RAND CORPORATION

Research Report

WILLIAM MARCELLINO, CHRISTIAN JOHNSON, MAREK N. POSARD, AND TODD C. HELMUS

Foreign Interference in the 2020 Election

Tools for Detecting Online Election Interference

Foreign election interference is a serious threat to U.S. democratic processes, something that became visible and received public attention in the wake of the 2016 U.S. general election. In the aftermath of that election, it became clear that agents acting on behalf of the Russian government went online and engaged in a very sophisticated malign information effort meant to sow chaos and inflame partisan divides in the U.S. electorate (Marcellino, Cox, et al., 2020). Because of the seriousness of the threat and concerns that such threats are likely to be ongoing, improving the detection of such efforts is critical. That desire to help bolster our democratic processes from illicit interference motivated our current study, which attempted to pilot improved detection methods prior to the 2020 election—we wanted to detect any such efforts in time to provide warning rather than post hoc.

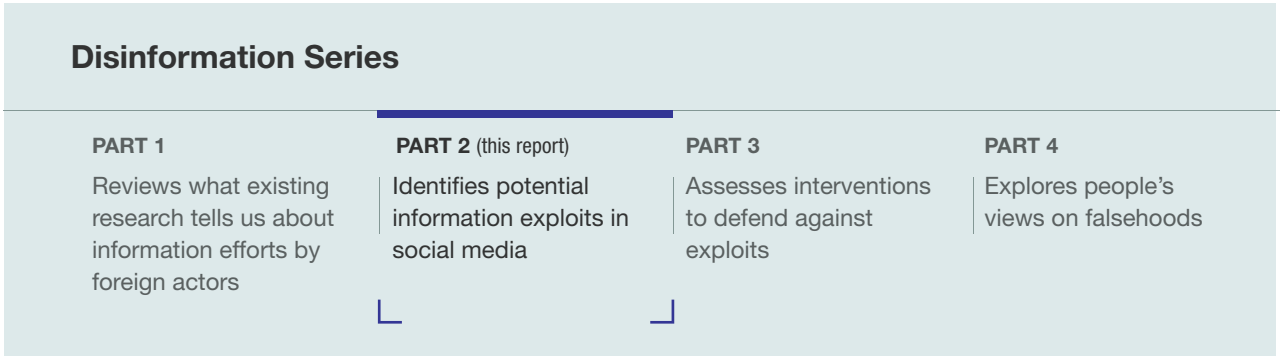
We found convincing evidence of a coordinated effort, likely foreign, to use social media to attempt to influence the U.S. presidential election. We examined two kinds of suspicious accounts working in concert toward this end: The first kind is *trolls*: fake personas spreading a variety of hyperpartisan themes.[1] The second kind is *superconnectors*: highly networked accounts that can spread messages effectively and quickly. Both kinds of accounts cluster only in certain online communities, engage both liberal and conservative audiences, and exacerbate political divisions in the United States.

This report is the second of a four-part series (Figure 1) for the California Governor's Office of Emergency Services designed to help analyze, forecast, and mitigate threats by foreign actors targeting local, state, and national

KEY FINDINGS

- We found credible evidence of interference in the 2020 election on Twitter.
- This interference includes posts from troll accounts (fake personas spreading hyperpartisan themes) and superconnector accounts that appear designed to spread information.
- This interference effort intends to sow division and undermine confidence in American democracy.
- This interference serves Russia's interests and matches Russia's interference playbook.
- Our methods can help identify online interference by foreign adversaries, allowing for proactive measures.

FIGURE 1
What This Series Covers



elections. This report describes what appears to be foreign online election interference and offers recommendations for response. Appendix A provides detailed descriptions of our methods.

Before laying out major findings, we want to acknowledge caveats to our study. First, our analysis is limited to Twitter data, which we chose both because of availability—such platforms as Facebook do not make user data public in the same way—and because the social nature of Twitter allowed us to use network analysis methods. In essence, *mentions* (replies and retweets) allow an algorithm to group Twitter users into communities according to their frequent interactions. In turn, that allows us to examine and compare communities—comparisons between communities can make suspicious accounts and behaviors clear and detectable.

Second, our choice of search terms shaped our data set and thus our results. We chose to use search terms aimed at capturing the broad election conversation and did not shape our query around various candidates. For example, we captured talk centered on major candidates, such as Vermont Senator Bernie Sanders, but did not capture a meaningful set of data centered on the campaigns of Minnesota Senator Amy Klobuchar or New York Mayor Michael Bloomberg. Thus, our findings about election interference regarding any given campaign come with the caveat that we don't know what we would have found if we focused on individual candidates instead of the broader conversation surrounding the presidential election.

Finally, we cannot firmly attribute election interference activity to a specific source, although the tactics we found do match Russia's prior efforts, and there is other evidence that election interference is being conducted by Russia (and possibly by other nations) (Select Committee on Intelligence of the United States Senate, 2019, undated; Office of the Director of National Intelligence, 2020; *United States v. Internet Research Agency LLC*, 2018). Although we feel confident that we discovered a coordinated effort, we cannot definitively attribute that effort to a specific actor.

Election Interference: Trolls and Superconnectors

In this section, we lay out the advocacy communities identified in our data that are arguing about the election, describe the two kinds of suspicious accounts we found in those communities, and give illustrative examples of how these accounts functioned.

Mapping Out the Rhetorical Battlefield

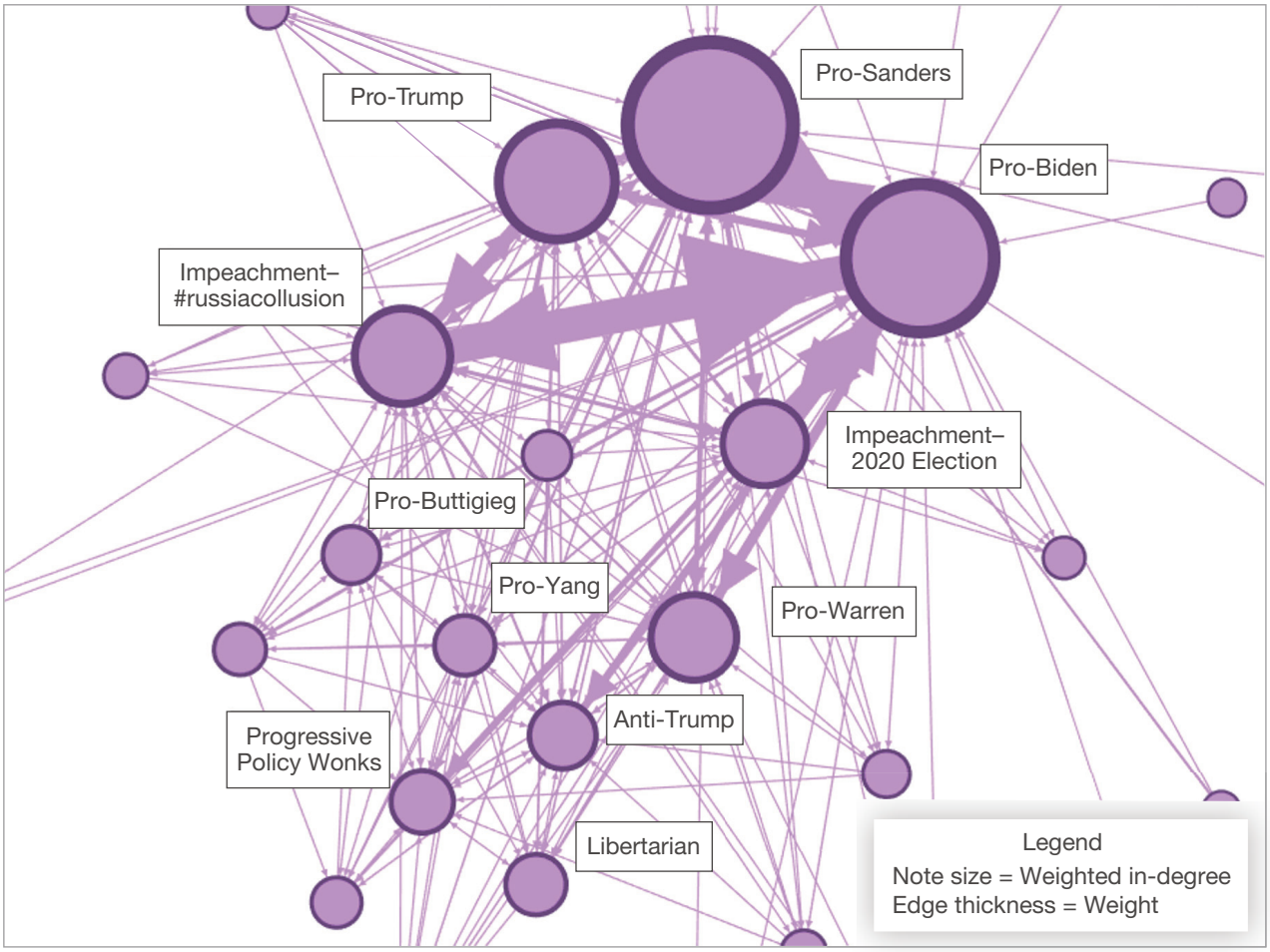
This work builds off of prior work (Marcellino, Cox, et al., 2020) for the United Kingdom's (UK) Ministry of Defence (MoD) piloting the detection of interference efforts through the use of both network analysis and machine learning (ML). That previous effort used data related to the 2016 U.S. general election, a known target of Russian interference efforts using trolls—in this case, social media accounts that appeared to be held by Americans talking about the presidential election but were fake personas controlled by workers at the Internet Research Agency.

We built off of that effort in two ways: (1) using the RAND Corporation's Community Lexical Analysis (CLA) method to identify advocacy communities discussing the 2020 U.S. general election on Twitter (Bodine-Baron et al., 2016; Marcellino, Marcinek, et al., 2020), and then (2) using ML to find trolls working in those communities.

CLA works by combining network analysis (discovering who is talking to whom) with text-analysis methods (understanding what those groups are talking about). For this, we used RAND-Lex,[2] a software suite that combines ML, network analysis, and computer-assisted text analysis. This allowed us to take a very large data set of 2.2 million tweets from 630,391 unique accounts collected between January 1 and May 6, 2020, and make sense of the online rhetorical battle over the upcoming election.

Figure 2 shows this rhetorical battlefield. Each node represents a community of Twitter accounts engaged in regular conversation with each other. The 11 largest communities, ranging in size from approximately 7,000 accounts to 150,000 accounts, have descriptive labels. Our figure shows the direction of connections in the network; the largest nodes are the most central as measured by incoming communication (in-degree), connected by many incoming connecting lines (edges). Those least connected are at the periphery. Each edge indicates interactions between communities, and the thicker (higher-weighted) the edge, the more interactions there are. Each edge is an arrow showing directions, but some are so small

FIGURE 2
Twitter Communities Discussing the 2020 General Election



SOURCE: RAND analysis of Twitter data, 2020.

that the point of the arrow is invisible. However, for the largest and most-central communities, the interactions are so dense that the point of the arrow is visible.

The community detection algorithm in RAND-Lex detects which accounts are in frequent communication, thus implying social membership, and then bins all the tweets from each community into data sets for follow-on characterization of each community via text-mining.[3] This allows a human analyst to make sense of the tweets in each community, which can number from the tens of thousands to the hundreds of thousands. Table 1 summarizes the communities.

Interfering with Both the Left and the Right, Along with Candidate Preference

In addition to understanding the rhetorical landscape, mapping out these communities was important because we found that trolls and superconnectors were clustered in specific communities. It is one thing to see trolls and superconnectors as general and consistent phenomena in political conversation. It is quite another thing to see that only a few communities have these suspicious accounts in high concentrations. In addition to identifying which communities were most targeted by trolls and superconnectors, we were able to measure the relative intensity of targeting between communities. A normal (baseline) percentage for superconnectors is 2.5 percent, and 5 percent would be an even distribution for troll accounts—numbers significantly higher than those are noteworthy concentrations.[4] Table 2 shows the distribution of both trolls and superconnectors by community, with the top three highest concentrations for each type bolded.

In Table 2, all of the communities have concentrations of superconnectors that are higher than the baseline, but the three that are bolded in each column have particularly high concentrations relative to the rest. The three communities with the highest troll concentrations are also bolded, although the differences in concentration are less pronounced: The community with the fourth-highest troll population (Pro-Buttigieg—at 5.98 percent), is only slightly

TABLE 1
Summary of Largest Communities

Community Label	Description
Pro-Biden	Broad discussion of former Vice President Joseph R. Biden and President Donald J. Trump in the election, generally pro-Biden
Pro-Sanders	Support for Sanders and progressive polices
Pro-Trump	Pro-Trump discussion, along with support for QAnon and deep state conspiracy theories[a]
Pro-Warren	Support for Massachusetts Senator Elizabeth Warren and progressive policies
Impeachment–#russiacollusion	Impeachment proceedings discussion, strong anti-Trump tenor
Impeachment–2020 Election	General discussion of how the impeachment would affect the election
Anti-Trump	Broad anti-Trump discussion on a variety of issues
Progressive Policy Wonks	Discussion focused on technical policy and budget, from progressive perspective
Libertarian	Libertarian discussion: counter-Democrat with some Trump support
Pro-Yang	Supportive of entrepreneur Andrew Yang and his policies discussion
Pro-Buttigieg	Supportive of former South Bend, Indiana, Mayor Pete Buttigieg and his policies discussion

SOURCE: RAND analysis of Twitter data, 2020.

NOTE: Communities are listed by size, as depicted in Figure 2.

[a] Adherents of the deep state conspiracy believe that a powerful cabal secretly controls the U.S. government and operates an international child sex-trafficking ring that serves powerful elites. QAnon is an anonymous online persona who claims to be a highly placed government insider, working with President Trump to expose and dismantle the secret deep state.

TABLE 2
Distribution of Suspicious Accounts by Community

Community	Accounts	Superconnectors (%)	High Troll Scores (%)
Pro-Biden	159,576	**10.96**	4.00
Pro-Sanders	91,241	3.90	2.68
Pro-Trump	87,712	**21.25**	**8.10**
Pro-Warren	26,454	2.91	4.50
Impeachment–#russiacollusion	23,858	**11.40**	**6.00**
Impeachment–2020 Election	16,631	6.48	2.28
Anti-Trump	13,647	5.01	2.01
Progressive Policy Wonks	7,359	4.38	2.77
Libertarian	4,832	3.83	**6.31**
Pro-Yang	4,478	4.49	2.57
Pro-Buttigieg	1,889	5.77	5.98

SOURCE: RAND analysis of Twitter data, 2020.
NOTE: Bolded items have particularly high concentrations.

lower than the community with the third highest troll population (Impeachment–#russiacollusion—at 6.00 percent).

Among these communities with the three highest superconnector and identified troll concentrations, there are two politically right-leaning (a Libertarian community, which had a high percentage of trolls, and the Pro-Trump community, which had the highest percentage of both trolls and superconnectors). Two politically left-leaning communities were also in the top three: the Pro-Biden community had a high number of superconnectors, and the Impeachment–#russiacollusion community had high numbers of both superconnectors and trolls.[5] Our interpretation is that election interference and manipulation is being directed toward both sides of the U.S. political spectrum. Such a strategy is consistent with prior Russian activity and Russia's theory of information conflict, but we cannot directly attribute these actions to Russia (Posard et al., 2020).

Troll and superconnector activity in these communities might have worked in favor of President Trump and against Biden. Accounts identified as likely trolls in the Pro-Trump community were strongly supportive of the President, QAnon content, and anti-Democrat content that favored the President's candidacy. In contrast, the Pro-Biden community overall strongly supported Biden, but the troll-identified accounts in that community were anti-Biden—that is, they either criticized Biden or praised Senator Sanders. We also found that trolls and superconnectors both boosted hashtags that worked against Biden's campaign. Based on this activity (and assuming the Pro-Sanders community support was not genuine but rather meant to hurt the Biden campaign), we infer there was a preference for Trump's campaign in this interference effort, which dovetails with other research on Russian interference with the 2020 election (Frenkel and Barnes, 2020). Our methods for finding trolls and superconnectors, and illustrative examples of their behavior, are detailed in following sections.

Trolls

Troll Hunting with Machine Learning

Mapping out the communities within the 2020 election discussion meant we could then look efficiently for online interference efforts. In the 2016 election, Russian interference was targeted at specific communities talking on Twitter, and we expected that this tactic might continue. We thought that having discrete data subsets (the different communities) might make interference efforts easier to detect by contrast,

and this proved to be the case: The ML model we used found what were likely troll accounts, and these accounts were clustered in specific communities.

We built our ML model using the Twitter output from approximately 800 known troll accounts from the 2016 election.[6] We used multiple combinations of features to build and test various models, using semantic content, linguistic style, and metadata.[7] We found that a hybrid model of semantic content and linguistic style performed the best,[8] improving performance from 80 percent to 97 percent on training data using only the semantic content.[9] Given that even small performance improvements of a few percentage points in this kind of ML task can be very difficult, this very large improvement was extremely promising.

However, we found this model was hindered by data limitations and thus not as useful for finding trolls in our 2020 data. The training data we built our model from consisted of every tweet from the known 2016 Russian trolls; the 2020 election data we analyzed was just one conversation—a slice of tweets from the more than 630,000 accounts we were analyzing. Using only data from this very heated partisan discussion resulted in many authentic accounts looking like trolls to the model. If we had full access to all tweets from each of those 630,000 accounts and sufficient computing resources, we then could have used the superior-quality hybrid model, but given these data restrictions, we selected a model that performed at a slightly lower level but seemed relatively robust to the differences between our data samples.[10]

Our ML model was trained as a *binary classifier*, but rather than design it to return a troll-or-not result, we set it up to return a likelihood match rating to the known trolls from the 2016 election that we used to build our model. We then verified how our model performed on a new data set by manually inspecting 130 accounts with the highest troll ratings across the communities. Accounts at the top of this range looked like trolls: They were hyperpartisan, posted very little or no original content, engaged in round-the-clock retweeting, and shared only political content (nothing about family, hobbies, etc.). At the low end of this range, accounts looked more like real people: They had some original content in idiomatic American English, expressed humor and responses that require American cultural knowledge, and shared some nonpolitical content. An important finding was that a few of the accounts that our model gave high troll scores were so hyperpartisan that they looked like trolls to our model but were identified as real people in the United States through manual inspection (and, in one case, Twitter verification).

Given that our prior research has shown that Russian trolls are trying to imitate U.S. political partisans at each extreme (Marcellino, Cox et al., 2020[11]), we think that our model is useful because the accounts to which it gave high troll ratings closely matched known Russian trolls in two distinct ways. The first way involves language: Our model's performance on data from verified Russian trolls from 2016 shows that posts by Russian trolls have distinctive language patterns, even though they might sound similar enough to other posts to fit in with a political community. The accounts we identified as highly troll-like used that distinctive language pattern in the 2020 election conversation. In addition to this ML matching for a specific conversation, we reviewed those accounts' output as a whole, looking to see whether these suspicious accounts broadly acted like trolls: whether they posted original content or exclusively retweets, whether they ever tweeted about family or nonpolitical themes, whether there were breaks in activity or nonstop, day-and-night tweeting, and whether they shared extreme partisan content. Of the 130 accounts we manually inspected, only two appeared to belong to real people. For those two apparent false positive identification as trolls, intense partisanship matched the first tell (the linguistic match in the conversation captured in our data set), but the overall accounts did not look like those of trolls. The other 128 accounts did match on both fronts. Our model probably does have some false positives (mainly because we analyzed a partisan conversation rather than the entire output of each account), but it worked well overall.

What Themes Are Troll Accounts Pushing?

Our ML text analysis and human qualitative review indicated that these troll accounts were characterized by a variety of partisan themes. The qualitative review involved looking up these suspected accounts on Twitter and assessing their "Tweets & Replies"

for patterns that would summarize the qualities of that account's content. For example, an account that posted negative content about "anti-American Marxists," "socialists," "leftists," "ANTIFA thugs," or "communists," was tagged for qualitative review as "Democrats are Communists/Socialists." This kind of coding is inductive and emerged as more accounts were reviewed to detect common themes. Figure 3 summarizes some of the prominent themes from these troll accounts.

Many of the themes in Figure 3 have parallel qualities. Both types of trolls floated conspiracy theories that Jeffrey Epstein (a now-deceased financier charged with sex trafficking of minors) had incriminating evidence about powerful members of the opposite party that would soon come to light (U.S. Department of Justice, 2019). Coronavirus theories also had a kind of parallel promotion: for politically left-leaning accounts, that the pandemic would be used to install fascism or that federal responses were directed to undermine Democratic or minority communities; for politically right-leaning accounts, that the pandemic was a hoax or exaggerated to influence the election. Finally, both types of trolls expressed either support (politically left-leaning) or opposition (politically right-leaning[12]) for Black Lives Matter activists.

However, there were other themes that were parallel in function but more partisan in specifics. Troll accounts from politically right-leaning and politically left-leaning accounts shared content, largely manipulated photographs, that was critical and disparaging of former First Lady Michele Obama, First Lady Melania Trump, and President Trump's daughter, Ivanka, as a kind of indirect criticism of their ideological opponents. Troll accounts also posted material that constructed the opposition in hyperpartisan terms: Politically left-leaning trolls talked about Republicans or political conservatives as fascists or Nazis; politically right-leaning accounts talked about progressives as communists or socialists. Politically right-leaning trolls spoke frequently about the *deep state*—an amorphous conspiracy theory that describes relationships among a variety of national security and law enforcement agencies—and plans to confront it, mixed with statements of support and reference to QAnon claims. These politically right-leaning accounts also shared a mix of candidate-specific criticism and mockery of former Vice President Biden, Senator Sanders, former Secretary of State Hillary Clinton, and former President Barack Obama. Politically left-leaning troll accounts shared specific themes around Trump as a Russian-owned traitor, criticism of Biden as insufficiently progressive (these trolls were in the Biden community), and messages sharing "peace," "love," and "good vibes." Although this last theme might seem oddly nonpolitical, that positive-affect language was a hallmark of politically left-leaning trolls in the 2016 election interference (Marcellino, Cox, et al., 2020).

Trolls and Boosting Candidates

In addition to spreading hyperpartisan themes, trolls appeared to engage in coordinated campaigns to

FIGURE 3
Examples of Troll Account Themes by Political Affiliation

POLITICALLY LEFT-LEANING THEMES

- Black Lives Matter (pro)
- Jeffrey Epstein connected to powerful Republicans
- Coronavirus theories
- Anti–Melania Trump or anti–Ivanka Trump
- Republicans are "Nazis" or "fascists"
- Trump as traitor
- Biden as insufficiently liberal
- Sending love and "positive vibes"

POLITICALLY RIGHT-LEANING THEMES

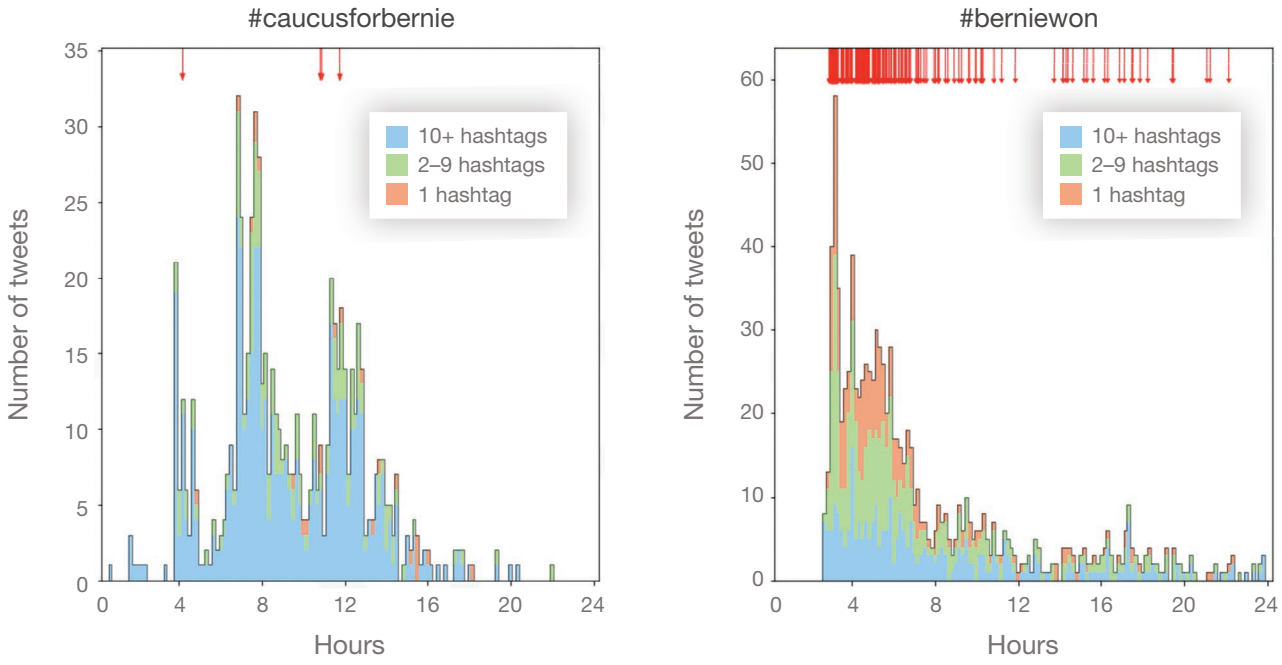
- Black Lives Matter (anti)
- Jeffrey Epstein connected to powerful Democrats
- Coronavirus theories
- Anti–Michele Obama
- Democrats are "communists" or "socialists"
- Deep-state conspiracy theories
- QAnon conspiracy theories
- Anti-Biden, Sanders, or Hillary Clinton

SOURCE: RAND analysis of Twitter data, 2020.

support candidates by boosting hashtags—another activity that we detected as an orchestrated election interference effort. Authentic Twitter users tend to use multiple hashtags (for example, if an authentic poster used #caucusforbernie, that poster was also likely to use other related hashtags). We found that specific, potentially strategic hashtags were suddenly boosted on troll accounts as the only top-trending hashtag they used. In Figure 4, we compare two Sanders-specific hashtags: #caucusforbernie (referring to the Iowa caucuses on February 3, 2020) and #berniewon (which appeared shortly after the Associated Press declared Senator Sanders the victor in the Nevada caucuses on February 23, 2020). The left plot of Figure 4 is normal: The overwhelming majority of accounts that tweeted the hashtag are blue, meaning they used ten or more of the 234 top-trending hashtags in our data. The one on the right is abnormal, dominated by a separate population of accounts (many of them trolls) that used only a single hashtag (#berniewon) over the period that our data covered.[13]

Tweets in Figure 4 are colored by the number of other prominent hashtags that was used by each user—orange indicates that a particular user only used that hashtag; blue indicates that a user frequently tweeted other prominent hashtags. It is possible that the orange population depicted in the right plot, which used only a single hashtag in this conversation between January and May, is a subcommunity of the broader Pro-Sanders community that behaves very distinctly. Our data only cover a specific conversation about the 2020 election—our point is that, within this election conversation, one population of users employed the use of a single hashtag, which was very unusual. In addition, the #berniewon population was dominated by users that were given high troll scores; the #caucusforbernie population showed very few accounts with high troll scores. Therefore, we think

FIGURE 4
Normal Versus Troll-Boosted Hashtags



SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This figure presents a comparison of tweet frequency for two pro-Sanders hashtags. The left panel reflects frequency for #caucusforbernie (referring to the Iowa caucuses on February 3, 2020); the right panel reflects frequency for #berniewon (referring to Sanders' victory in the Nevada caucuses on February 23, 2020). Note that the vertical axis is different between the two plots. Tweets are binned into 10-minute intervals and colored by the prevalence of other fast-rising hashtags that each account used over our observation period. The data show that accounts that tweeted #caucusforbernie were typically frequent users of fast-rising hashtags; #berniewon was mostly tweeted by accounts that rarely used fast-rising hashtags. The vertical red lines in each plot show the time of tweeting by an account that our ML model gave high troll scores; few such accounts tweeted #caucusforbernie, but they made up a significant proportion of accounts that tweeted #berniewon.

that the rise of the #berniewon hashtag could be the result of troll accounts that rarely use hashtags except to boost a particular narrative—in this case, to establish a Senator Sanders victory before the official results had been completely tabulated (this can be seen in the plot farthest to the right with the early stacking of single hashtag users around the 2-hour mark).

Superconnectors

The other kind of suspicious account we found were *superconnector* accounts: accounts with friend and follower numbers very close to or circumventing the restrictions that Twitter places on authentic accounts. (In the Twitter lexicon, *friends* are the accounts followed by the account in question; *followers* are the accounts that follow the account in question.) We found coordinated campaigns using superconnectors to support candidates by boosting hashtags.

To prevent manipulation, Twitter caps friends at 5,000 for most accounts, a limit that can be exceeded only if an account also has a large number of followers (verified public figures get a blue checkmark on their account and are not limited). Concentrated networks of this kind of account are particularly well suited to transmitting a large volume of messages, and can be engineered because the networked accounts follow each other. Although such highly connected accounts can happen naturally, they are relatively rare (as illustrated by the distributions in Table 2).

We first encountered superconnectors when investigating the accounts that were rated as most likely to be trolls by our ML model. A pattern quickly became apparent among these accounts: They were highly connected yet had few interactions with other accounts aside from their high frequency of retweets. Our suspicions deepened when we analyzed the distribution of these accounts and found that they were disproportionately common only in a few communities—the same communities, in fact, that showed a similarly high number of suspected trolls. We defined superconnectors as those accounts with more than 4,500 friends and fewer than 1.2 followers per friend, which captured much of the behavior we were interested in. In a sample of nonpolitical Twitter discourse (e.g., posts about sports, movies, and games) from the same January–May period of 2020, such superconnectors are dispersed and make up approximately 2.5 percent of the total accounts. In our election 2020 data set, the percentages were much higher in specific communities—for example, 21.25 percent in the Pro-Trump community and 11.4 percent in the Impeachment–#russiacollusion community.

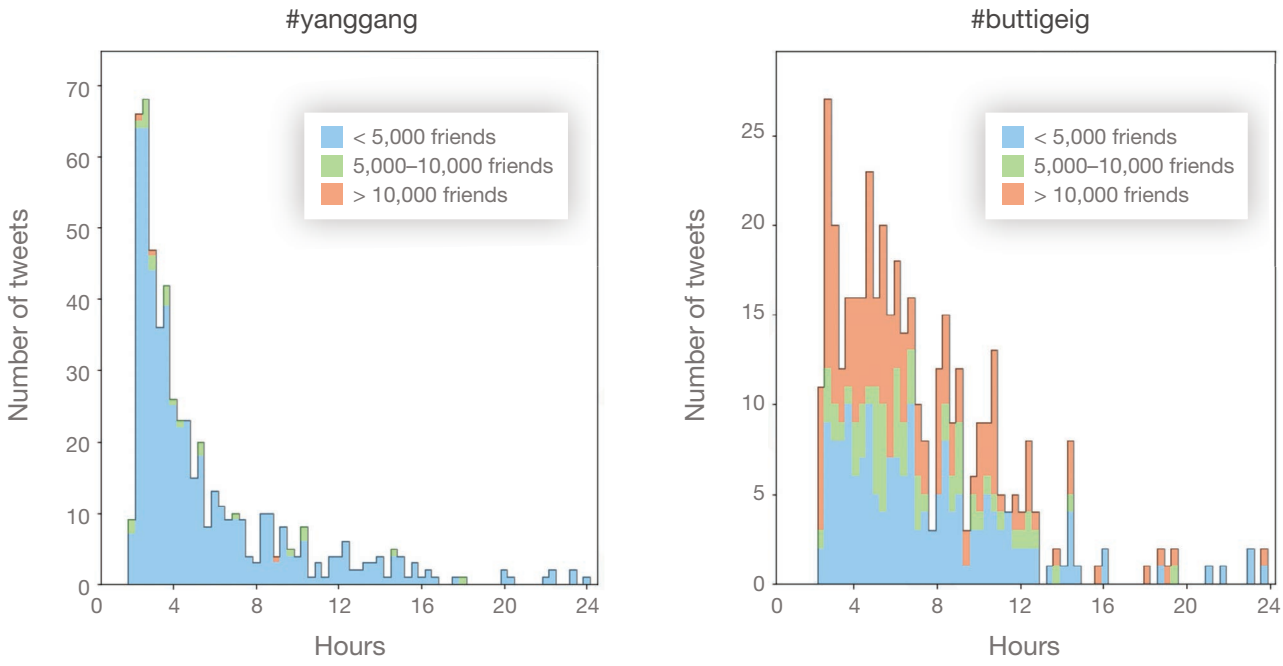
In addition to being so strongly concentrated in only a few communities, superconnectors in our data sometimes exhibited a suspicious pattern of boosting specific hashtags. As an example, the left-hand panel of Figure 5 illustrates a hashtag tweeted by accounts that have non-superconnector numbers of friends and followers (in blue), and the right-hand panel illustrates that of that one tweeted by superconnectors (in orange and green).

Figure 5 shows the spread of two candidate-specific hashtags: #yanggang and the *misspelled* #buttigeig. Each plot shows the number of tweets for each hashtag among our data set. The hashtag #yanggang spread mostly in the Pro-Biden community; #buttigeig spread mostly in the Pro-Trump community. We chose this example because #buttigeig in practice had little to do with the actual candidate Pete Buttigieg; instead, it was tacked on to a tweet regarding conspiracy theories surrounding Hunter Biden. An unusual fraction of the accounts that retweeted it are superconnectors (shown in orange), compared with the relatively innocuous #yanggang, which was mostly spread by authentic accounts. It is possible that people could retweet a misspelled hashtag, but the concentration of superconnectors sharing it to spread anti-Biden content in a targeted community is suspicious.

Conclusions

Our analysis of early 2020 Twitter discourse about the general election found two kinds of suspicious accounts: *trolls* (fake personas spreading a variety of hyperpartisan themes) and *superconnectors* (highly networked accounts that can spread messages effectively and quickly). We found both of these types of suspicious accounts to be overrepresented in specific communities (two politically right-leaning communities and two politically left-leaning ones). Troll

FIGURE 5
Normal Versus Superconnector Hashtag Boosting



SOURCE: RAND analysis of Twitter data, 2020.

accounts, with their nonstop partisan messaging, are well suited for stoking division; superconnectors, when they become active, are well suited for spreading messages because of the density of connections these accounts have.

An important caveat is that we cannot definitively conclude from any single part of our analysis that there was a coordinated foreign interference effort at the time we analyzed this particular data set. Our ML model was based on 2016 Russian tactics and those assumptions might not transfer fully if Russian tactics are dramatically different in 2020. Another possibility is that our model identified efforts that mimic 2016 Russian tactics. We also acknowledge that superconnectors occur naturally on Twitter, albeit in small numbers. We have inferred a coordinated effort based on the following intersecting findings:

- Our model showed trolls clustered in high numbers in three specific communities.
- The accounts with the highest troll ratings were engaged in activity consonant with Russian interference goals and tactics (e.g., spreading hyperpartisan themes, undermining or supporting specific candidates, and boosting certain hashtags).
- We found superconnectors clustered in high numbers in three specific communities.
- These superconnectors were also engaged in activity consonant with Russian interference goals and tactics (e.g., boosting hashtags and undermining or supporting specific candidates).

What we found dovetails with our prior and ongoing research: Russia seeks to boost existing political partisanship in the United States, and its strategy involves leveraging existing partisan tensions that already exist organically, helping to create an "us vs. them" political discourse (Marcellino, Cox, et al., 2020; Posard et al., 2020).

All of our findings make sense within a larger framework for malign Russian information efforts. Russia's highest aim in these efforts is to elicit strong partisan reactions and create a sense of disunity (although operators might have preferences for particular electoral outcomes).[14] A nation that is in conflict domestically (or at least talks as if it were) is less

able to exert influence and counter Russia's political goals. This is a longstanding Russian strategy, but social media has made it cheaper and easier than ever to conduct such efforts. For these reasons, we infer that there is ongoing election interference over social media for the 2020 election, and (based on how our findings reflect Russian practices and goals) it is possible that the effort we detected is part of a Russian information effort to sow chaos in the United States.

Our primary focus in this report is on the insights already mentioned, but our analysis indicates that our innovation in methods—specifically, using social network analysis to map out communities and improving ML through hybrid models—is also important. In the first case, using network analysis to create smaller data sets by community allowed us to find suspicious activity by comparing those communities. An important piece of evidence indicating that there was a coordinated election interference effort was finding superconnectors that worked in concert to boost hashtags in support of specific candidates. Although superconnectors can happen naturally, they are rare and dispersed, and thus would have been background noise in our original data set. We were only able to notice them through the community network analysis: visualizing that they were concentrated in specific communities was the clue we needed to start examining them (see Figure A.3).

Additionally, combining semantic content and style to create a hybrid model is a potentially powerful advance in ML detection of interference. We were not able to take advantage of our high-performance hybrid ML model because we lacked full access to Twitter data, but that would not be a barrier for social media platforms that want to better guard against interference. Continued innovation, such as these novel combinations of methods, could be fruitful in the battle between concealment and detection—offense and defense—in election interference.

Recommendations

This study was part of a larger effort designed to test possible protective interventions and provide a broad framework for responding to election interference (Posard et al., 2020; Helmus, Marrone, and Posard, forthcoming). Our aim was to advance election interference detection methods, and our recommendations reflect that. Although our study was sponsored by the state of California, other states and the U.S. government could benefit from our work. We recommend several measures as part of a holistic effort to help protect vulnerable populations from manipulation.

Continue to Innovate in Methods

By combining network analytics and ML, we were able to uncover an election interference effort that had not been detected. Our methods are repeatable, and our advances in hybrid modeling could improve ML detection even further with full access to data. We hope social media platforms will respond to the effort we detected and will be open to adapting these and other emerging methods. Although technological innovation enables the spread of malign influence efforts and the increasing ease with which such efforts are conducted, technical innovation can also work to combat these efforts. We recommend continuing to experiment with and innovate technical solutions to counter the scale of this sort of malign effort.

Continue to Publicize the Threat, Targets, and Tactics

Other interference efforts have been discovered and publicized, and we recommend continuing that practice. We recommend publicizing the threat broadly using multiple channels (e.g., radio, print, TV) to help alert Californians (and the American public) to ongoing—and, most likely, foreign—efforts at manipulation that undermine confidence in democracy. Publicizing the effort should feature details, such as the target audiences (e.g., supporters of Trump and Biden), and specific tactics (e.g., sharing attack memes of first ladies or normalizing the words "fascist" and "socialist" to describe other Americans). A nonthreatening way to help defend against interference efforts would be to issue a warning to Californians and the rest of the nation's citizens that that they are still being targeted for manipulation.

Appendix A. Description of Data and Modeling of Trolls

Model Building and Training

The main goal of our modeling efforts was to reveal descriptive features of Twitter trolls participating in discourse related to the 2020 U.S. election. There are many kinds of inauthentic accounts—e.g., foreign trolls, domestic trolls, bots—in contrast to authentic accounts that appear to be run by humans acting in good faith.

To study various inauthentic accounts, we devised a strategy with two main components: First, we built and trained ML models to classify Twitter accounts as trolls or non-trolls, using training data (output from known Russian trolls interfering with the 2016 election) from 2015 and 2016 that had been compiled previously by academic researchers. Second, we applied a selected ML model to our 2020 data set and manually inspected the accounts with the highest estimated probability of being trolls, hoping that these accounts would be representative of inauthentic accounts writ large. During this process, we discovered that certain features of these accounts' metadata (numbers of friends, followers, and retweets) might be indicative of inauthentic behavior. As a control, we examined the metadata of Twitter accounts participating in nonpolitical discourse, assuming that these were mostly authentic users. Finally, we combined what we had learned with the results of RAND's community detection method (described later in this appendix) to chart where potentially inauthentic activity was taking place within the broader political discourse.

For this effort, we brought all those previous methods to bear on social media data collected between January and May 2020. One line of effort was to combine networks analytics, text-mining, and qualitative analysis to map out the online argument space on Twitter. This allowed us to visualize and understand the various online stakeholder groups that support, oppose and argue about various candidates running for president. The second line of effort was to adapt and improve the "troll hunter" (ML) model developed in our UK MoD study (Marcellino, Cox, et al., 2020).

Data

We used Twitter data from 2020 about the U.S. general election. We used the commercial service Brandwatch (undated) to collect tweets between January 1 and May 6 that contained the following search terms:

> ("US general election" OR "US presidential election") OR (("2020 Election" OR "Election 2020" OR "national election" OR "general election") AND (Trump OR Biden OR Sanders OR Warren))

In essence, we queried for variations of "US general election" combined with talk about the most-

To study various inauthentic accounts, we devised a strategy with two main components: First, we built and trained ML models to classify Twitter accounts as trolls or non-trolls. Second, we applied a selected ML model to our 2020 data set and manually inspected the accounts with the highest estimated probability of being trolls.

prominent candidates. Earlier query attempts had indicated that omitting candidate names increased the likelihood of picking up international conversations not relevant to our study. As it was, we still collected a large set of English-language discussion about the U.S. general election that text-mining revealed was clearly from Indian English speakers talking about how the election might affect India. (After conducting the network analysis, we excluded this data from our study.)

Our query yielded 2.2 million tweets from 630,391 unique accounts.[15] The number of tweets per account was very unbalanced: The top 1,000 accounts (just 0.2 percent of the total) generated 5 percent of all the tweets in the data set. By contrast, nearly 600,000 accounts had fewer than 10 tweets in the data set. Because of the search method we used, these figures do not reflect all of the tweets that each of the 630,000 accounts sent out, nor do they reflect the total number of people tweeting about political matters during the first three months of 2020: They reflect only those tweets that met the query criteria.

Research Overview

We combined two lines of effort to detect election interference. We thought of this as being a problem fundamentally about detecting a faint signal against loud background noise: a relatively small number of inauthentic accounts hiding in a sea of authentic ones. Such data reduction approaches to amplifying signal have been used in other efforts—for example, to detect information operations aimed at specific audiences within much larger social media conversations (Marcellino, Cox, et al., 2020).

Our first line of effort was to combine networks analytics, text-mining, and human qualitative analysis to map out the online argument space on Twitter. Doing this allowed us to visualize and understand the various online stakeholder groups supporting and arguing about various candidates running for president. This approach also functioned as a strategy for data reduction: In place of an enormous unsorted pile of 2.2 million tweets, we identified the ten largest communities engaged in arguments about the U.S. general election, grouped by their social interactions, specific concerns, and ways of speaking.[16]

The second line of effort was to adapt and improve the "troll hunter" ML model developed in our previous UK MoD study. For this, we used an existing data set of U.S. government–identified Russian trolls that engaged in election interference in 2016. Working backward from posts by those trolls, we were able to capture how they attempted to influence the 2016 election conversation on Twitter and build an effective ML model using only linguistic stance that identified trolls in that data set. We hoped to build off of that work and improve it using various ML approaches, notably a novel hybrid approach that combined deep neural network embeddings with linguistic stance. Although this hybrid approach did improve performance significantly on our 2016 training data, it did not generalize well to our new data, likely because our data sets differed in completeness. In the previous study, verified identification of approximately 800 trolls meant we could harvest all of their data; we only had a fraction of the content for each account involved in this specific conversation on the 2020 election. Stance measurements improve performance of the model by picking up on some of the features of partisan political discourse, but adding linguistic stance in this very partisan political discourse selection increased false positives: Many partisans were falsely identified as trolls. However, we did find that the deep word embedding approach worked well, and segmentation of the data into communities helped us discover another kind of suspicious account: highly networked superconnectors clustered in specific communities.

Detecting Communities Via Network Analysis

We used RAND-Lex, RAND's proprietary text and social media analysis software suite, to conduct this step of the analysis (Irving, 2017). The community detection algorithm in RAND-Lex detects which accounts are in frequent communication, thus implying social membership, and then bins all the tweets from each community into data sets for follow-on characterization of the communities via text-mining. The broad outlines of this step are as follows:

1. The software first built a mentions network to determine all interactions within the data set.[17]

2. Using Louvain modularity, the software then applied a community detection algorithm to infer communities from the relative frequency of those interactions (Blondel et al., 2008).
3. The tweets from each community were grouped, based on community membership, into data subsets, allowing for human qualitative text analysis.
4. We then focused on community characterization. This detection step discovers the communities in the data set but tells us nothing about the discussion or character of the community. We used a mixed-method analysis of each community's texts, combining human qualitative and machine text analysis. Two text-mining methods apply thresholds for frequency and distribution to help better characterize the community as a whole:
   a. *Keyness testing* finds words that are conspicuously overpresent or underpresent in a text collection compared with a baseline collection. For CLA, the baseline comparison is all the other communities' tweets—a one-against-many comparison that shows what words are characteristic of a given community (Scott, 1996, 1997, 2015).
   b. *Collocate extraction* identifies word pairs and triplets that occur near each other nonrandomly in a text collection, within a given window (a seven-word phrasal length in this study). Collocates are often abstractions, personal and place names, or habitual turns of phrase (Xiao and McEnery, 2006), and they are an important complement to keyness testing (Marcellino, 2019; Wenger et al., 2019).
5. The next step involved in-context viewing to qualitatively understand keywords (also called *concordance view*). RAND-Lex can automatically show how a given word is used in ten-word, in-context table entries to give a qualitative sense of usage.
6. The last step in the workflow was network visualization—using visualization software to create a visual representation of the network structure. Although RAND-Lex has built-in network visualization capabilities, we used Gephi to make higher quality graphics for publication purposes.[18]

Modeling Online Trolls and Understanding Their Context

Our labeled training (what we used to teach our ML model) was U.S. general election talk on Twitter from 2015 and 2016. It consisted of tweet text from known Russian trolls. Because this is a forensic task to identify bad actors (not bad tweets), we concatenated each troll's tweets into bundles with a maximum of 800 words.[19] By applying RAND-Lex to the text, we also generated stance vectors for each 800-word bundle, in essence converting the stance and style content of each bundle into a string of coordinates (a vector) that can be plotted in an *N*-dimensional space.[20] Our training data came from a Clemson University research team that applied one of four labels for each unique author according to qualitative analysis: politically right troll, left troll, conservative authentic, or liberal authentic (Linvill and Warren, 2018). Because our goal for this project was to study inauthentic behavior across the political spectrum, we focused our model-building on the task of binary classification, distinguishing only between trolls and non-trolls.

Visual inspection of the text associated with each type of account raised an immediate issue: Because the troll and authentic accounts had been acquired separately, there were formatting differences between them. Hyperlinks, for instance, had been added to each of our troll tweets that were not present in the original tweet. An artifact of how the data was formatted can allow an ML model to "cheat" (for example, an ML model could learn to predict cancer accurately not by looking at the scans it is being trained on but because technicians had stamped "CANCER" on all the images that were positive: The stamp is an effective tell, but it would not work in the wild with new, unlabeled data). To prevent this kind of data leakage, we cleaned the text of both types of accounts (removing hyperlinks, special characters, retweet markers, and usernames), so that the model could learn only from the text. We acknowledge that doing so might remove some legitimate discriminat-

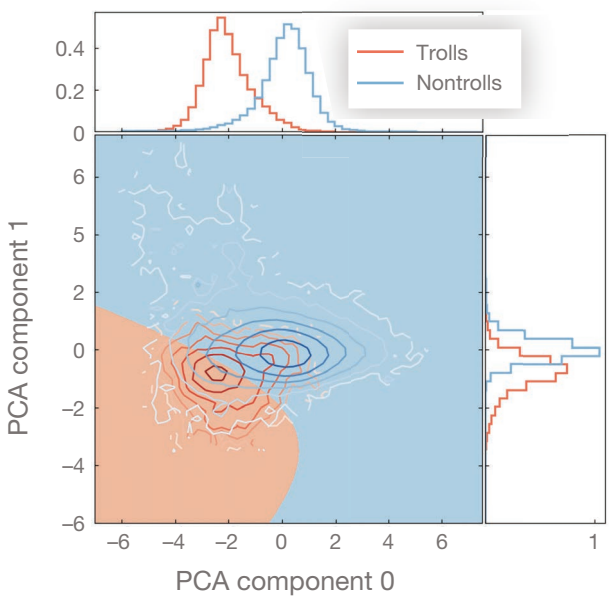
ing characteristics of trolls and non-trolls, but it also reduces the possibility of the model learning features that are nothing more than data artifacts. Manual inspection of the remaining text confirmed that our formatting procedure had not removed a significant amount of informative text.

The stance vectors, which are essentially arrays of numbers, are the most straightforward target for ML algorithms. Building off of previous RAND work (Marcellino, Cox, et al., 2020[21]), we first applied principal component analysis (PCA) to the stance vectors as a form of dimensionality reduction. PCA is a method of data reduction that builds new variables out of linear combinations of the original variables. The new variables are designed to capture a high degree of information that can be helpful in distinguishing populations in data sets.

In agreement with previous results, we saw clear distinctions between just the first two PCA components of the stances for trolls and non-trolls. We started with a very simple, foundational ML algorithm, a simple support vector machine (SVM) model with an order-2 polynomial kernel on the two most significant PCA variables as a test, which had reasonable success distinguishing between trolls and non-trolls (MCC = 0.678[22]). The decision boundary, and histograms for the first two PCA components, are shown in Figure A.1.

Once we saw that using the most important stance features could do a decent job classifying output as generated by a troll (or not), we sought to build a more powerful model that would yield better performance. Although clearly useful, stance is computed from a predefined dictionary, and we hypothesized that it might miss key features of trolls. We therefore turned to natural language processing (NLP) models, the most advanced of which rely on deep recurrent neural networks. Recent advances in NLP have yielded algorithms capable of a wide variety of tasks, from language translation to question answering, and we hoped that the relatively simple task of text classification would be feasible for an out-of-the-box model. We decided to use BERT (Devlin et al., 2018), a well-known model that uses semantic content to classify text. BERT is a multipurpose Deep Neural Network (DNN) NLP model, pretrained on text from Wikipedia and a corpus of about 11,000 publicly available books. Although it is recommended to fine-tune NLP models on sample text, we found good performance from the base BERT DNN model, perhaps because our data did not contain an abundance of specialized vocabulary.

FIGURE A.1
Stance PCA Reduction



SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This two-dimensional histogram (contours on a logarithmic scale) illustrates the first two PCA components of stance in our 2015–2016 data. Trolls are represented by the color orange; non-trolls are represented by the color blue. Our PCA components are zero-indexed, so the first component is on the x-axis and the second is on the y-axis. Histograms show the distribution along each axis and are normalized to unity. The SVM decision boundary is seen in the blue and orange background. The histogram clearly shows linguistic differences between trolls and non-trolls that motivate our next modeling steps.

In training our model, we were cognizant of the limits of our data. Our training tweets, for instance, were written in 2015 and 2016; the tweets of interest for this report were pulled from the early months of 2020. The parameters of the data pull were also slightly different for trolls and for non-trolls; they also differed for the 2015–2016 and the 2020 sets. We therefore anticipated that our model, although well trained to detect trolls in our training data, would not perform as well on the 2020 data, but we hoped that generic troll features would be present in both sets of data, allowing the model to generalize to unseen data.

Building a Hybrid Model for Improved Performance

From there, we employed a hybrid approach: Raw text extracted from tweets was analyzed by the BERT DNN, and stance features were extracted into a vector. We also were able to extract some limited pieces of metadata from our sample of known trolls: numbers of friends,[23] followers, and total retweets sent. Different combinations of the different features (DNN embeddings, stance vectors, and metadata) were then fed into a logistic regression classifier that delivered predictions for each author. We hypothesized that the stance features would be more readily translatable to the new data—trolls seeking to spread discord might use the same type of inflammatory language across conversations, even as the content of a conversation changes over time. We hoped the NLP model would learn to find subtler language features that would also translate well to the new data. In any case, we believed that a heterogeneous model that was reliant on multiple types of features would be more robust in dealing with changing data and therefore yield the best performance.

We split our data into training and evaluation sets, with a ratio of approximately 80:20 for training to testing. The ratio was not exactly 80:20 because many authors—especially those with lots of tweets—are represented across multiple rows, so drawing a random sample of rows for training or testing sets might result in the same author ending up in both sets. If this happened, the model might simply learn features of individual authors, instead of the generic troll characteristics we were hoping to understand. Therefore, we were careful to split the data so that no author appeared in both training and test sets. We used a single test set and trained the models without cross-validation because we saw no evidence of overfitting on our training set. It is possible that some of the troll authors were pseudonyms of the same human, in which case such data leakage is inevitable.

Testing Different Models

We sought to build a relatively straightforward model with results that would transfer well from training to test data. This led us to make several design choices: first, we used a miniature version of BERT called DistilBERT (Sanh et al., 2019), which has decreased computing requirements while maintaining most of the accuracy of the full BERT DNN model. Second, we used this DNN only to generate embeddings of our text samples—no fine-tuning was done. The embeddings, which were vectors of length 768, were taken directly from the last hidden layer in the DNN model after the text was passed through the neural network and fed into a logistic regression classifier. To reduce our computing requirements, we only passed the first 1,000 characters of each text sample to the DNN model, corresponding to about 160 words per sample. We also regularized our logistic regression classifier by setting an L2 penalty of 0.05, which we found was sufficient to reduce overfitting without adversely affecting performance. Our samples were also weighted in inverse proportion to the class frequency to address the imbalance between trolls and non-trolls. We used the Transformers and Scikit-learn (version 0.22.1) libraries for our implementations of the DNN and the logistic regression classifiers, respectively (Wolf et al., 2019; Pedregosa et al., 2011). We trained our models on several combinations of the data: the DNN embeddings alone, the stance vectors alone, the concatenated DNN and stance vectors, and the DNN vector concatenated with metadata values. The logistic regression classifiers return both predicted labels (troll or not-troll) and raw prediction scores for each author.

Our results, shown in Table A.1, support our hypothesis that a hybrid model would be the most powerful because it consistently yielded superior performance on both training data and evaluation data. The jump from an MCC score of 0.80 for the DNN model (just the semantic content) to 0.97 for the hybrid DNN + Stance is truly noteworthy because it involved no fine-tuning of the model. The stance taxonomy we used is generic, just as BERT is somewhat generic (trained on Wikipedia entries). However, this generic stance taxonomy clearly captures something important about the stylistic choices of Russian trolls. It is quite possible that a specialty DNN model might capture the same feature set, but this out-of-the-box capability is a kind of shortcut. Moreover, the stance vector only contains 110 variables, meaning quite a bit of information is being carried in a relatively

TABLE A.1
Model Performance

Model Description	Data Set	MCC Score	True Positive	False Positive	False Negative	True Negative
Support Vector Machine (SVM)	Training[a]	0.678	1,577	1,000	496	46,926
	Evaluation[a]	0.677	614	402	183	18,801
Deep Neural Network (DNN)	Training	0.788	13,757	6,970	441	259,729
	Evaluation	0.798	3,826	1,746	188	64,629
Stance	Training	0.783	13,650	7,016	548	259,683
	Evaluation	0.812	3,890	1,676	124	64,699
Hybrid: DNN + Stance	Training	0.967	14,155	936	43	265,763
	Evaluation	0.968	3,962	214	52	66,161
DNN + Metadata	Training	0.541	12,890	20,555	1,308	246,144
	Evaluation	0.584	3,678	4,898	336	61,477

SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This table illustrates performance of different text modeling approaches to classifying a Twitter author as a troll or non-troll. MCC score is the Matthews correlation coefficient, a measure that considers binary classification performance on unbalanced data (higher scores indicate better performance). The SVM model was trained on a slightly different training-and-testing splits, which is why the rows do not add to the same values. The best-performing model is the hybrid DNN + Stance model, which outperforms all the other models by a considerable amount. However, we use the DNN + Metadata model for the results throughout this paper, for reasons we explained in the text.

[a] Using only the first two principal components of the stance vector.

small vector; in this sense, generating stance vectors can be considered a sort of dimensionality reduction technique.

The rate of false positives was consistently higher than that of false negatives across all the models we considered, sometimes significantly so. In this sense, each of our models is like an overvigilant detective, which indicates their use would be limited in individual-level detection. We hypothesize that the reason for this discrepancy is that troll language is often quite similar to authentic hyperpartisan discourse, which presents difficulties in distinguishing between the two. This conclusion is consistent with previous work, which found that there was strong overlap between the language of trolls and hyperpartisan non-trolls.

Because the DNN + Metadata model only added three extra features to the DNN-alone model, we expected their performances to be similar. Surprisingly, the DNN + Metadata model performed significantly worse than the DNN model alone, mostly because of an increased rate of false positives. It is odd for a model with more input features to result in worse performance; if the new features simply had little predictive power, the model would learn to ignore them and the performance should be unchanged. It is possible that the addition of the extra features was causing the model to overfit to the training data, which might result in poor performance. But the DNN + Metadata model achieved better performance on the evaluation data than on the training data, and we used a relatively strong L2 penalty for its logistic regression, both of which indicated that overfitting was not the cause.

What seems most likely, instead, is that the text-only models might have latched onto certain keywords or phrases indicative of trolls, and the inclusion of non-text data acted as further regularization by providing a non–text-based feature to the model. If this were the case, we would expect the model that uses the more-heterogeneous data to generalize better to unseen data than a text-only model, insofar as metadata is actually indicative of trolls. To investigate, we compared the distribution of raw scores returned by each model on the test set and on the 2020 data set. In agreement with our hypothesis, the DNN-alone, Hybrid DNN + Stance, and Stance-alone models had dramatically different raw score distributions on the two data sets. The distribution of scores from the text-only models in particular was much

wider, yielding values that are far more extreme than in the 2016 data. Meanwhile, the DNN + Metadata model returned relatively similar distributions on the two data sets. Our expectation was that the distributions should not change wildly between the two data sets; therefore, we interpreted this result as an indication that the DNN + Metadata model generalized best to the unseen data. This was the main reason we chose to use this model for the rest of the analysis. Despite the higher false-positive rate in our training and evaluation data sets, the DNN + Metadata model predicted fewer accounts in our 2020 data set to be trolls: about 10 percent, compared with the DNN-alone model prediction of about 26 percent, and far less than the Stance-alone prediction of 54 percent and the hybrid prediction of 47 percent.

Further investigation pointed to a possible reason why some of our models did not seem to generalize well to the 2020 data. The training data set we used contained all the tweets for each author, but our 2020 data consisted only of tweets that mentioned election-related terms. Although the two sets of text were processed identically, the content was markedly different, which became clear when observing the distribution of different stance variables. Future research might be able to overcome this problem with a more targeted data strategy; for our purposes, the DNN + Metadata model appeared to generalize adequately.

We concluded that Stance-alone and DNN-alone models, although powerful, might be somewhat brittle and require text more precisely formatted than that to which we had access. Therefore, we used the DNN + Metadata results to conduct our analysis and draw our conclusions. An examination of analogous results based on our other models ended up pointing toward virtually identical conclusions, so we do not report those results here.

Applying Our Model to 2020 Data

We passed the text from each author in our 2020 data set through our DistilBERT model and then applied our trained logistic regression classifiers. Because the various models classified an overwhelming majority of the 2020 authors as non-trolls, we were most interested in the raw scores, instead of the predicted label. We hypothesized that the accounts with the highest scores (regardless of whether that score was above or below the threshold of troll or not-troll) would be the most-likely inauthentic accounts—or at the very least, the accounts with the behavior that was most troll-like. We anticipated that the overwhelming majority of accounts were authentic, so we focused our analysis on the accounts with raw scores in the 90th to 95th percentiles.

The authors with the highest raw scores next underwent human qualitative analysis (as described in the section on election interference). Although there appeared to be a significant number of false positives during this process—an account that was verified by Twitter, for example—we also found several suspicious accounts within our sample that were promising troll candidates. The most striking feature of these accounts was their seemingly inauthentic interactions with their followers and friends; accounts that had thousands of followers had few or no likes, retweets, or responses to any of their tweets. Often, they had a relatively similar numbers of followers and friends, often nearly exactly so. We suspected that troll accounts might be artificially boosting their popularity with bot followers to appear more authentic. The other feature that was conspicuous among these candidates was their constant retweeting; authentic accounts also retweet often, but the lack of original tweets from our candidate accounts was notable. A high volume of retweets makes the task of discriminating between authentic and inauthentic accounts difficult, but we believed that at least some of the accounts we examined might have been inauthentic: One account, for instance, tweeted or retweeted constantly throughout April—2,846 times over one week we observed, about once every four minutes—then ceased completely on April 27 and at the time of this writing, has not tweeted since.

With these key insights, we turned to analyzing the metadata—interactions and behavior beyond text content—of the Twitter accounts we observed. We first sought to understand some of the immediately apparent features in the data, such as prominent peaks in the histogram of friends at 2,000 (in the 2015–2016 data set) and 5,000 (all data sets). We learned that these accounts were pushing up against

limitations that Twitter places on the number of accounts that can be followed: As of 2020, the limit is 5,000 friends unless the account has a similarly large number of followers, in which case there are no limits to the number of friends an account can have. We broadly refer to these accounts with high numbers of friends as superconnectors.

Superconnectors

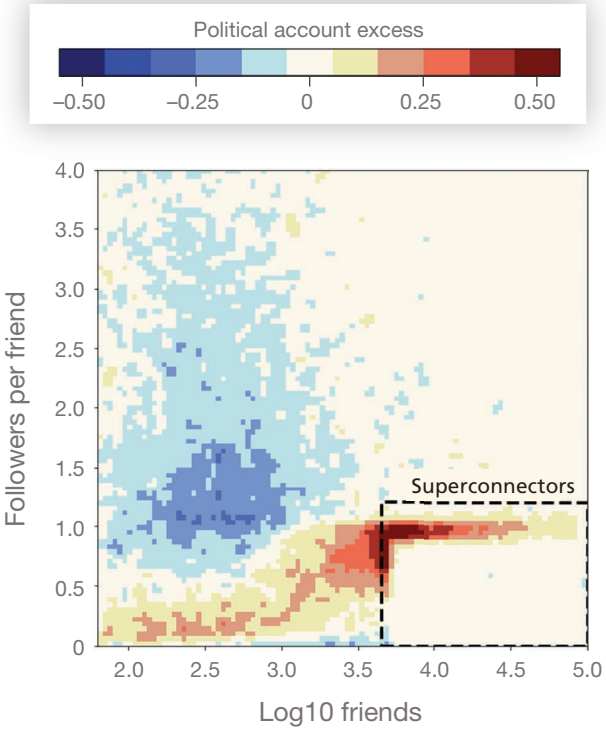
Superconnectors are not necessarily suspicious in and of themselves; our nonpolitical sample of Twitter users also had a population of these accounts.[24] However, we found that the prevalence of superconnectors varied widely among different communities. Only about 2.5 percent of nonpolitical accounts had more than 4,500 friends and less than 1.2 followers per friend;[25] the same fraction in our political sample was about 10 percent. The relative account frequency between the two data sets is displayed Figure A.2, and a breakdown by community is shown in Figure A.3.

One plausible explanation for this discrepancy could be the nature of the conversation taking place; perhaps political discussions naturally lead to more engagement than nonpolitical ones. However, the observed excess is only apparent in three of our identified communities (see Figure A.2) and appears to be almost completely absent elsewhere. This suggests that the excess is not natural, which might lead to a skewing of the conversation in these communities.

Participation in Discourse

The next step in our analysis was to cross-reference the metadata already described and the results of our linguistic modeling with the community membership derived in the previous section. Because of some technical issues with the matching of accounts,[26] the number of accounts that we were able to cross-reference was slightly decreased (which is why the values in Table A.2, which we will discuss later, do not match exactly with those shown for the communities previously). We discovered that only three communities seemed to contribute to the observed excess in Figure A.2; the others showed only moderate excesses and deficits. Moreover, the accounts with highest friend and follower counts in our data set belonged disproportionately to a single community (Pro-Trump); the other communities (such as Pro-Biden and Pro-Sanders) that we investigated showed little evidence of the same type of clustering.

FIGURE A.2
Excess of Superconnector Political Accounts



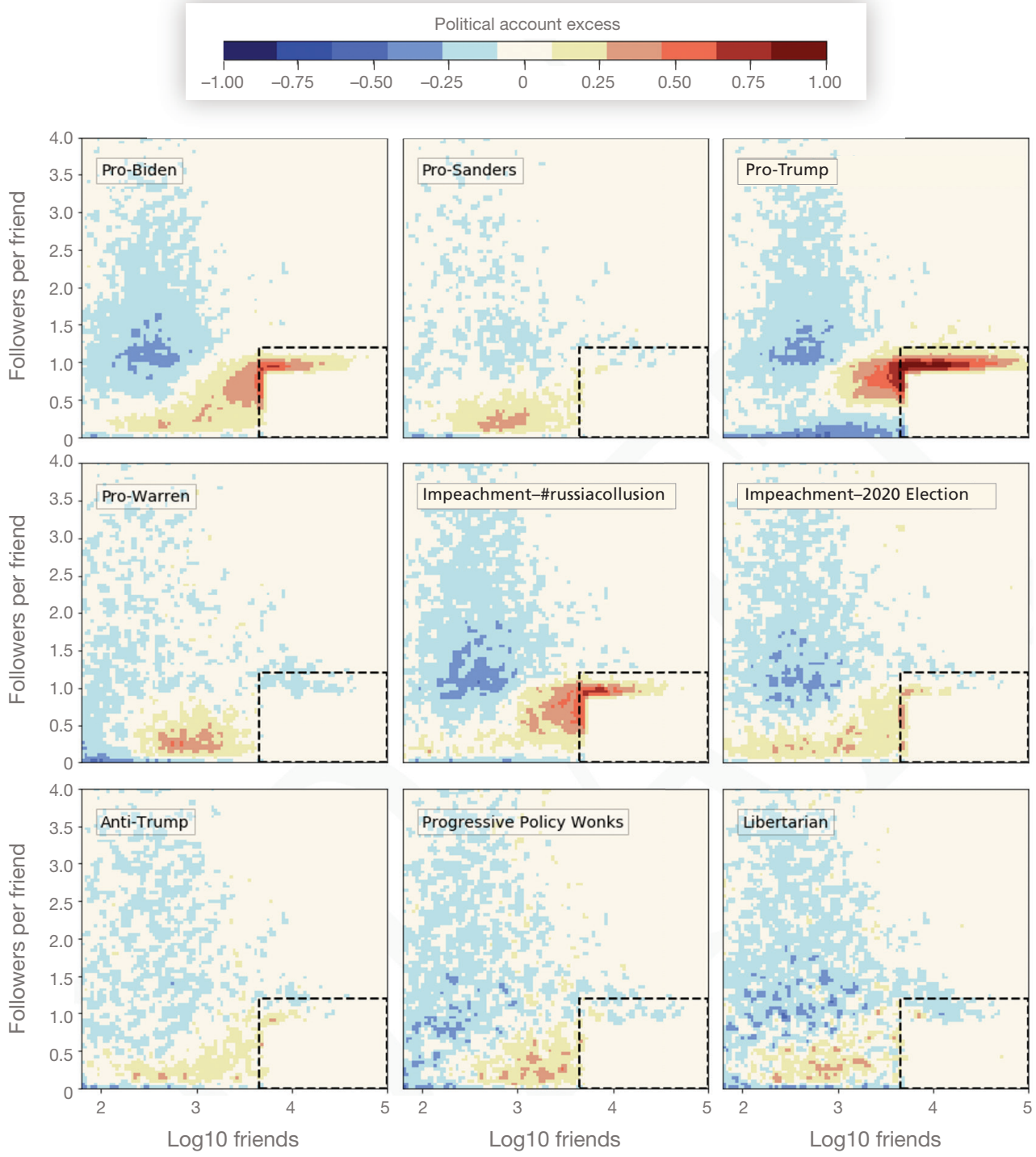
SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This figure illustrates the relative frequency of political accounts compared with nonpolitical accounts, in friends or followers space. The relative frequency metric is computed in each pixel as $(P_c - N)/\sqrt{(P_c + N)}$, where P_c is the density of political accounts in community C, and N is the density of nonpolitical accounts. Because there are different numbers of accounts in the political and nonpolitical communities, we normalize the density by dividing the raw counts in each pixel by the total number of accounts in each sample. Therefore, this is an "apples-to-apples" comparison of where political accounts are more frequent (in red), and where nonpolitical accounts are more frequent (in blue). The data has been smoothed with a Gaussian filter to reduce noise. The box in the lower right corner shows a cutoff for superconnectors (more than 4,500 friends; fewer than 1.2 followers per friend). The sharp boundary inside this box, at 5,000 friends and approximately 0.9 followers per friend, are the result of Twitter restrictions on the number of accounts that might be followed.

Another way of looking at the data is shown in Figure A.4, which displays the fraction of accounts that belong to each of our three most popular communities. A surprisingly high fraction of superconnectors belong to the Pro-Trump community, and the fraction increases as the accounts become more connected. The total number of accounts also

FIGURE A.3

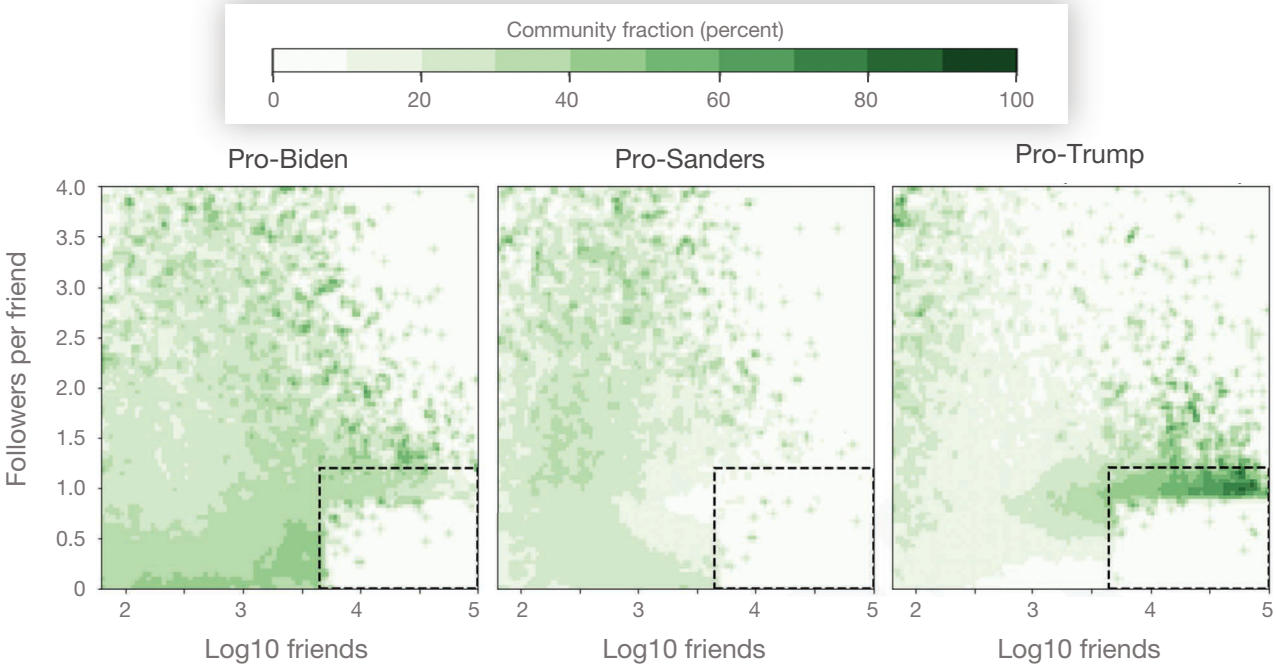
Superconnectors Clustered in Specific Communities



SOURCE: RAND analysis of Twitter data, 2020.

NOTES: Each panel is a representation of one of the nine communities with the highest membership in the friend or follower space, relative to our sample of nonpolitical accounts. The box in the lower right-hand corner represents the same superconnector definition as in Figure A.2.

FIGURE A.4
Community Membership Fraction



SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This figure illustrates the fraction of accounts in each bin belonging to each of the three largest communities identified. The Pro-Biden and Pro-Sanders accounts appear to be relatively uniform in their distribution; among accounts with more than 10,000 friends, the Pro-Trump community is dominant. As in the previous figures, the box in the lower right-hand corner signifies the superconnectors.

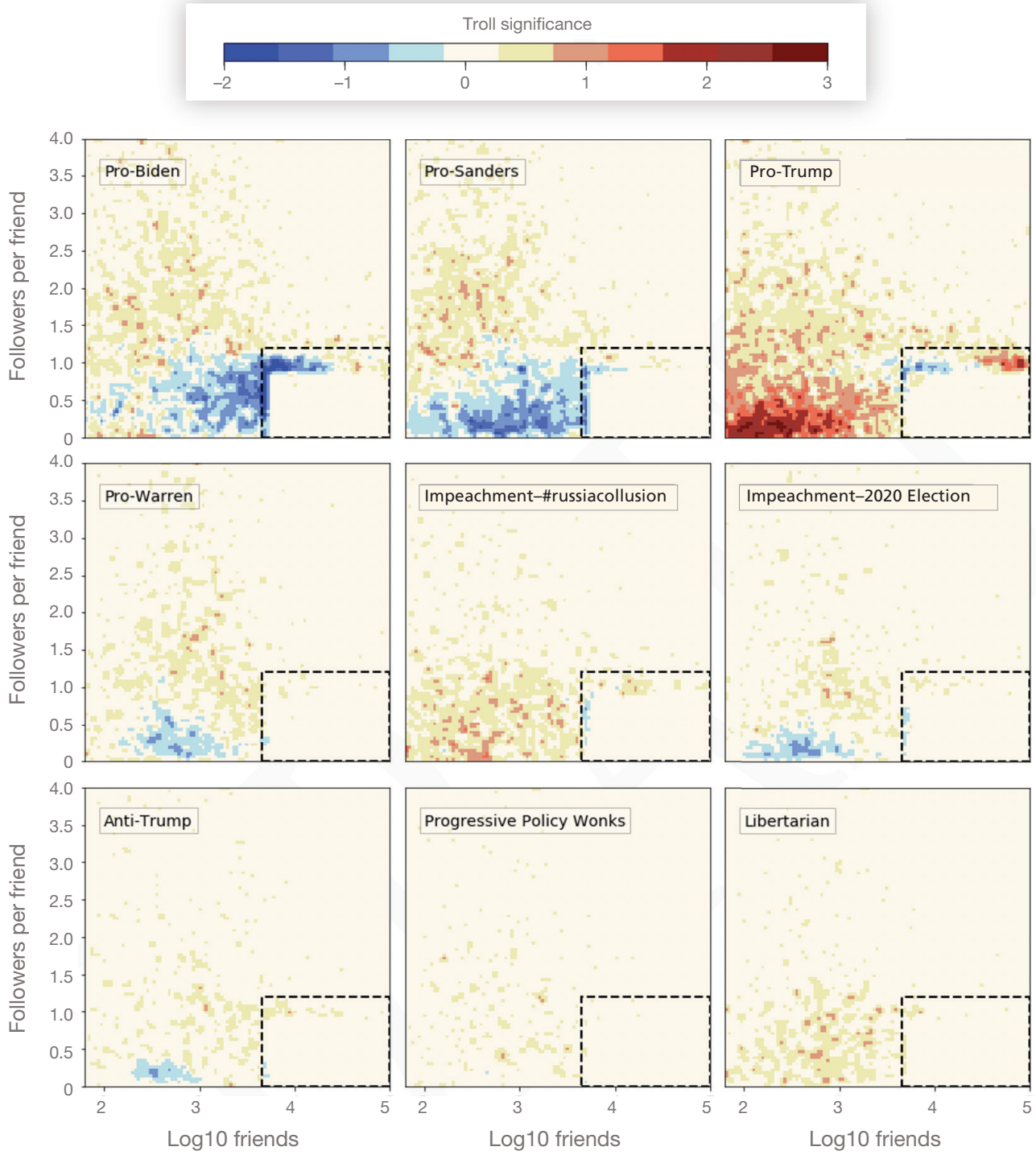
decreases as connectedness increases, but the pattern is significant: although the Pro-Trump community makes up a little less than 20 percent of the total number of accounts we considered, it makes up more than 80 percent of accounts in some regions of the superconnector space. In other words, a random political Twitter account that has more than 50,000 friends has a very high chance of belonging to the Pro-Trump community.

We found a similar result when we broke down the results of our linguistic troll model by community. Again, the accounts with the highest troll ratings landed disproportionately in the Pro-Trump community; the other communities showed only modest excesses (or, more often, deficits) of accounts with high troll ratings (Figure A.5). The pattern of placement for these accounts was somewhat different than that of the superconnectors: Most of the accounts with the highest troll ratings had modest friend and follower numbers, though the accounts with more than 20,000 friends also showed a moderate excess.

The correlation between these two markers of possible inauthentic accounts is clear when comparing the results by community (Figure A.6). Together, the data suggest that there was a possible relationship between the linguistic model results and the observed metadata excess, although the exact nature of that relationship remains somewhat unclear. A summary of the data is also available in Table A.2.

FIGURE A.5

High Troll Scores by Community

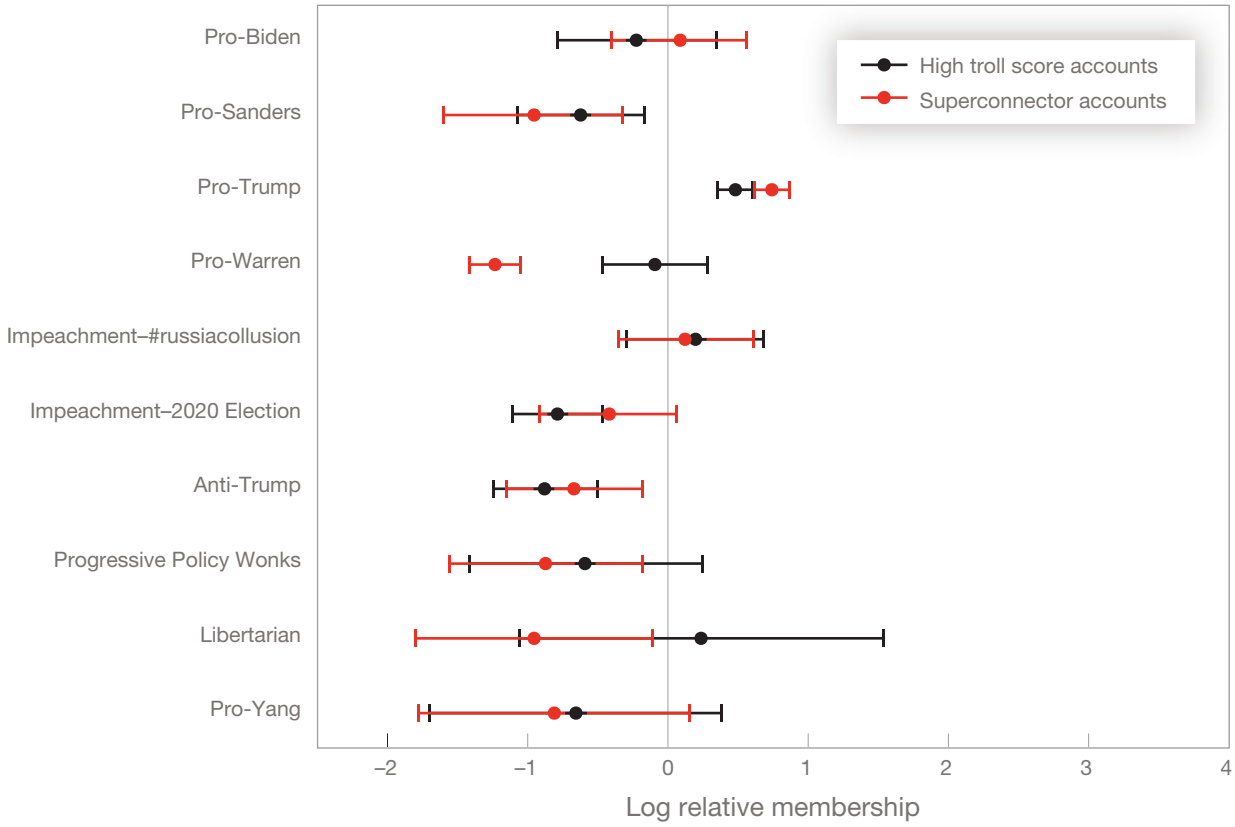


SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This figure depicts troll significance by community among our sample of election-related accounts. The troll significance in bin *i* is defined as $〚TS〛_i=(〚(N〛_i^{90}-0.1\times N_i))/\sqrt{(0.1\times N_i)}$, where N_i is the number of accounts in that bin, and N_i^{90} is the number of accounts in the bin with troll score above the 90th percentile. Assuming the number of accounts is Poisson-distributed, if N_i is not O(1), this metric is then roughly the number of standard deviations above the mean. Were troll scores evenly distributed across bins and communities, we would expect to see small fluctuations about 0 significance across the plot; instead, there is a clear excess in the Pro-Trump community. The excess is observed mostly at low friend and follower counts, though there is also a smaller excess at very high friend counts. Conversely, we see a deficit in the Pro-Biden troll scores for superconnectors.

FIGURE A.6

Relative Community Representation of Trolls, Superconnectors



SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This figure illustrates relative community membership by accounts with high troll scores and near maximal friend and follower numbers. The log relative membership score is calculated by dividing the fraction of suspicious accounts in each community—either those with troll scores above than the 95th percentile, or with more than 4,500 friends and fewer than 1.2 followers per friend—by the expected fraction and taking the natural logarithm. If troll scores were randomly assigned, one would expect about 5 percent of accounts to have troll scores above the 95th percentile. Instead, we see that more than 8 percent of accounts in the Pro-Trump community have troll scores above than the 95th percentile, so the third black point lies to the right of the dashed line. The error bars are computed by assuming a Poisson distribution of accounts in each community, so there is more uncertainty in the relative membership for smaller communities. We note a strong correlation between the two markers of possible inauthentic behavior: The Pro-Sanders community, for example, is underrepresented by both measures while the Pro-Trump community is significantly overrepresented by both.

TABLE A.2

Distribution of Suspicious Accounts by Community

Community	Accounts	Near Maximal Accounts [%]	Trolls in Top 95th Percentile [%]
Nonpolitical (comparison group)	*90,720*	*2.44*	*Not applicable*
Pro-Biden	159,576	**10.96**	4.00
Pro-Sanders	91,241	3.90	2.68
Pro-Trump	87,712	**21.25**	**8.10**
Pro-Warren	26,454	2.91	4.50
Impeachment–#russiacollusion	23,858	**11.40**	**6.00**
Impeachment–2020 Election	16,631	6.48	2.28
Anti-Trump	13,647	5.01	2.01
Progressive Policy Wonks	7,359	4.38	2.77
Libertarian	4,832	3.83	**6.31**
Pro-Yang	4,478	4.49	2.57

SOURCE: RAND analysis of Twitter data, 2020.

NOTES: This table illustrates markers of potentially inauthentic behavior by Twitter community membership. The three highest values in each column are bolded. Superconnectors are those with more than 4,500 friends and fewer than 1.2 followers per friend. Given our sample of nonpolitical accounts, we expect that a small percentage of accounts will lie within this boundary, but a few communities—the Pro-Biden, Pro-Trump, and Impeachment–#russiacollusion communities—are significantly overrepresented in this region. The column farthest to the right is the percentage of accounts in each community that has a raw troll score greater than the 95th percentile for the entire sample; if troll scores were randomly distributed, we would expect most communities to have 5 percent of accounts fall into this category. However, some communities (particularly the Pro-Trump community) are significantly above this 5-percent baseline.

Appendix B. Top-Trending Hashtags

The top-trending hashtags were identified by ordering the hashtags according to their largest one-day and one-week increases in usage. The lists of the top 200 daily trending hashtags and top 200 weekly trending hashtags were combined. (Setting the list at a length of 200 was arbitrary.) Because most hashtags on one list were also on the other, the combined list had 234 unique hashtags. These are listed in order of total usage.

Hashtag	Total Usage
#trump	20,103
#trump2020	14,228
#votebymail	11,654
#russiancollusion	9,904
#2020	9,546
#dems	9,124
#trishregan	8,747
#icymi	8,006
#iowa	7,759
#kag	7,738
#impeachment	7,148
#russia	7,113
#maga	6,340
#wtpteam	6,279
#coronavirus	6,278
#joebiden	5,627
#demdebate	5,614
#berniebeatstrump	5,584
#notmeus	5,107
#onevoice1	4,911
#trumpisarussianasset	4,838
#biden2020	4,737
#russianinterference	4,704
#clinton	4,360
#supertuesday	4,358
#trumprussia	4,224
#biden	4,222
#kag2020	4,102
#trumppressbriefing	4,020
#votebluenomatterwho	4,017
#bernie2020	3,929
#qanon	3,455
#berniesanders	3,451
#hd28	3,345
#coronavirusliar	3,244
#covid19	3,184
#iacaucus	3,129
#donaldtrump	2,984
#treason	2,900
#bernie	2,809
#complicitgop	2,803
#losewithbiden	2,677
#democrats	2,652
#wwg1wga	2,593
#trump2020landslide	2,372
#iranattack	2,359
#yanggang	2,295
#election2020	2,201
#beafraid	2,160
#google-and-the-gang's	2,158
#dementia	2,131
#onlybernie	2,015
#mog	1,957
#bidenscognitivedecline	1,758
#voteforbernie	1,666
#presidenttrump	1,596
#sotu	1,589
#fakenews	1,589
#moscowmitch	1,585
#bluewave2020	1,551

Hashtag	Total Usage
#2020election	1,475
#gop	1,467
#sanders	1,446
#trumpownseverydeath	1,437
#saturdaymotivation	1,426
#covid2019	1,419
#abuseofpower	1,406
#berniewon	1,400
#aoc	1,387
#miniaoc	1,372
#twgrp	1,368
#dropoutbernie	1,342
#resist	1,285
#generalstrike	1,277
#amjoy	1,273
#iowacaucuses	1,271
#putin	1,268
#trump2020landslidevictory	1,256
#voteandlive	1,208
#preventfraud	1,208
#stayhome	1,202
#notdying4wallstreet	1,198
#democraticdebate	1,196
#trump's	1,178
#covid	1,124
#joementum	1,119
#americaneedsyang	1,105
#impeachmenttrial	1,099
#traitor	1,091
#yangbeatstrump	1,087
#nhprimary	1,076
#presidentcheat	1,075
#mayorcheat	1,074
#maga2020	1,064
#impeachbarrnow	1,054

Hashtag	Total Usage
#creepyjoebiden	1,049
#trumpslushfund	1,047
#voteblue2020	1,020
#foxnews	1,009
#bidenbeatstrump	1,007
#protectourcare	1,001
#usa	992
#nevertrump	986
#theresistance	985
#bidenlosestotrump	984
#neverbiden	977
#wtp2020	959
#bernieknew	955
#nevadacaucus	930
#voteredtosaveamerica	921
#preexistingconditions	921
#maddow	917
#caucusforbernie	909
#trumpistheworstpresidentever	908
#iowacaucas	890
#barr	872
#impeachpelosi	848
#votebluetosaveamerica	847
#neverbernie	841
#susancollins	839
#cnn	837
#wisconsin	836
#teamjoe	820
#trumpfearsbernie	809
#tqphpoll	800
#china	794
#southcarolinaprimary	783
#qanon2020	783
#qproof	782
#iowacaucus	777

Hashtag	Total Usage
#qanon2018	774
#penceknew	769
#presidentpelosi	765
#hollywood	763
#joebiden2020	757
#impeachmenthoax	756
#dncrigging	743
#cancelbiden	738
#levspeaks	737
#trumphotel	732
#coronaviruspandemic	726
#christian	724
#onevoice1lgbtq	708
#votebluenomatterwho2020	705
#parnasdocs	703
#medicareforall	703
#breaking	701
#trumpbeatsbloomberg	686
#sc2020	684
#texas	679
#smartnews	673
#iowacaucasdisaster	667
#klobuchar	650
#voteredtosaveamerica2020	640
#supertuesday2020	637
#hypocrisy	637
#demsaffairwithayatollahs	636
#new	634
#warren2020	631
#socialsecurity	626
#defendourdemocracy	625
#sotswamp	624
#walkaway	623
#mighty200	621
#republicans	621

Hashtag	Total Usage
#thursdaywisdom	611
#thursdayvibes	611
#supertuesday3	607
#kag2020landslidevictory	603
#chinacollusion	603
#throwbackthursday	602
#nevada	599
#demexit	596
#votered	595
#superbowl	593
#kaga2020	592
#trump2020nowmorethanever	582
#coronvirus	579
#potus	578
#witnessesnow	577
#democratsaredestroyingamerica	576
#voteblue	574
#obummers	572
#spotlight	572
#mtpdaily	568
#buttigeig	566
#alandershowitz	563
#tuesdaythoughts	558
#votethemout	557
#unitewithbernie	554
#draintheswamp	549
#americafirst	544
#newhampshire	537
#politics	524
#muellerreport	523
#newhampshireprimary	517
#berniebros	513
#shahidvspelosi	499
#realtalk	499
#votebluetosavetheplanet	498

Hashtag	Total Usage
#uniteanddefend	498
#ibelievetarareade	497
#wtp271	495
#133	481
#joementia	461
#ridinwithbiden	461
#voterfraud	444
#ruststatebelt	435
#toledo	435
#magarollercoaster	431
#trumpslump	430
#mtp	428
#joebiden4china	427
#michiganprimary	426
#michiganvotes	417
#nhprimary2020	414
#thesepeoplearestupid	401
#supertuesday2	400
#trumps	399
#cult45	395
#nv3	395
#danrodimer	390
#nevada3	390
#secureourborders	384
#iowacaucusdisaster	380
#coronavid19	369
#russiahoax	367
#trumprallynh	351
#madking	346
#firefauci	341
#stopvoterfraud	335
#hydroxychloriquine	332
#democratcorruption	331
#unendorsebiden	331

Notes

[1] In this report, we use the term *troll* to refer to fake personas engaged in political manipulation as part of a malign influence campaign, not the broad vernacular meaning of someone on the internet who acts provocatively to elicit anger and frustration.

[2] RAND-Lex is RAND's proprietary text and social media analysis software platform: a scalable cloud-based analytics suite with network analytics and visualizations, a variety of text-mining methods, and ML approaches. For example, see Kavanagh et al., 2019.

[3] RAND-Lex uses a version of Louvain modularity (see Blondel et al., 2008).

[4] We looked at the relative overrepresentation and underrepresentation of both superconnector accounts and the troll score returned by our lexical model. As a control, we found the average for superconnectors in nonpolitical Twitter discourse was 2.44 percent (generally, 2.5 percent is a baseline for this kind of account). For the troll scores, we care about the relative abundance of high-scoring accounts, so we compare the percentage of accounts with scores above the 95th percentile of scores for the overall population. (We found that the 95th percentile was a good balance between volume of accounts and specificity.) If high-scoring troll accounts were evenly distributed (or if our model returned random values), we would expect this value to be approximately 5 percent for each community. We discuss this in more detail in Appendix A.

[5] By *politically right-leaning*, we mean accounts that were in the Pro-Trump or Libertarian communities and shared content that promoted the GOP and President Trump and that disparaged Democrats, "leftists," "communists," and "socialists." By *politically left-leaning*, we mean accounts that were in either the Pro-Biden or Impeachment–#russiacollusion communities and that shared content promoting progressive policies and candidates and condemning President Trump, the GOP, and conservatives, "fascists," and "Nazis."

[6] We acknowledge a meaningful difference between trolls (human-run inauthentic accounts) and bots (automated inauthentic accounts), but in this study we did not attempt to determine whether the suspicious accounts we found were human-run or automated. For purposes of this study, the specifics of automation levels were less important than identifying and describing suspicious activity: what was being done, and who was being targeted.

[7] By *semantic content*, we mean informational content of text—in essence, the words and their relationship to other words, captured by a powerful modeling technique known as word embeddings. By *linguistic style*, we mean the stance or attitudinal content of text, captured through a set of expert dictionaries of rhetorical choices. By *metadata*, we mean data associated with the text: that account's number of friends, followers, and number of retweets. All these are discussed in detail in Appendix A.

[8] To capture style, we used a taxonomy of the rhetorical functions of language developed at Carnegie Mellon University (Beigman Klebanov et al., 2019; Ringler, Beigman Klebanov, and Kaufer, 2018).

[9] Performance was measured by Matthews Correlation Coefficient (MCC), a widely used measure of the quality of a binary predictor model. MCC is calculated out of 1.00. (We converted these figures to simple percentages in the report text to be more interpretable to general readers.) ML model performance and performance measures are discussed in detail in Table A.1.

[10] Our model of choice had an MCC of 0.58 on our evaluation set, compared with a coefficient of 0.97 for our best-performing model. More discussion of our modeling choices can be found in Appendix A.

[11] This research looked for Russian trolls (i.e., state-sponsored social media accounts masquerading as authentic members of the U.S. polity).

[12] A particular tactic for politically right-leaning troll accounts was to post content condemning Black Lives Matter protests as leftist violence, punctuated with content showing criminal or violent behavior involving black Americans outside a political context. We think this content is meant to foment division along perceived racial lines.

[13] Because of logistical difficulties, the official Nevada results did not come in until about two days later, at which point #berniewon did not trend, although other pro-Sanders hashtags did.

[14] Russia's information competition framework and reflexive control theory are explained in more detail in Posard et al., 2020.

[15] There were 635,000 screen names associated with these 630,391 accounts because some users changed their screen names, one as many as 11 times during the course of the data pull.

[16] We chose to examine the top ten largest communities because communities quickly scaled down in size and dropped precipitously after the tenth community.

[17] Every retweet and "@so-and-so" on Twitter is a mention of another account. A mentions network involves drawing lines between accounts for every mention. Our algorithm resolves this large tangle of mentions into membership based on the preponderance of interactions.

[18] Gephi is a popular network visualization tool. For more detail, see Gephi, undated.

[19] More precisely, the bundles had a maximum of 800 tokens; these were identified using the Tweet Tokenizer module, which is part of the NLTK library. (NLTK is a Python library for text analysis.)

[20] Because individual tweets are so short, we concatenated tweets from each author into 800-word bundles to get a larger unit of analysis.

[21] This research effort looked for Russian trolls (i.e., state-sponsored social media accounts masquerading as authentic members of the U.S. polity).

[22] MCC is the Matthews correlation coefficient, a metric for binary classification that performs well on imbalanced data sets. For our purposes, it ranges from 0 (no better than random chance) to 1 (perfect accuracy).

[23] In the Twitter lexicon, the *friends* of an account are the people being followed by that account.

[24] To better understand how our general election 2020 data set (explicitly political) might differ from Twitter talk in general that was nonpolitical, we gathered a comparison corpus of sports,

games, and movie talk from the same period, using the following search terms: "MLB OR #MLB OR 'Baseball' OR Movies OR #Movies OR Movie OR Film OR #Film OR Cinema OR Gaming OR Gamers OR Gaming OR #Gaming."

[25] This captured most of the behavior that we were interested in near the boundary that Twitter enforces for each account, but the results are not strongly dependent on the exact value.

[26] Some of our data were indexed by the user's Twitter handle; some were indexed by user identification.

References

Beigman Klebanov, Beata, Chaitanya Ramineni, David Kaufer, Paul Yeoh, and Suguru Ishizaki, "Advancing the Validity Argument for Standardized Writing Tests Using Quantitative Rhetorical Analysis," *Language Testing*, Vol. 36, No. 1, 2019, pp. 125–144.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre, "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment*, Vol. 2008, No. 10, 2008.

Bodine-Baron, Elizabeth, Todd C. Helmus, Madeline Magnuson, and Zev Winkelman, *Examining ISIS Support and Opposition Networks on Twitter*, Santa Monica, Calif.: RAND Corporation, RR-1328-RC, 2016. As of July 31, 2020: https://www.rand.org/pubs/research_reports/RR1328.html

Brandwatch, homepage, undated. As of September 17, 2020: https://www.brandwatch.com

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, BERT: P*re-Training of Deep Bidirectional Transformers for Language Understanding*, arXiv.org, October 11, 2018.

Frenkel, Sheera, and Julian E. Barnes, "Russians Again Targeting Americans with Disinformation, Facebook and Twitter Say," *New York Times*, September 1, 2020.

Gephi, homepage, undated. As of September 17, 2020: https://gephi.org

Helmus, Todd C., James V. Marrone, and Marek N. Posard, *Russian Propaganda Hits Its Mark: Experimentally Testing the Impact of Russian Propaganda and Counter-Interventions*, Santa Monica, Calif.: RAND Corporation, RR-A704-3, forthcoming.

Irving, Doug, "Big Data, Big Questions," *RAND Blog*, October 16, 2017. As of September 17, 2020: https://www.rand.org/blog/rand-review/2017/10/big-data-big-questions.html

Kavanagh, Jennifer, William Marcellino, Jonathan S. Blake, Shawn Smith, Steven Davenport, and Mahlet G. Tebeka, *News in a Digital Age: Comparing the Presentation of News Information over Time and Across Media Platforms*, Santa Monica, Calif.: RAND Corporation, RR-2960-RC, 2019. As of July 31, 2020: https://www.rand.org/pubs/research_reports/RR2960.html

Linvill, Darren L., and Patrick L. Warren, *Troll Factories: The Internet Research Agency and State-Sponsored Agenda Building*, Leipzig, Germany: Resource Centre on Media Freedom in Europe, July 2018. As of November 22, 2019: https://www.rcmediafreedom.eu/Publications/Academic-sources/Troll-Factories-The-Internet-Research-Agency-and-State-Sponsored-Agenda-Building

Marcellino, William, "Seniority in Writing Studies: A Corpus Analysis," *Journal of Writing Analytics*, Vol. 3, 2019, pp. 183–205. As of September 17, 2020: https://wac.colostate.edu/docs/jwa/vol3/marcellino.pdf

Marcellino, William, Kate Cox, Katerina Galai, Linda Slapakova, Amber Jaycocks, and Ruth Harris, *Human-Machine Detection of Online-Based Malign Information*, Santa Monica, Calif.: RAND Corporation, RR-A519-1, 2020. As of July 20, 2020: https://www.rand.org/pubs/research_reports/RRA519-1.html

Marcellino, William, Krystyna Marcinek, Stephanie Pezard, and Miriam Matthews, *Detecting Malign or Subversive Information Efforts over Social Media: Scalable Analytics for Early Warning*, Santa Monica, Calif.: RAND Corporation, RR-4192-EUCOM, 2020. As of August 18, 2020: https://www.rand.org/pubs/research_reports/RR4192.html

Office of the Director of National Intelligence, "Statement by NCSC Director William Evanina: Election Threat Update for the American Public," press release, August 7, 2020. As of September 17, 2020: https://www.dni.gov/index.php/newsroom/press-releases/item/2139-statement-by-ncsc-director-william-evanina-election-threat-update-for-the-american-public

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al., "Scikit-Learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825–2830.

Posard, Marek N., Marta Kepe, Hilary Reininger, James V. Marrone, Todd C. Helmus, and Jordan R. Reimer, *From Consensus to Conflict: Understanding Foreign Measures Targeting U.S. Elections*, Santa Monica, Calif.: RAND Corporation, RR-A704-1, 2020. As of October 1, 2020: https://www.rand.org/pubs/research_reports/RRA704-1.html

Ringler, Hannah, Beata Beigman Klebanov, and David Kaufer, "Placing Writing Tasks in Local and Global Contexts: The Case of Argumentative Writing," *Journal of Writing Analytics*, Vol. 2, 2018, pp. 34–77.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf, *DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter*, arXiv.org, October 2, 2019.

Scott, Mike, *WordSmith Tools Manual*, Oxford, UK: Oxford University Press, 1996.

Scott, Mike, "PC Analysis of Key Words—and Key Key Words," *System*, Vol. 25, No. 2, 1997, pp. 233–245.

Scott, Mike, *WordSmith Tools Manual*, Version 6, Stroud, Gloucestershire, UK: Lexical Analysis Software, 2015.

Select Committee on Intelligence of the United States Senate, *Report on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*, Vol. 1: *Russian Efforts Against Election Infrastructure with Additional Views*, 116th Congress, Report 116-XX, 2019. As of July 29, 2020: https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf

Select Committee on Intelligence of the United States Senate, *(U) Report on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election*, Vol. 2: *Russia's Use of Social Media with Additional Views*, 116th Congress, Report 116-XX, undated. As of July 29, 2020: https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf

United States v. Internet Research Agency LLC, 2018 W.L. 914777, 2018. As of July 29, 2020: https://www.justice.gov/file/1035477/download

U.S. Department of Justice, U.S. Attorney's Office for the Southern District of New York, "Jeffrey Epstein Charged in Manhattan Federal Court with Sex Trafficking of Minors," July 8, 2019. As of August 5, 2020: https://www.justice.gov/usao-sdny/pr/jeffrey-epstein-charged-manhattan-federal-court-sex-trafficking-minors

Wenger, Jennie W., Heather Krull, Elizabeth Bodine-Baron, Eric V. Larson, Joshua Mendelsohn, Tepring Piquado, and Christine Anne Vaughan, *Social Media and the Army: Implications for Outreach and Recruiting,* Santa Monica, Calif.: RAND Corporation, RR-2686-A, 2019. As of September 17, 2020: https://www.rand.org/pubs/research_reports/RR2686.html

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., *HuggingFace's Transformers: State-of-the-art Natural Language Processing*, arXiv.org, October 9, 2019.

Xiao, Richard, and Tony McEnery, "Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective," *Applied Linguistics*, Vol. 27, No. 1, 2006, pp. 103–129.

About This Report

Foreign election interference is a serious threat to U.S. democratic processes, something that became visible and received public attention in the wake of the 2016 U.S. general election. In the aftermath of that election, it became clear that agents acting on behalf of the Russian government went online and engaged in a very sophisticated malign information effort meant to sow chaos and inflame partisan divides in the U.S. electorate (Marcellino, Cox, et al., 2020). Because of the seriousness of the threat and concerns that such threats are likely to be ongoing, improving the detection of such efforts is critical. That desire to help bolster our democratic processes from illicit interference motivated our current study, which attempted to pilot improved detection methods prior to the 2020 election—we wanted to detect any such efforts in time to provide warning rather than post hoc. We found convincing evidence of a coordinated effort, likely foreign, to use social media to attempt to influence the U.S. presidential election.

This report is the second of a four-part series for the California Governor's Office of Emergency Services (Cal OES) designed to help analyze, forecast, and mitigate threats by foreign actors targeting local, state, and national elections.

We would like to thank our sponsors at Cal OES. We are grateful to Agnes Schaefer and Pete Schirmer of the RAND Corporation for their dedicated work supporting this study. Finally, we thank Zev Winkelman and Joshua Kerrigan of RAND for their thoughtful reviews.

This research was sponsored by Cal OES and conducted within the International Security and Defense Policy Center of the RAND National Security Research Division (NSRD). NSRD conducts research and analysis for the Office of the Secretary of Defense, the Joint Staff, the Unified Combatant Commands, the defense agencies, the Navy, the Marine Corps, the U.S. Coast Guard, the U.S. Intelligence Community, allied foreign governments, and foundations.

For more information on the RAND International Security and Defense Policy Center, see www.rand.org/nsrd/isdp or contact the director (contact information is provided on the webpage).

For more information on this publication, visit www.rand.org/t/RRA704-2.

www.rand.org

RR-A704-2