest.

Here, these variables of interest were the top dimensions of the output. The partisan dimension shown in the main text is the 2nd dimension in Table A5 and the social de-mobilization dimension shown in the main text is 1st dimension plus the 2nd dimension in Table A6. As shown in Figures A9, we observe the same over-time patterns (and similar keywords, see Table A6) in both the 1st dimension and 2nd dimension of the liberal cluster text.

In the tables below, we show the keywords for each of those top two dimensions.[23] In Section C, we validate our labels for the dimensions using the crowd-sourced coding of tweets.

[23]Keywords are estimated using left singular vectors of the transformed word co-occurrence matrix described in the previous section. See Hobbs (2019) for details.

B.2.1 Text Scaling: Overall 2016 - keywords

Dimension 1		Dimension 2	
		Conservative	**Liberal**
giselleevns	gerfingerpoken	trumpforpresident	unarmed
ihatepokemongobecause	thinker	makeamericagreatagain	fatally
danageezus	clinton	perfectsliders	police
hashtag	tcot	imvotingbecause	officer
worldofhashtags	httpstcojeaacre	trumpk	benandjerrysnewflavor
midnight	ccot	trumppence	bleepthepolice
eat	joeamerica	trumptrain	cop
thingseveryboywantstohear	maga	hillaryforprison	policebrutality
chrixmorgan	lnyhbt	votetrump	blacklivesmatter
pokemon	petefrt	gopdebatesc	acab
playing	tlot	draintheswamp	shot
ruinadinnerinonephrase	trumptrain	giselleevns	pauloneal
ihateitwhen	pjnet	johnatsrs	btp
onewordoffmoviequotes	poll	maga	shooting
boothprince	rasmussen	lockherup	trueblacknews

Table A5: *2016 Overall Keywords*

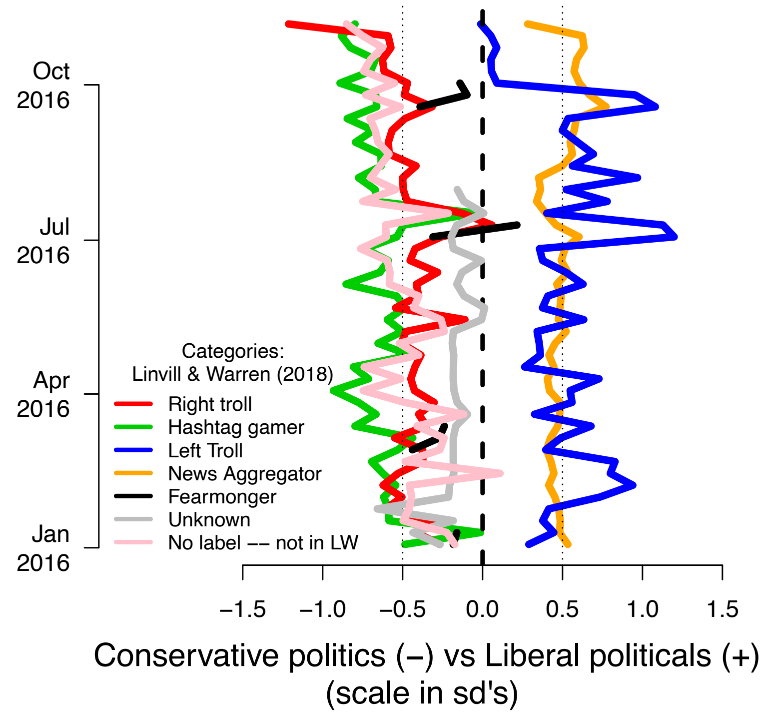

Figure A1: *Polarization Over Time – all categories.* Note that the categories "fearmonger", "unknown", and not labeled at all by Linvill Warren rarely posted in 2016. See Figure A10 for a visualization of their activity counts.

B.2.2 Text Scaling: Left trolls 2016 - keywords and over-time plots

Dimension 1		Dimension 2	
Mobilization	**De-Mobilization**	**Mobilization**	**De-Mobilization**
fatally	indieradioplay	blackskinisnotacrime	rapstationradio
unarmed	httpstcoemxjgtvv	chaimgoldberg	feat
shooting	tycashh	blackoncampus	torae
officer	sinice	red-pilled	hiphop
benandjerrysnewflavor	thetrudz	nowadays	barz
charges	playing	diminish	-fr-o
charged	music	antipolicebrutalityday	scarface
police	nowplaying	istartcryingwhen	nowplaying
pauloneal	listen	beingblackis	checkitout
cop	nineoh	fggot	october
shot	rapstationradio	altonsterling	contest
fixthepolice	boogsmalone	philandocastile	reks
dashcam	recklessdondon	wearhoodiefortrayvon	kass
bulldoze	rdeyeplug	oscarhasnocolor	xzibit
fatal	ogiiiiy	blackpowerbaby	mixtapemppromo

Table A6: *Left Troll Dimension 1 and 2 Over Time.*

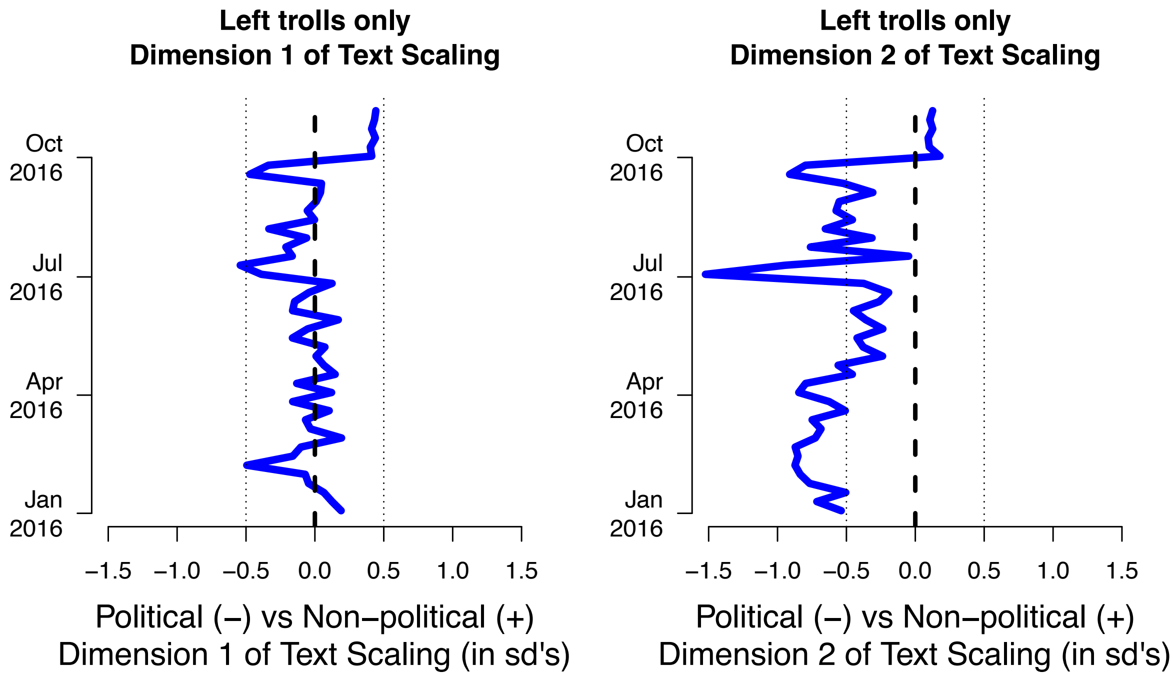

Figure A2: *Left Trolls, Over Time*

B.2.3 Text Scaling: Right trolls 2016 - keywords and over-time plot

Dimension 1		Dimension 2	
renewus	adjusted	afight	islamkills
islamkills	rasmussen	trumpisright	brussels
afight	incite	crookedcruz	oscarhasnocolor
stopislam	bribe	readily	prayforbrussels
cosproject	lester	trumpwillwin	oscars
jstines	mcclatchy	rkba	stopislam
brussels	statespoll	ctot	refugees
pjnet	ppollingnumbers	lnyhbt	honorforthebrave
irishjoeharriso	holt	perfectlylaura	oscarssowhite
readily	emails	trumparmy	europe
cruzcrew	manager	tgdn	oscar
molonlabe	aide	irishjoeharriso	nocybercensorship
ccot	overcharging	ppsellsbabyparts	terrorists
nra	probe	noliberalbias	texit
makedclisten	allegations	defundpp	religionofpeace

Table A7: *Right Trolls 2016: Keywords*

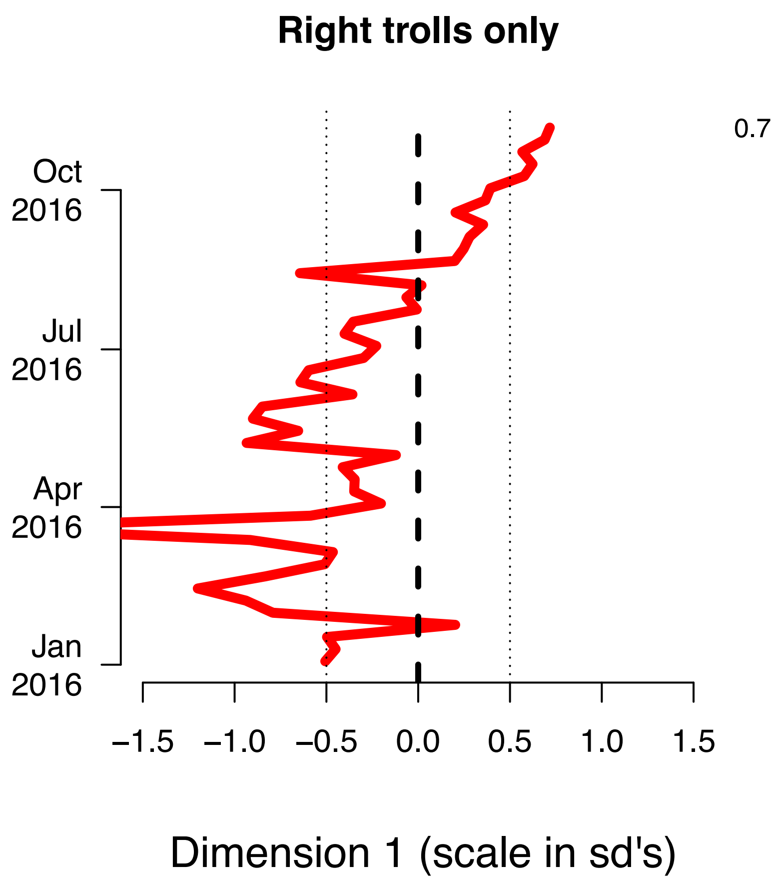

Figure A3: *Maga Imitators Over Time – dimension 1 in Table A7.*

C Hand Coding and Supervised Labeling

For our hand coded analysis, we need to validate that out-of-the-loop human readers identify the same "entertainment" vs "politics" distinction when actually reading the tweet texts. We also seek to place our text scaling estimates onto a more interpretable scale – the proportion of documents about politics or entertainment.

C.1 Hand Coding: Tweet Sampling and Coding Instructions

To validate the text scaling measure of apolitical tweets, we designed a coding exercise using the research platform Figure Eight. This platform uses Amazon Mechanical Turk, which is an online crowdsourcing service where anonymous workers complete tasks online for small sums of money.

In our task, human coders were first given a set of detailed instructions (see Figure A4 below), and then were given selection of individual tweets. We asked human coders to read each tweet, and assign each tweet to one of four distinct categories: i) Politics and Elections, ii) Social Justice, and Race Relations, iii) Entertainment, and iv) Unclear/Other. Coders were given descriptions of each category as well as example tweets in the instructions, and were instructed that if a tweet falls into two or more categories, to just choose one. If a coder selected the "Other" category, they also had the ability to explain their rationale using an open-ended response text box.

We added the social justice/race relations category both because it is a dominant factor in our text scaling and because these discussions would not necessarily be coded (partisan) *politics* – yet, would nonetheless be relevant to political mobilization on the left and right. Prior studies have established that trolls talked about Black Lives Matter and other social justice topics on the left (Arif, Stewart and Starbird 2018).

We randomly sampled 450 tweets from the left trolls and 450 tweets from the right trolls for crowd-sourced human coding. Given our interest in late campaign shifts, we sampled those tweets from June 1, 2016 through November 8, 2016. Each tweet was categorized by three independent individuals, who were based in the United States and were ranked as high quality workers by Figure Eight. We assigned a tweet to a topic if two out of three coders chose that topic. The coding task took place on November 23, 2019.

Although we were specifically interested in general campaign messaging in this paper, we nonetheless label all tweets 2015 through 2016 using these labels (see Section C.2).

Coders received the following instructions:

Classify Tweets [T2]

Instructions ▴

Overview

In this job, you will be presented with various tweets, taken from Twitter. We are asking you to sort those tweets into categories

Steps

- Read the text of the tweet.
- Determine which category the tweet best fits into.
  - **PLEASE PAY ATTENTION TO HOW WE DEFINE THE CATEGORIES BELOW.**

Categories

POLITICS/ELECTIONS

These tweets will reference politics, partisanship, and/or elections -- particularly US politics, US presidential elections, and political candidates (e.g. Donald Trump or Hillary Clinton), as well as partisan media focused on these topics.

Some examples:

- RT @foxandfriends: .@JudgeJeanine: Does Clinton ever answer a question directly? https://t.co/lld75eyQEe https://t.co/E4b0O2wO37
- RT @annieb0823: Roger Stone puts this campaign in perspective.. worth watching! https://t.co/4TkzdnkifH https://t.co/0gb6GblH8R
- Trump is so poor that by next year he'll be living under public housing. https://t.co/WCu9bqyGo2
- RT @Dbargen: Arms dealer says admin made him scapegoat on Libya operation to 'protect' Clinton @@FoxNews #TCOT #MAGA #LNYHBT https://t.co/â€¦

SOCIAL JUSTICE/RACE RELATIONS

These tweets will reference issues relating to race relations, social justice, activism, and protest -- particularly in a US context, and including the Black Lives Matter movement.

Some examples:

- A police-involved shooting in Mount Greenwood, Chicago leaves a Black man dead. https://t.co/HicNIGCRHj
- RT @USVeteran2: Two Hospitalized White Girls Prove Black Lives Matter Is A Racist Movement https://t.co/Gz3MsDd6UR
- FUTURE UNDERSTANDS #STAYWOKE https://t.co/bUhuSEouBT
- RT @547George: Do normal people agree #blacklivesmatter? Its just a meme promoted by those who believe #kalergi plan of #whitegenocide mattâ€¦

ENTERTAINMENT

These tweets will often reference music, celebrities, popular shows, movies, sports and things of this nature. They may also be content that you might expect to see in a tabloid.

Some examples:

- RT @CINESHARES: .@ShadowAndAct: Terry Crews Will Save Christmas in Week-Long Holiday-Themed Series for The CW https://t.co/msQsFMerwI
- RT @ZerlinaMaxwell: âš¡ï¸ â€œThe Empire cast is with herâ€ https://t.co/Q2xey6gLR0
- RT @PROMO4LIFEBIZ: This my new favorite song! @WhoTFisVon https://t.co/xb4eliBOUp
- RT @PettyIdol: At the #AmySchumerGottaGoParty likeâ€¦ https://t.co/Lwj8DorbPr

*** Some tweets might fall into two or more of the categories above, please choose the ONE that is best fits into.***

FINALLY, THERE WILL BE TWEETS THAT ARE (1) HARD TO DECIPHER, (2) DO NOT FALL INTO ANY OF THE THREE CATEGORIES ABOVE. THOSE BELONG IN THE FINAL CATEGORY.

UNCLEAR/OTHER

These tweets will not fall into any of the three categories above.

Some examples:

- RT @FeministaJones: Yes I can. Humans are trash. Don't need a religious text to explain this. https://t.co/uvS3fEBf5w
- RT @JamilahLemieux: Stop to smell the flowers today! #minimilah https://t.co/SAckVplmYs
- RT @redrivergrl: Walking it back, walking it back... https://t.co/Vhvgkp46Km
- RT @Delo_Taylor: Maybe he's a Vampire? ðŸ‘¿ðŸ‘¿ðŸ‘¿ https://t.co/xDo21Vg1rM

Figure A4: FigureEight Coding Exercise: Instructions Given to Workers

C.2 Hand Coding: Human Coder and Supervised Model Evaluations

We first evaluate inter-coder reliability among all human coders and then evaluate "inter-coder" reliability between 2 out of 3 human coders and our (test set) machine predictions. We consider the human coder - machine inter-reliability anticipating that some fraction of the human coders answered randomly, and that using 2 out 3 coders will be more reliable. A machine should be able to pick up on systematic, non-random patterns in training data to predict the categorization of the 2 out of 3 coders in test data.

Inter-coder reliability evaluates how precise our hand labeled data and machine predictions are. The *calibration* of the machine predictions is more directly relevant to social science than accuracy, however, since we are typically interested in aggregate proportions rather than the classification of individual documents (Hopkins and King 2010; Card and North 2018). In this, for example, approximately 60% of tweets assigned a predicted probability of 60% for being about politics should actually be labeled "politics".

For the calibration evaluations, we present two forms of evidence:

1. we display our results using both machine prediction and hand coded averages (showing that they do not substantively differ), and
2. we display calibration plots.

C.2.1 Hand Coding: Supervised Models

For machine predictions, we use Lassos (*l1* penalized logistic regressions) (Tibshirani 1996) and the first 10 dimensions of our PCA-based word embeddings (see Section B.1) to predict *each* of the categories. Logistic regression is well-calibrated compared to more complex models (Niculescu-Mizil and Caruana 2005; Card and North 2018), and the Lasso in particular has few researcher selected tuning parameters, especially when compared to neural nets and random forests. The sole penalization term in the Lasso is selected automatically using cross-validation in standard software packages (we use the R package "glmnet"[24]). As a robustness check, we also show results predicting hand labels using GloVe word embeddings trained on the troll data.

The dependent variable in each of the models is an indicator for whether 2 or more human coders labeled a tweet a given category (e.g. for entertainment, whether 2 or more coders labeled the tweet "entertainment"). In analyses using these predictions, we use predicted probabilities from the models.

C.2.2 Hand Coding: Interrater Reliability and Prediction Accuracy

In Table A8, we show inter-rater reliability (Fleiss' Kappa) for the 3 labels on each tweet. These calculations use the kappa.fleiss command in the R package "irr"[25] and the confusion-Matrix command in the R package "caret"[26].

We anticipated some fraction of the Amazon Turk workers' submissions to be random *or* for the tweet itself to be uninterpretable, and so had 3 workers code each tweet. In Table A9, we evaluate hand labels for tweets where 2 or more of the coders agreed on a label.

In Table A9, we show inter-rater reliability for the 2 or more coder agreement labels compared to dichotomized machine predictions in a holdout set. For this procedure, we randomly subset our data into approximately 50/50 splits, trained a Lasso on one half of the hand labels, and then evaluated those machine predictions on the remainder of the hand labels. We repeated that procedure 1000 times and report the average of those Kappas, as well as intraclass correlation for continuous predictions and the fraction of hand label for a

[24] `https://cran.r-project.org/web/packages/glmnet/index.html` and it by default selects the penalization term using minimum misclassification error in cross-validation

[25] https://cran.r-project.org/package=irr

[26] https://cran.r-project.org/package=caret

	Left Troll	Right Troll
Entertainment	0.37	0.29
Politics	0.47	0.55
Social Justice and Race Relations	0.42	0.32
Other	0.14	0.23
Overall human inter-rater reliability 0.42 (Fleiss' Kappa) 0.42 (Krippendorff's Alpha)		
In our analyses, we use the hand labels where at least 2 coders agreed, and re-label remaining tweets "other/no agreement". We evaluate those labels below.		

Table A8: *Inter-rater reliability (Kappa, 3 human raters)*

given category.

The continuous intraclass correlations reflect greater accuracy for labels with 100% agreement and somewhat lower accuracy for mixed labels, which are changed to 100% agreement if 2 out of 3 coders agree for the dichotomized evaluation.

All of the categories of interest have moderately high inter-rater reliability.[27] We are only unable to predict the "other/no coder agreement" tweets, suggesting that the texts not considered in our analyses lack systematic patterns to distinguish them from other tweets.

These human-machine Kappas are included here as a comparison for the hand label inter-rater reliability shown in Table A8. In Sections C.2.4 and C.2.5, we show more standard machine learning evaluations for calibration (calibration plots), sensitivity vs specificity (receiver operating characteristic curves), and, across the 1000 replicates shown above, area under the ROC curve.

[27]Note that a reliable "politics" category for the Right Trolls is sufficient to establish that explicitly political content was common relative to other types of content, including entertainment.

	Left Troll	Right Troll	Combined
Kappa: Dichotomized Labels/Predictions			
Entertainment	0.47	0.33	0.49
Politics	0.53	0.58	0.67
Social Justice and Race Relations	0.53	0.36	0.47
Other	0.01	0.01	0.01
Intra-Class Correlation: Fraction with Label/Continuous Predictions			
Entertainment	0.61	0.54	0.65
Politics	0.61	0.67	0.74
Social Justice and Race Relations	0.59	0.54	0.58
Other	0.05	0.12	0.09
Note: training in this evaluation is based on 50% of data to allow for training-test split.			

Table A9: *Inter-rater reliability (Kappa or ICC in test set, 2 out 3 human raters - machine predictions) – see Figure A11 for AUCs.* Test set human-machine reliability here suggests that the 2 (or more) out of 3 agreement among coders picks up on systematic variation in the text. Training sets are approximately 450 observations, while test sets are subset from the remaining observations down to the relevant category (i.e. there are fewer observations in test sets for the left and right troll evaluations). Note that our actual analyses use the entire labeled data set in training – 900 tweets: 208 entertainment, 310 politics, 221 social justice and the remaining designated "other/no coder agreement".

C.2.3 Hand Coding: Comparison of Analyses Based on Hand Labels and Machine Predictions

In each of the figures below, we show a) proportions of topics from the hand-coded tweets (with tweets categorized in a topic when 2 out of 3 coders agreed on that topic), and b) proportions of topics from a supervised model trained on the hand-coded tweets.

The Lasso on the PCA-based word embeddings (Section B.1) closely matched the hand-coded proportions but machine predictions did appear to *under*estimate the shift from politics to entertainment seen in the hand-coded data. As a reminder, the Lasso for each of these models used a logistic regression, and the regularization term was selected using minimum misclassification error in cross-validation (the default for binomial models in the "glmnet" R package).

The underestimation of the politics to entertainment shift and the perhaps smoother shift in content do not affect our interpretation of the overall results.

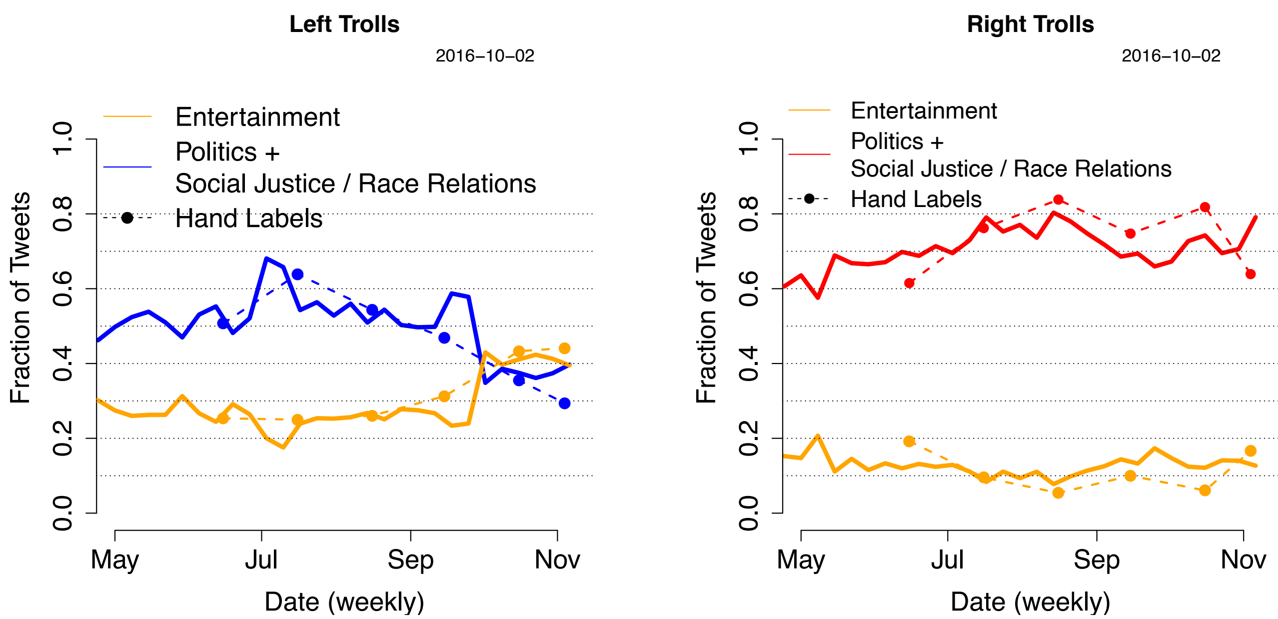

Figure A5: *Coding Validation Results from FigureEight.* (Note: repeated from main text.) This figure shows the results from applying a supervised model to label the full corpus based on a sample of hand-coded tweets. Solid lines are proportions from the supervised model, while dotted lines (and points) are from the raw hand-coded data. Our model only slightly underestimates the fraction of entertainment content in the left-leaning sample. There is some limited evidence of an increase in entertainment content prior to the spike in left-leaning activity.

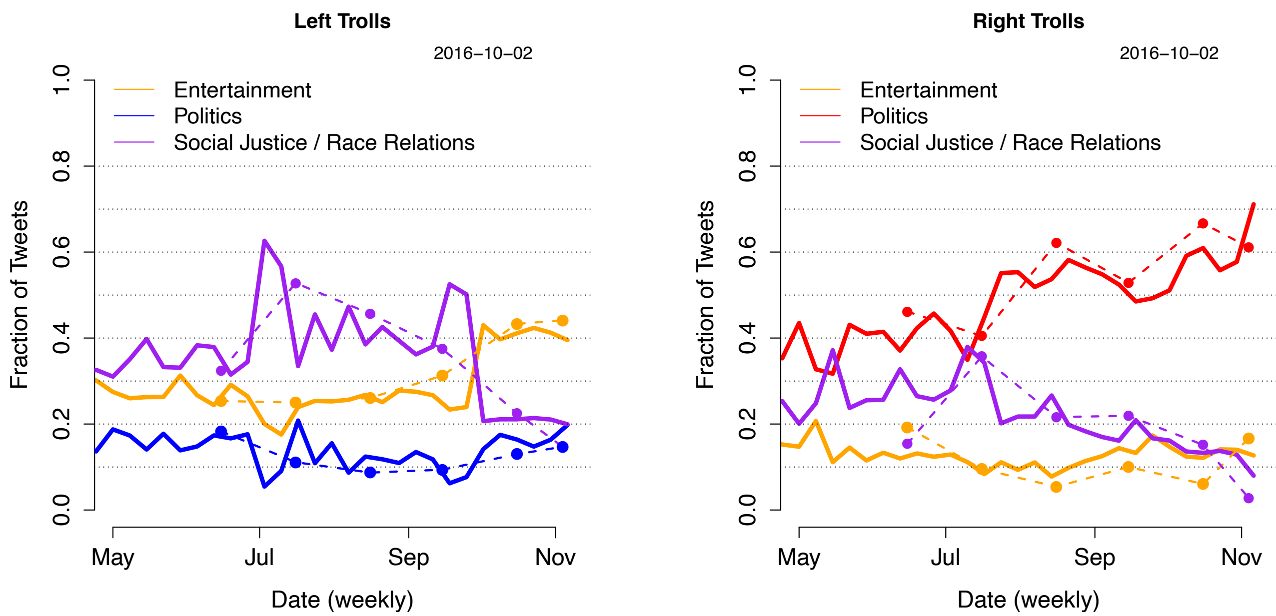

Figure A6: *Coding Validation Results from FigureEight.* This figure shows the results from applying a supervised model to label the full corpus based on a sample of hand-coded tweets. Solid lines are proportions from the supervised model, while dotted lines (and points) are from the raw hand-coded data. This figures separates the social justice category from the politics and election category. Our model slightly underestimates the fraction of entertainment content in the left-leaning sample.

C.2.4 Hand Coding: Calibration Plots

In Figure A7, we show calibration plots for each of our category predictions. For all of these predictions, our estimated probabilities approximate the actual proportion of tweets. In these plots, the x-axis is our predicted probability and the y-axis is the proportion of tweets with the hand label of interest. Predictions are placed into 5 bins in order to evaluate proportions based on binary labels.

These plots are best interpreted as assessments of model fit rather than plots representing the accuracy of the model, since these visualization are not based on training-test splits (as shown in Section C.2.2). The figure shows that we were able to fit assigned probabilities to the actual probabilities, despite binomial outcomes.

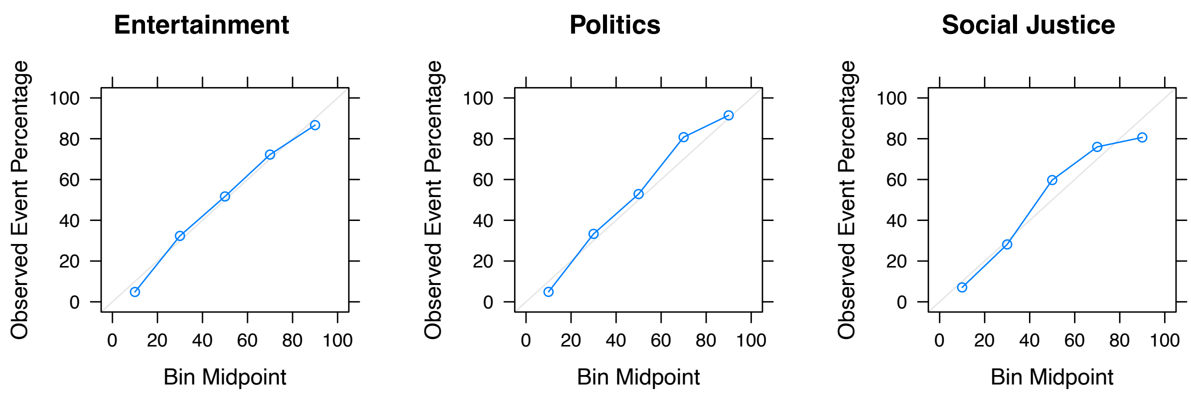

Figure A7: *Model calibration.*

C.2.5 Hand Coding: ROC Curves and Confusion Matrices

In Figure A8, we show ROC curves and confusion matrices for each of our category predictions. In the ROC curves, a line at 45 degrees indicates predictions no better than chance. The x axis is the false positive rate (e.g. machine labels "entertainment" while the human coder does not) and the y axis is the true positive rate (e.g. human coder labels "entertainment", machine also labels "entertainment"). ROC curves are helpful for evaluating binary predictions using data with unbalanced data. Other evaluation methods might, for example, score a model well for predicting that a rare event never occurs – such behavior would be readily apparent in the ROC curve, as well as the confusion matrix.

These plots are best interpreted as assessments of model fit rather than plots representing the accuracy of the model, since these visualization are not based on training-test splits (see Figure A11). The plots show balanced true positive and false positive rates – e.g. though not a major concern in our data, we nevertheless show that we are not achieving high accuracy through predicting all 1 or 0s for common / rare outcomes respectively.

Area under the ROC curve statistics for a 50/50 training-test split are shown in Table A11. This analysis dichotomizes both the hand labels (2 out of 3 or greater agreement) and predictions (greater than or equal to 50% probability).

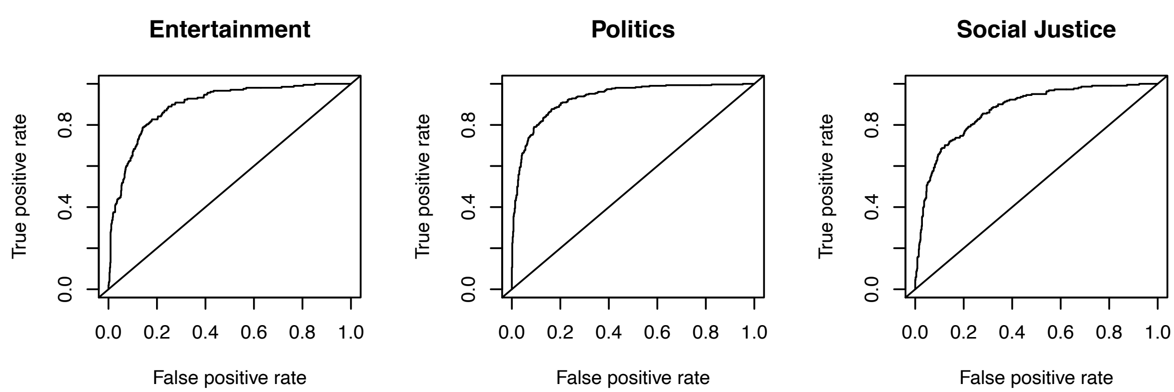

Figure A8: *Model sensitivity and specificity.*

In the confusion matrices, rather than use the predicted probabilities (as we use for the main analyses), we must dichotomize our predicted values. For this, we assigned the predictions 1 for probabilities greater than 50 percent and 0 otherwise. Keep in mind here that this analysis tells us how precise our predictions *above vs below* probability 50 percent. The top rows here are single points in the ROC curves – the top-right of the matrices x and the top-left of the matrices y. The ROC curves are more informative for true positive and false positive raters over many thresholds. Further, the calibration plots (and comparisons to hand labels over time) are more substantively important, since they inform whether our averages (the quantity of interest in our analyses) well-approximate the reference averages.

	Entertainment Reference (hand labels)			Politics Reference (hand labels)			Social Justice Reference (hand labels)		
		1	0		1	0		1	0
Prediction	1	0.55	0.07	1	0.74	0.07	1	0.48	0.05
(dichotomized)	0	0.45	0.93	0	0.26	0.93	0	0.52	0.95
	Sum	208	692	Sum	310	590	Sum	221	679

Table A10: *Confusion matrices.* This table shows confusion matrices for each of our hand labels and their corresponding predictions. To show proportions matching a single point on the ROC curves above, the reference columns are divided by the total number of hand labels assigned the category (or not).

	Left Troll	Right Troll	Combined
Entertainment	0.86	0.85	0.89
Politics	0.91	0.87	0.92
Social Justice and Race Relations	0.87	0.83	0.86
Other	0.59	0.71	0.66

Note: training in this evaluation is based on 50% of data to allow for training-test split.

Table A11: *Area Under the Receiver Operating Characteristic Curve (dichotomized labels)*. The data in this figure are drawn from the same 1000 replicates as shown in Figure A9.

C.2.6 Hand Coding: Predictions Using GloVe Word Embeddings

Predictions based on GloVe word embeddings (Pennington, Socher and Manning 2014) tended to underestimate over-time changes in the tweet contents compared to the hand labels, and predicted evenly distributed labels. We nonetheless see similar patterns in the predictions.

In these figures, the dotted lines are the hand labels, and the solid lines are the supervised model fits to that data.

The word embeddings here were estimated using the R package "text2vec,"[28] with word vectors set to size 100, window size to 5, and alpha 0.5.

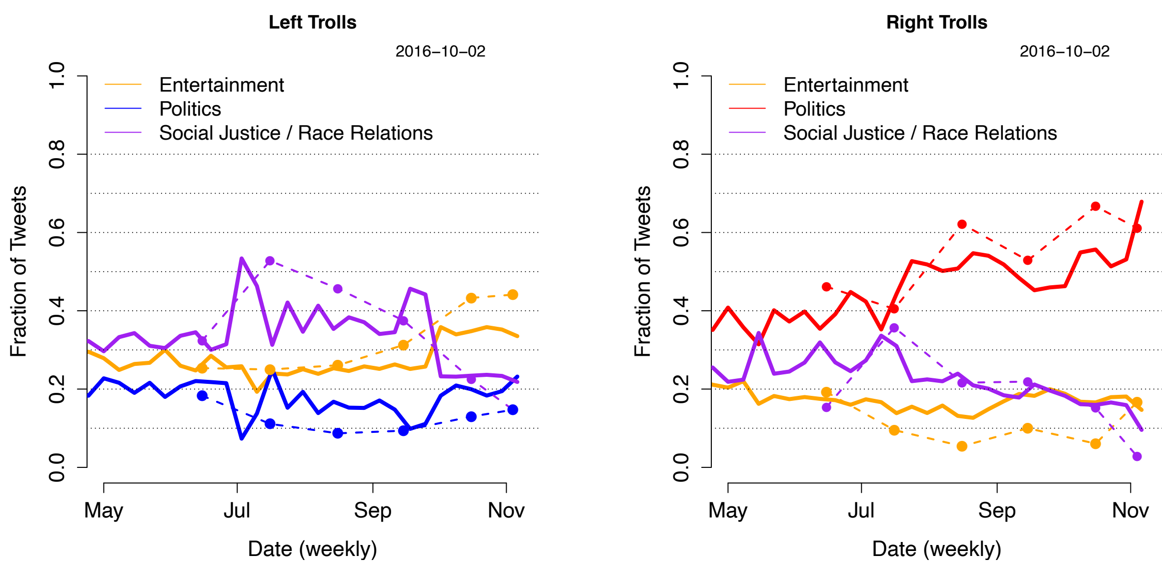

Figure A9:

[28] https://cran.r-project.org/package=text2vec

D Additional Results and Robustness Checks

D.1 Account Activity Timelines

Figure A10 plots the number of user mentions in tweets per account type from June 2014 through the election 2016. The spike in activity among polarized (i.e. left or right troll) accounts in 2015 occurred prior to the first Republican presidential debate, as shown in the bottom panel of Figure A11.

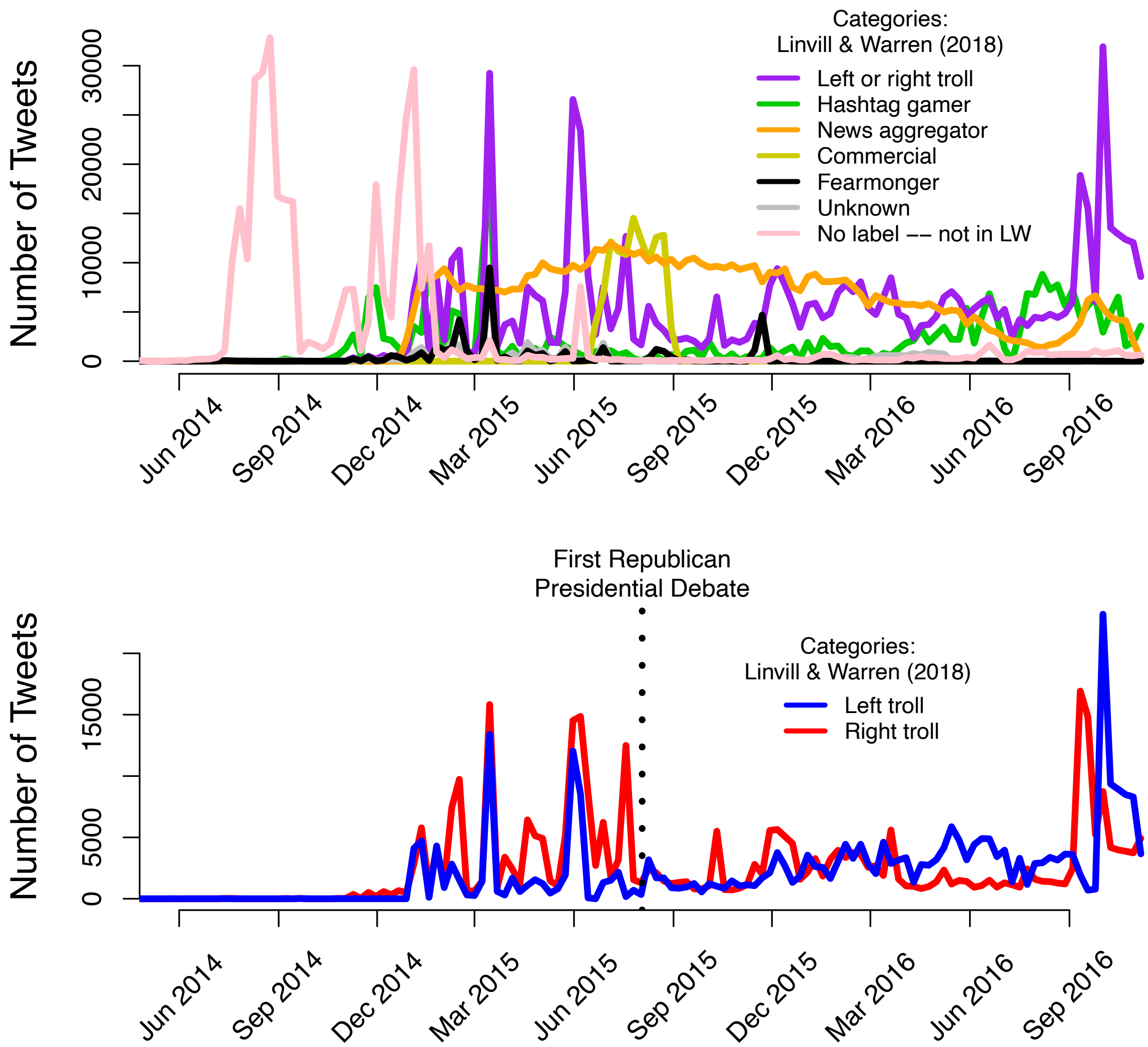

Figure A10: *Changes in use of different types of troll accounts.* Accounts tweeting local news reduced activity from 2015 into 2016, while accounts using polarized, partisan identities dramatically increased activity close to the 2016 election. Left and right trolls are presented together in the top panel of this figure and separately in the bottom panel. Note that less than 0.1% of the all tweets were posted prior to June 2014.

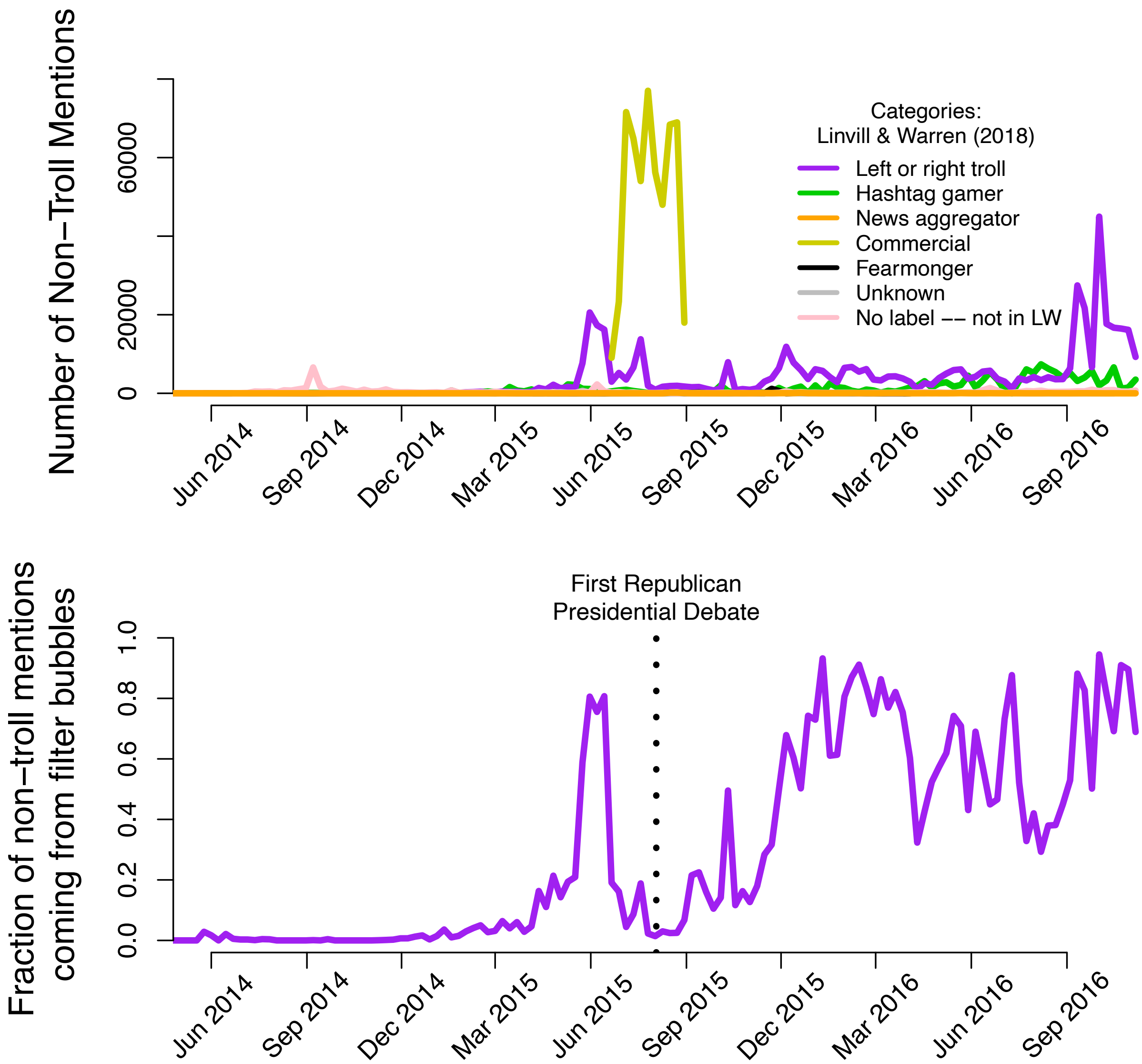

Figure A11: *Changes in use of different types of troll accounts – user mentions.*

D.2 Messaging Shifts Within Accounts

Figures A12 and A13 below repeat the main analyses with each troll account centered at its mean. These results mirror the findings in the main text – close to the election, left trolls shifted from discussing politics / social justice and race relations to entertainment. We also see a large shift from social justice / race relation to *politics* within accounts on the right as the 2016 election approaches.

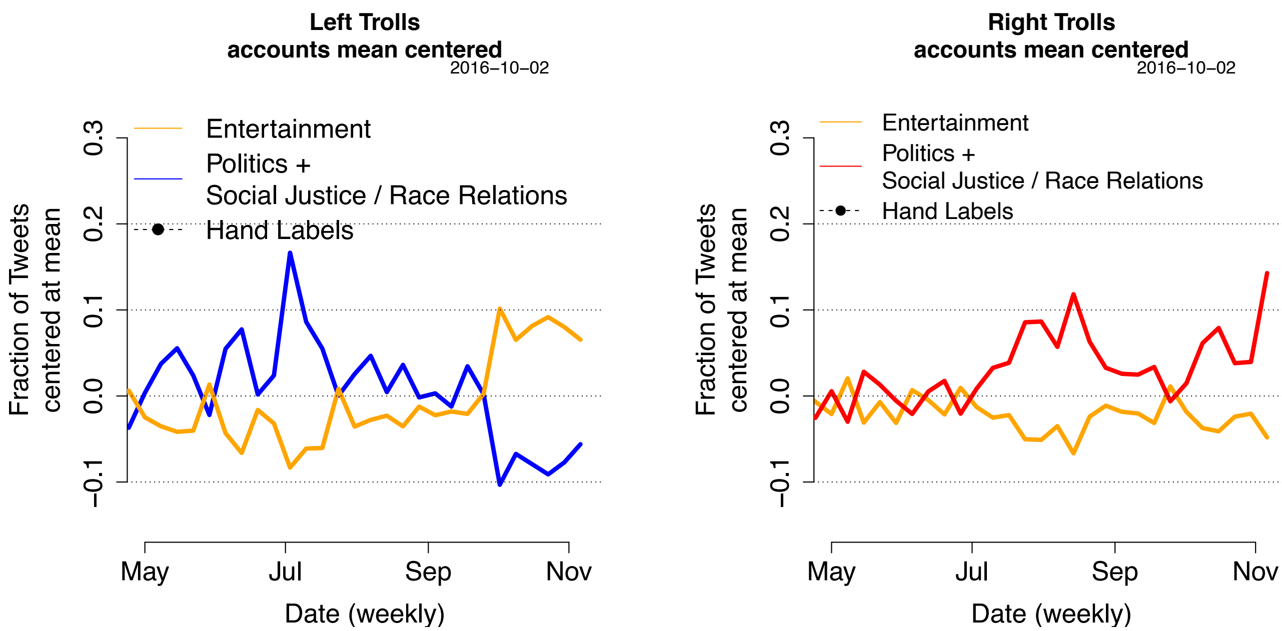

Figure A12: *Coding Validation Results from FigureEight – centered at mean by account.*

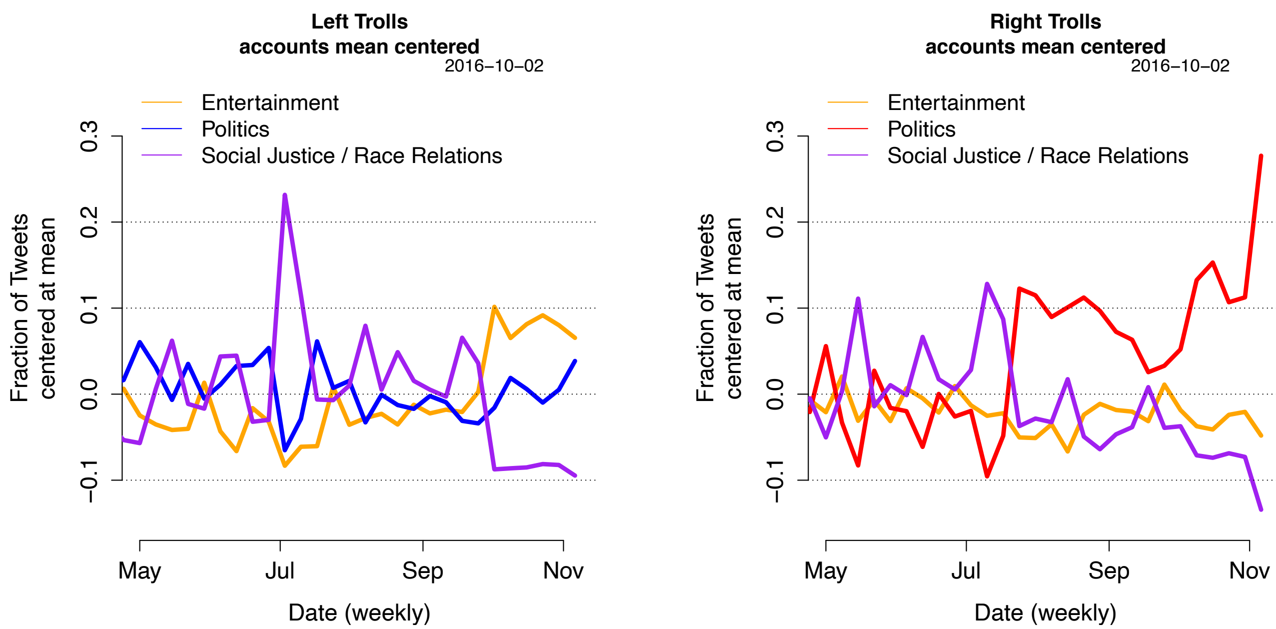

Figure A13: *Coding Validation Results from FigureEight – centered at mean by account.*

D.3 Results for 2015 through 2016

Figures A14 and A15 below repeat the main analyses for tweets going back to 2015.

The hand label predictions for 2015 should be interpreted with caution because we did not hand label any data from 2015. In particular, the increasingly political tweets from conservative accounts could reflect either a shift in topics not detectable with our 2016 hand labels, or a genuine politicization among those accounts. Either way, it is perhaps instructive to see that there was no shift among the liberal accounts, other than the very sparse and noisy activity in the first half of 2015 and earlier.

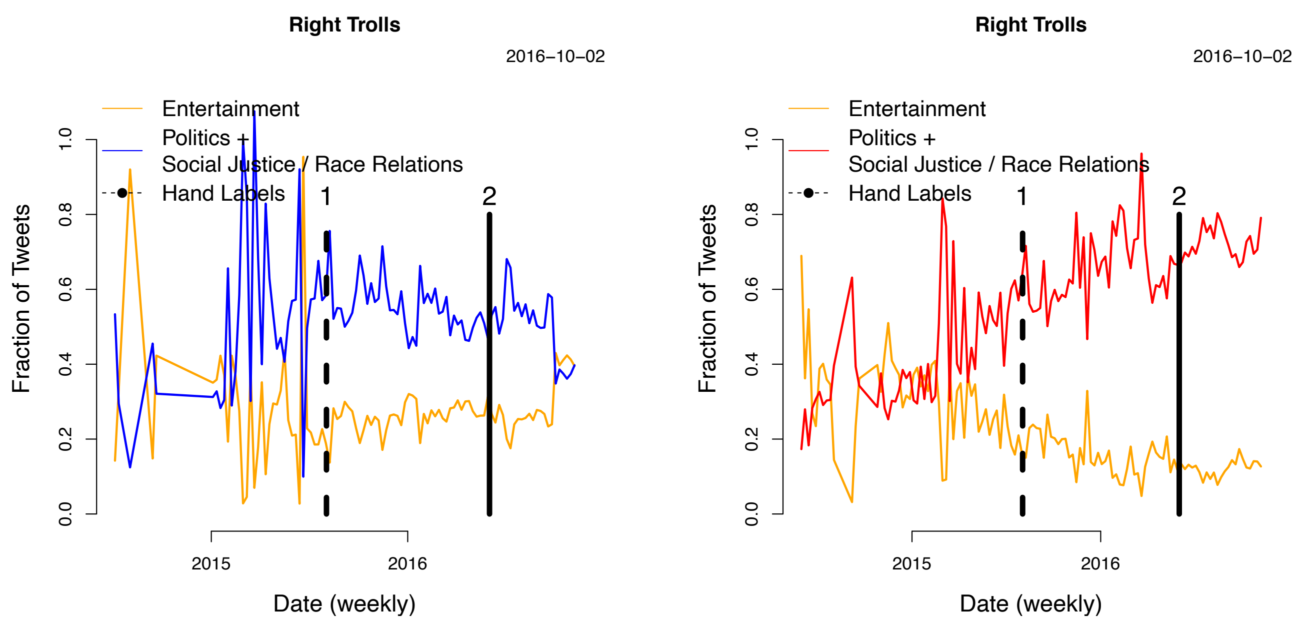

Figure A14: *Coding Validation Results from FigureEight – extended to 2014 through 2016.* Vertical lines in this figure are 1) the first Republican presidential debate (August 3, 2015) and 2) the earliest tweets hand-coded. We show the first Republican debate line in the activity figures as well – Figures A10 and A11.

The figure below shows that the within account shifts from entertainment to politics occurred within accounts from 2015 through 2016.

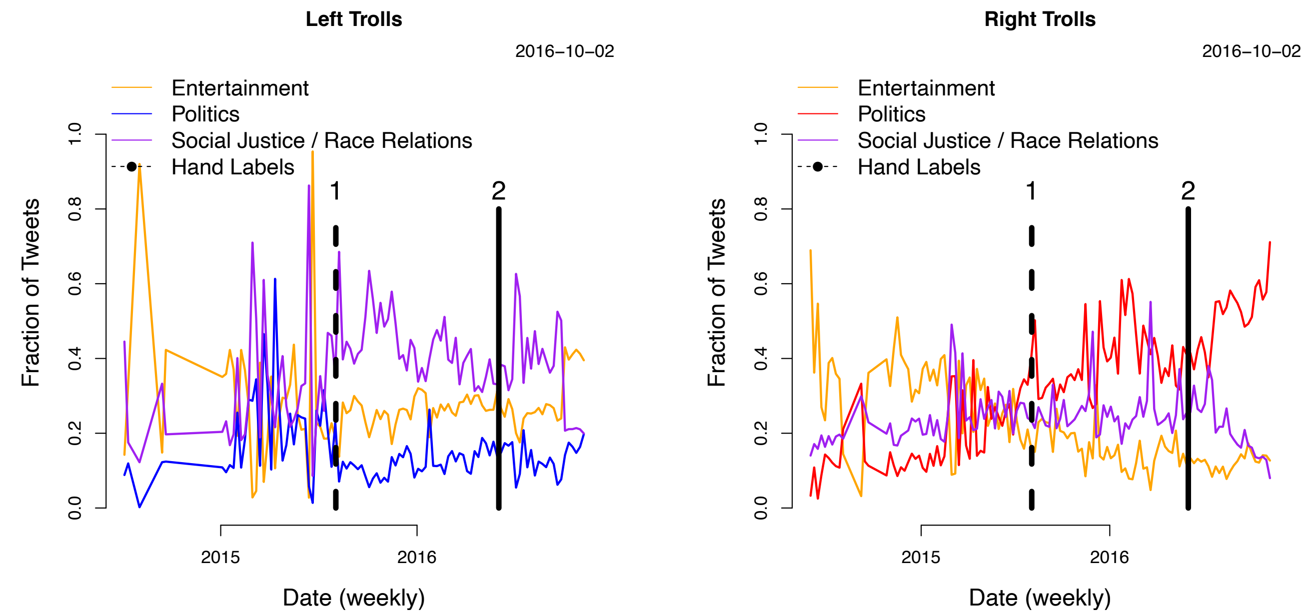

Figure A15: *Coding Validation Results from FigureEight – extended to 2014 through 2016.* Vertical lines in this figure are 1) the first Republican presidential debate (August 3, 2015) and 2) the earliest tweets hand-coded.

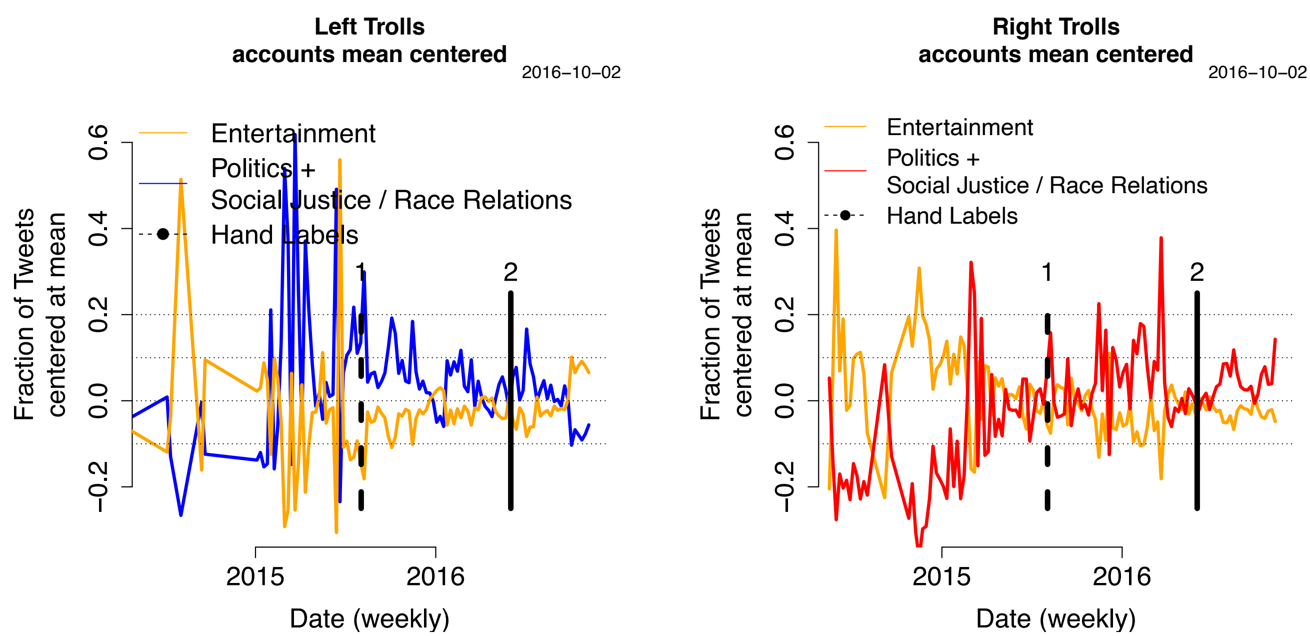

Figure A16: *Coding Validation Results from FigureEight – extended to 2014 through 2016, accounts centered at means.* Vertical lines in this figure are 1) the first Republican presidential debate (August 3, 2015) and 2) the earliest tweets hand-coded.

D.4 Comparisons of 2015 and 2016 Tweets Using Mutual Information

We focus our main analyses on tweets posted in 2016, but some of the troll accounts were active well prior to 2016. In this section, we use mutual information (Manning, Raghavan and Schütze 2008) to calculate what words distinguish the 2015 content from the 2016 content. Mutual information here measures which words provide the largest amount of information about whether their containing tweets were posted in 2015 or 2016.

This analysis illustrates differences in messaging from 2015 to 2016 that we likely do not address in our main analyses. In the 2015 words, we see discussion of the Fukushima Daiichi nuclear disaster and Ukraine-related news or propaganda.

In the following tables, we repeat this 2016 vs 2015 analysis by account category, and further calculate which words best distinguish a given account category from others.

All Trolls	
2016 words	2015 words
trump	fukushima
black	love
https	ukraine
hillary	httpt
clinton	chernobyl
blacklivesmatter	quote
httpst	true
gloedup	rap
islamkills	nuclear
giselleevns	npp
white	ukrainian
danageezus	life
pjnet	imho
httpstco	lentaruofficial
tcot	httptco

Table A12: *Distinctive words 2016 vs 2015, by mutual information.*

Left Trolls

Distinctively Left Troll Words (vs. other trolls in 2016)	2016 words (vs. 2015)	2015 words (vs.2016)
blacklivesmatter	black	news
black	https	baltimorepost
gloedup	trump	sports
police	blicqer	braveconwarrior
blacktwitter	gloedup	ebbdfcfdeaaedadadfefbeafdce
staywoke	talibkweli	local
cops	amp	bbsp
blicqer	httpst	independent
policebrutality	white	politics
white	httpstco	httpt
bleepthepolice	nowplaying	httptco
blm	blackhistorymonth	blackpeopletwitter
trayneshacole	btp	isis
talibkweli	beingblackis	chris
https	thehill	httptc

Table A13: *Left troll distinctive words 2016 vs 2015 and Left Troll vs others, by mutual information.*

Right Trolls

Distinctively Right Troll Words (vs. other trolls in 2016)	2016 words (vs. 2015)	2015 words (vs.2016)
hillary	trump	news
trump	hillary	braveconwarrior
tcot	islamkills	sports
pjnet	clinton	independent
obama	https	local
clinton	pjnet	chris
news	tcot	money
realdonaldtrump	ccot	love
ccot	amp	bbsp
wakeupamerica	brussels	life
refugees	httpst	httptco
isis	stopislam	make
maga	trumpforpresident	httpt
via	vote	selfie
hillaryclinton	hillaryforprison	politweecs

Table A14: *Right troll distinctive words 2016 vs 2015 and Right Troll vs others, by mutual information.*

Hashtag Gamers

Distinctively Hashtag Gamer Words (vs. other trolls in 2016)	2016 words (vs. 2015)	2015 words (vs.2016)
midnight	giselleevns	love
giselleevns	danageezus	true
danageezus	chrixmorgan	life
mustbebanned	boothprince	rap
igetdepressedwhen	trump	quote
ihatepokemongobecause	worldofhashtags	usa
chrixmorgan	amp	heart
rejecteddebatetopics	phonline	never
istartcryingwhen	bunniboila	quotes
toavoidworki	midnight	happiness
tofeelbetteri	kattfunny	sometimes
myolympicsportwouldbe	andyhashtagger	will
betteralternativetodebates	annogalactic	success
andyhashtagger	https	things
donttellanyonebut	gamiliell	mind

Table A15: *Hashtag gamer distinctive words 2016 vs 2015 and Hashtag Gamer vs others, by mutual information.*

News trolls

Distinctively News Troll Words (vs. other trolls in 2016)	2016 words (vs. 2015)	2015 words (vs.2016)
news	world	chicago
sports	trump	news
politics	zika	breaking
local	aleppo	local
business	warfareww	showbiz
foke	environment	newyork
chicago	sanders	foke
health	topnews	texas
topnews	syria	baseball
police	clinton	houston
world	tech	orleans
texas	cruz	atlanta
breaking	brexit	reuters
tech	rio	detroit
says	mosul	astros

Table A16: *Newsfeed Troll distinctive words 2016 vs 2015 and Newsfeed Trolls vs others, by mutual information.*

D.5 Network Community Detection

We use the account categories of Linvill and Warren (2020) in our main analyses, but these categories can also be identified using network community detection (Fortunato 2010). Communities are a natural feature of social networks, in that social networks have clusters with high connectivity within a group and low connectivity to others outside the group. For the trolls, the promotion of the same Twitter accounts and of each others' Twitter accounts would both increase their reach and, if distinct from the rest of activity on Twitter, likely increase the probability of discovery by Twitter itself (or, at least, raise the probability of a review of the clusters).

In Figure A17, the colors on the left correspond to clusters derived from a commonly applied community detection algorithm (Clauset, Newman and Moore 2004) and the colors on the right correspond to the account categories of Linvill and Warren (2020).

The categories correspond to highly clustered communities of interactions. As shown in prior work (Stewart, Arif and Starbird 2018), the trolls retweeted and mentioned relatively non-overlapping accounts. We show the limited overlap in clusters here both to validate the hand labeled categories and also to justify the cluster-specific text analyses, since we expect vocabulary to be distinct across clusters as well.

In the pre-processing for the network clustering figure, the user-mention data are represented using rows for the mentioned users and columns for the tweeting users, with the number of user-target mentions as elements. We standardize the rows of this matrix by dividing their row-wise sum[29] and then use the cross-product of that matrix as the graph for the network detection algorithm. This graph represents the co-mentions of trolls conditioning on the overall mentions of the targeted accounts.

[29]With this standardization, the clustering algorithm does not strongly prefer to optimize connectivity for only the most activity accounts, and instead treats accounts relatively equally.

To identify clusters much like Linvill and Warren's, we can use the fast greedy algorithm (Clauset, Newman and Moore 2004) implemented in igraph to maximize modularity in the troll-to-troll graph. Modulary maximization algorithms select a number of clusters and cluster assignments that maximizes the number of within cluster connections and minimizes the number of across cluster connections. Colors are each assigned to the community containing the largest number of a given Linvill Warren category.

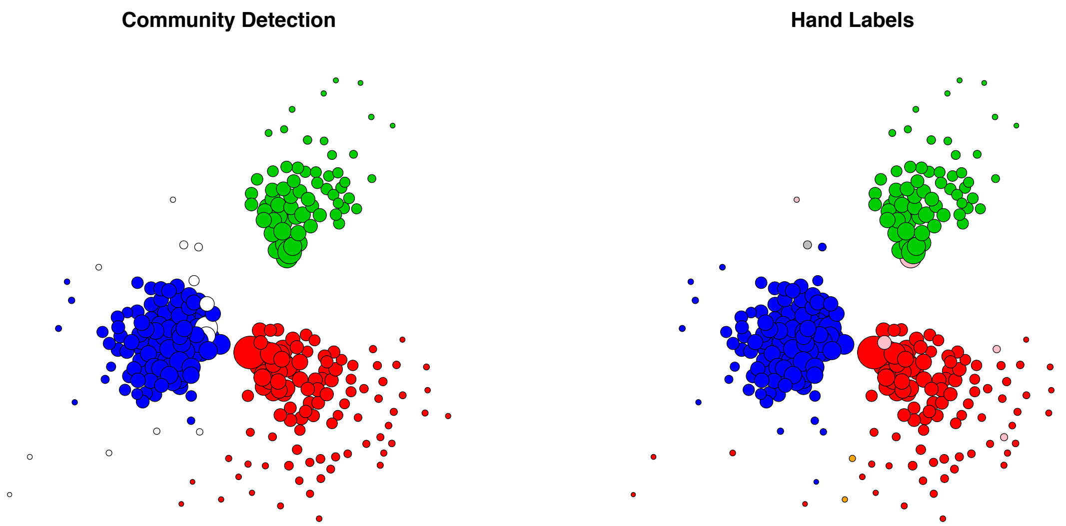

Figure A17: The left panel of this figure shows troll clustering (sharing tweets from the same accounts) using automated community detection while the right panel shows clustering using Linvill and Warren (2020)'s hand coding. Consistent with prior work, account categories can be easily separated using community detection algorithms. Colors are each assigned to the community containing the largest number of a given Linvill Warren category. Accounts with activity below the 50th percentile are not shown in this visualization.

D.6 Voter Suppression

We have presented the strategic use of apolitical content as one strategy to demobilize potential voters on social media. An alternative strategy is voter suppression, or tweets that actively discourage users to participate in the election. For example, this could include tweets saying "boycott the election" or "do not vote".

A handful of studies have documented this behavior during the 2016 presidential campaign. The white paper by cybersecurity firm New Knowledge, commissioned by the Senate, first documented the use of voter suppression tactics across multiple platforms (DiResta et al. 2019). Similarly, a report from the Computational Propaganda Research Project at Oxford also noted some troll activity involved campaigning for African American voters to boycott election on Twitter (Howard et al. 2018). Kim (2018) looks at sponsored advertising discouraging voting on Facebook and Instagram, and finds evidence that these ads targeted nonwhites or likely Clinton voters. However, the focus of prior research has not necessarily been to identify the frequency of voter suppression tweets.

We can use our data to explore to what extent the IRA used a strategy of voter suppression, in addition to distraction from flooding. To do so, we look for any mention of "vote"/"voting"/"voted", "election", "support"/"supported" (i.e. any characters matching "vote", "voting", "election", "support"), as well as negation (to be inclusive here, any characters matching: "not", "n't", "boycott", "sit out", "truth", "rigged", "before", "illegal", "deserv", "fuck"). The additional negation words cover phrases identified by prior studies (DiResta et al. 2019; Howard et al. 2018; Kim 2018) as examples of demobilization from suppression: boycott, don't vote, do not vote, didn't vote, sit out the election, fuck the election, do not support, don't support, can't support, not voting, rigged, before you vote, illegally voted, truth about the election, deserve our vote (presumably implying *don't* deserve). We do not explicitly search for complex and malicious information about the voting process (for more on election incidents using Twitter, see Mebane et al. (2018)).

Beyond this, we use the average sentiment of tweets using the AFINN sentiment lexicon (Nielsen 2011), and categorize the voting tweets above as "negative" for average less than 0 (each word in this lexicon is scored from -5 to 5, with words less than 0 negative). We can also look to what extent this strategy was used by conservative or BLM leaning troll accounts.

Figure A18 documents our findings. While we find some evidence of voter suppression tweets, they are rare, especially in comparison to flooding of entertainment content. Mentions of voting at all are a small fraction of tweets until the last week of the election, very few tweets include the negation and suppression words, and left trolls were no more likely than right trolls to negate or use negative sentiment in voting tweets.

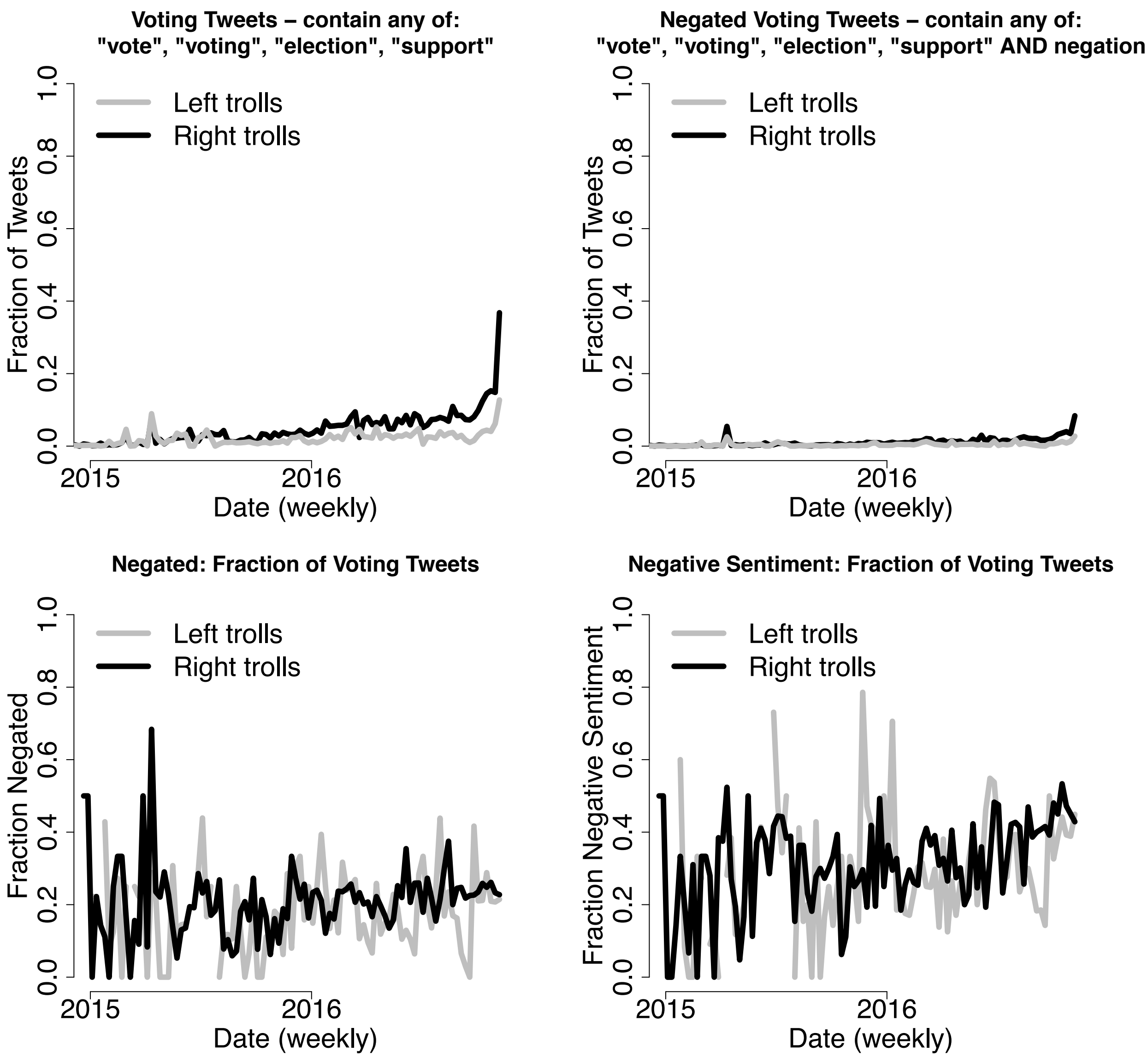

Figure A18: *Voting and voter suppression.* This figure shows that the right trolls mentioned "vote", "election", "support" in around 35% of tweets in the week leading up to the election, while the left trolls tweeted these words in slightly over 10% of tweets. Left trolls were not more likely to negate or use negative sentiment in their tweets about voting.

References

Arif, Ahmer, Leo G. Stewart and Kate Starbird. 2018. “Acting the Part: Examining Information Operations Within BlackLivesMatter Discourse.” *Proceedings of the ACM on Human-Computer Interaction* 2(20).

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. “Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data.” *The American Political Science Review* 110(2):278–295.

Card, Dallas and Noah A North. 2018. The importance of calibration for estimating proportions from annotations. In *NAACL*. New Orleans, Louisiana: pp. 1636–1646.

Clauset, Aaron, MEJ Newman and Cristopher Moore. 2004. “Finding community structure in very large networks.” *Physical Review E* 70(6):066111.

Dawson, Andrew and Martin Innes. 2019. “How Russia’s Internet Research Agency Built its Disinformation Campaign.” *The Political Quarterly* 90(2):245–256.

Deerwester, Scott, Susan T Dumais, George W Furnas, Thomas K Landauer and Richard Harshman. 1990. “Indexing by Latent Semantic Analysis.” *Journal of the American Society for Information Science* 41(6):391–407.

DiResta, Renee, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright and Ben Johnson. 2019. “The tactics & tropes of the Internet Research Agency.”.

Fortunato, Santo. 2010. “Community detection in graphs.” *Physics Reports* 486(3-5):75–174.

Hobbs, William R. 2019. “Text Scaling for Open-Ended Survey Responses and Social Media Posts.” *https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3044864* .

Hopkins, Daniel J and Gary King. 2010. “A Method of Automated Nonparametric Content Analysis for Social Science.” *American Journal of Political Science* 54(1):229–247.

Howard, Philip N., Bharath Ganesh, Dimitra Liotsiou, John Kelly and Camille François. 2018. “The IRA, Social Media and Political Polarization in the United States, 2012-2018.” *Oxford Project on Computational Propaganda Report* .

Kim, Young Mie. 2018. “Uncover: Strategies and Tactics of Russian Interference in US Elections.” *Working Paper* .

King, Gary, Jennifer Pan and Margaret E. Roberts. 2017. “How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument.” *American Political Science Review* 111(3):484–501.

Linvill, Darren L. and Patrick L. Warren. 2020. “Troll Factories: Manufacturing Specialized Disinformation on Twitter.” *Political Communication* 37(4):447–467.

Manning, Christopher D, Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge: Cambridge University Press.

Mebane Jr, Walter R, Patrick Wu, Logan Woods, Joseph Klaver, Alejandro Pineda and Blake Miller. 2018. “Observing Election Incidents in the United States via Twitter: Does Who Observes Matter?” *Working Paper* .

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado and Jeffrey Dean. 2013. “Distributed representations of words and phrases and their compositionality.” *NIPS* pp. 3111—-3119.

Niculescu-Mizil, Alexandru and Rich Caruana. 2005. “Predicting good probabilities with supervised learning.” *ICML* pp. 625–632.

Nielsen, Finn Årup. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs .".

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. "Glove: Global Vectors for Word Representation." *EMNLP* 14:1532–1543.

Stewart, Leo Graiden, Ahmer Arif and Kate Starbird. 2018. Examining Trolls and Polarization with a Retweet Network. In *MIS2*.

Tibshirani, Robert. 1996. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society Series B (Methodological)* 58(1):267–288.

Tucker, Joshua A, Andrew Guess, Pablo Barberá, Cristian Vaccari, Alexandra Siegel, Sergey Sanovich, Denis Stukal and Brendan Nyhan. 2018. "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature." *William and Flora Hewlett Foundation* .

Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, Xueqi Cheng and 2013. 2013. "A biterm topic model for short texts." *ICML* pp. 1445–1456.

Zhang, Youwei and Laurent E Ghaoui. 2011. "Large-Scale Sparse Principal Component Analysis with Application to Text Data." pp. 532–539.

Zou, Hui, Trevor Hastie and Robert Tibshirani. 2006. "Sparse Principal Component Analysis." *Journal of Computational and Graphical Statistics* 15(2):265–286.

Zuo, Yuan, Jichang Zhao and Ke Xu. 2015. "Word network topic model: a simple but general solution for short and imbalanced texts." *Knowledge and Information Systems* 48(2):379–398.