SOCIAL SCIENCES

Content-based features predict social media influence operations

Meysam Alizadeh[1]*, Jacob N. Shapiro[1], Cody Buntain[2], Joshua A. Tucker[3]

We study how easy it is to distinguish influence operations from organic social media activity by assessing the performance of a platform-agnostic machine learning approach. Our method uses public activity to detect content that is part of coordinated influence operations based on human-interpretable features derived solely from content. We test this method on publicly available Twitter data on Chinese, Russian, and Venezuelan troll activity targeting the United States, as well as the Reddit dataset of Russian influence efforts. To assess how well content-based features distinguish these influence operations from random samples of general and political American users, we train and test classifiers on a monthly basis for each campaign across five prediction tasks. Content-based features perform well across period, country, platform, and prediction task. Industrialized production of influence campaign content leaves a distinctive signal in user-generated content that allows tracking of campaigns from month to month and across different accounts.

INTRODUCTION

The same features that make social media useful to activists—low barriers to entry, scalability, easy division of labor, and freedom to produce media targeted at any given country from almost anywhere in the world (*1*, *2*)—also render it vulnerable to industrialized manipulation campaigns by well-resourced actors, including domestic and foreign governments (*3*). We define coordinated influence operation as (i) coordinated campaigns by one organization, party, or state to affect one or more specific aspects of politics in domestic or another state, and (ii) through social media, by (iii) producing content designed to appear indigenous to the target audience or state. South Korea conducted the first documented coordinated influence operation on social media in 2012 (*4*). Since then, what some term political astroturfing has spread widely [on this phenomenon in U.S. domestic politics, see (*5*)]. There were at least 53 such influence efforts targeting 24 countries around the world from 2013 to 2018 (*3*).

One well-covered example of these campaigns is the alleged effort by Russia's Internet Research Agency (IRA) to shape American politics, for which it was indicted by the U.S. government in February 2018. Social media platforms have worked to limit coordinated influence operations and published reports on campaigns on Facebook [from Egypt, United Arab Emirates (UAE), and Saudi Arabia], Reddit (from Russia), and Twitter (from Bangladesh, China, Ecuador, Iran, Russia, Saudi Arabia, Spain, UAE, and Venezuela) (*6*).

Previous academic work on the topic has focused on characterizing influence campaigns, including describing the methods used by IRA trolls to affect public opinion (*7*, *8*), inferring influence effort strategies and tactics based on observed behaviors (*9*–*12*), assessing their political influence (*13*), showing how other Twitter users interacted with IRA trolls (*14*), exploring the co-occurrence of images shared by the IRA trolls and real-world users (*15*), and identifying which kinds of users were likely to spread IRA content (*16*).

One key open scientific and policy question is how easy it is to distinguish industrialized information campaigns from organic social media activity. This issue is different in critical respects from the well-studied issue of bot detection [e.g., (*17*–*19*)]. First, influence operations typically involve a mix of manual and automated activity (i.e., not all participating accounts are bots), and automation is widely used for other purposes (i.e., not all bots are part of influence campaigns). Second, the key feature of an influence operation is the coordinated behavior across multiple accounts, as opposed to the behavior of individual accounts.

A small body of work has shown that it is possible to find coordinated influence efforts on social media using unsupervised (*4*, *20*) or supervised machine learning (*21*, *22*). Unsupervised approaches usually involve leveraging some external intelligence to narrow down tracking, constructing networks, cluster analysis of accounts (also known as community detection) to identify coordination, and manually inspecting each cluster (*20*). Unsupervised approaches have two drawbacks: (i) they can only identify coordination after it has happened, and (ii) they are not scalable. Extant supervised models do not answer whether such activity generally leaves a discernible signature, how the detectability of influence operations varies over time and across campaigns, or why it might do so. See section S1 for a detailed review of related work.

We address those gaps by fixing a platform-agnostic supervised machine learning approach and systematically studying performance over time on unseen, out-of-sample data across multiple platforms, campaigns, and prediction tasks. Our unit of analysis is the post-URL (universal resource locator) pair, an object that exists on almost all social media (a post can be a tweet, Reddit comment, Facebook status update, etc.), making our approach platform agnostic. However, it does not mean that we filter out those social media posts without any URL. If a post does not include a URL, then we capture that as a separate feature (i.e., number of URLs in a post) and put zeros for its URL-related features. Our test data include posts from coordinated influence campaigns and those by random samples of American users and random samples of politically engaged Americans.

Because user-level data are often hard for platforms to share, and because calculating features from users' friendship network

[1]School of Public and International Affairs and Department of Politics, Princeton University, Princeton, NJ 08540, USA. [2]Department of Informatics, New Jersey Institute of Technology, Newark, NJ 07102, USA. [3]Department of Politics and Center for Social Media and Politics, New York University, New York, NY 10012, USA.
*Corresponding author. Email: alizadeh@princeton.edu

(e.g., followers and friends in Twitter) are computationally intensive for large networks (and require customization depending on how the specific platform lets users communicate), we rely only on features that can be calculated from a given post-URL pair (e.g., a tweet containing a URL), what we term "content-based features." These include both characteristics of the post itself—e.g., timing, word count, or if the URL domain is a news website—and how material in a given post relates to other material in that period—e.g., if the URL domain is in the top 25 political domains shared by trolls in the training data—what we term "meta content." Our classifier does not use any historical or friendship network features of users when deciding on a given content (i.e., post-URL pair).

To systematically assess how well these content-based features distinguish coordinated influence activity from normal user activity, we conduct a rich set of experiments that systematically vary: (i) the country responsible, by using all known English-language social media influence campaigns attributed to China, Russia, and Venezuela; (ii) the time period, by running each experiment once a month over the longest feasible period (36 months for most tests); and (iii) the difficulty of the classification challenge. Specifically, we consider four increasingly difficult tasks:

Task 1: Distinguish influence-effort activity in month t from normal activity using data on only a portion of troll activity in month t. This standard train/test split experiment allows us to assess the predictability of influence operations activity over time. It simulates a system in which one received data about an influence campaign (perhaps through the platforms' ex post forensics using backend signals unavailable to users) and tried to detect other posts related to that campaign in the same month.

Task 2: Identify social media posts from troll accounts in month t using data on troll activity in month $t - 1$. This experiment tests how consistent influence campaigns are over time in terms of their public activity. We do not use account creation date–related features (the only user-level feature set we have) to avoid easy detection based on user-identifiable features. Had we not removed such features, this would probably be the easiest prediction task due to having content produced by same users in train and test months.

Task 3: Find social media posts from troll accounts in month t that have not posted in month $t - 1$ using data on troll activity in month $t - 1$. This test examines how similar past content is to the content produced by new troll accounts in the current period, which is effectively an indicator of whether those operating new accounts are using the same tactics, techniques, and procedures as those operating existing accounts. This experiment also highlights the utility of content-based features for finding new users in the next month that are part of a given campaign.

Task 4: Detect activity across different data releases by platforms that were leveraging backend signals and manual investigation to find influence campaigns. Twitter released two major datasets related to Russian activity (October 2018 and January 2019) and two on Venezuelan campaigns (January and June 2019). In both cases, there were trolls identified in the second release who were active before the first release. For each country, we study whether content-based features could detect tweets from trolls in the second data release in month t using data from the first release on troll activity in month t. This test addresses the question of whether content-based features could have complemented the platform's detection approaches (at the time of their first release at least).

Task 5: Identify social media posts from trolls in month t on a given platform using data on troll activity in month t on another platform. Russian trolls were active on both Twitter and Reddit. We test whether classifiers trained on Twitter (Reddit) data in month t using the 859 mutual features can detect social media posts from trolls on Reddit (Twitter) in the same month. This experiment tests whether data about influence operations on one platform can help find them on another. While previous research found that "IRA Reddit activity granger caused IRA Twitter activity within a one-week lag" in terms of the variations in the number of their social media posts (*23*), the extent to which it represents content similarity across the two platforms remains unclear.

Running each task over multiple short time periods provides several important advances over previous work. First, we can assess how distinctive influence campaign content is over time by asking how well content-based features distinguish it from normal activity as new political issues arise, as the platforms modified their content moderation standards, and as the would-be influencers adapted their strategies. Second, we can assess how the features that distinguish coordinated influence campaigns from other activity change over time, providing insight into tactics. Third, because the volume of influence activity varies over time, these tests also provide evidence on how much activity is needed to make which kinds of predictions.

Across 14 experiments on tasks 1 to 4 (excluding task 5), an out-of-the-box random forest classifier applied to a rich vector of human-interpretable content-based features performs well at distinguishing influence operation activity from normal users. Average F1 scores at the monthly level vary from a high of 0.99 in the case of a 50/50 train/test split on Venezuelan operations to a low of 0.74 in the case of using last month's influence operation activity to identify activity by previously unseen accounts that are part of the Russian campaign on Reddit. As for task 5, we obtained average monthly F1 scores of 0.60 for training on Twitter and testing on Reddit and 0.38 for training on Reddit and testing on Twitter, which suggest that a considerable share of Twitter content was generated without testing on Reddit and deployed without coordination with the Reddit effort. The features that distinguish coordinated influence operation's content are quite dynamic, suggesting that any application of machine learning to this detection challenge must involve frequent retraining.

Moving beyond our specific tests, this study shows that content-based prediction of coordinated influence efforts has a wide range of potential uses. A system that learns from a previously identified coordinated activity could (i) enable warnings to users in near real time when they are posting content similar to that of ongoing campaigns; (ii) facilitate estimation of aggregate effort by influencers, which can further be disaggregated by issue area and target geography; and (iii) enable a retrospective assessment of how much content spread by influencers was shaping others' postings. In addition, because it will inevitably take time for platforms to release datasets of newly detected coordinated operations, content-based methods can contribute to detecting ongoing operations and potentially help inform the public about some share of ongoing efforts not yet reported. Last, while the platforms have not released their detection steps (see section S2 for a review of public available information on how Twitter detects and handles these inauthentic coordinated activities), we expect that it most likely includes an unsupervised detection phase in which they identify a set of suspicious accounts, and a forensic testing phase in which they check the specific indicators of compromise left behind by the suspicious accounts [e.g., internet protocol

(IP) address and shared infrastructure such as internet service provider]. In this hypothetical scenario, the approach we evaluate here can be used to complement existing detection procedures.

RESULTS

We study the performance over time of a content-based classifier at distinguishing coordinated influence activity from that of normal users on four different out-of-sample tests across data on Chinese, Russian, and Venezuelan troll activity targeting the United States on two platforms.

This approach differs from previous studies, which have mostly focused on training one classifier on the entire dataset. By training classifiers on a monthly basis and testing on unseen, out-of-sample data, as described above in tasks 1 to 5, we are able to evaluate how well content-based approaches work, how their performance varies over time as the political issue environment changes and tactics of influencers shift, and how stable or dynamic the information that differentiates them from normal activity is.

Because our goal is to assess the basic scientific question of how well content-based features predict social media influence operations over time and across campaigns, we do not optimize the machine learning stage of our process. This ensures that we have the same parameters for all classifiers, making our tests apples-to-apples and oranges-to-oranges comparisons. Instead, we use an out-of-the-box random forest algorithm, learn only on 1 month of training data, use the default classification threshold of 0.5, and do no hyperparameter tuning. The results in this section therefore represent a lower bound on the performance of content-based classifiers.

We first present performance results and important features across influence campaigns over time for each of our four tests. We then examine what drives variance in classifier performance, our proxy for how easy it is to distinguish influence operation content from normal activity. We then focus on the Russian IRA operation on Twitter more deeply, focusing on what changes in feature importance over time tell us about shifts in their tactics, techniques, and procedures.

Performance

Performance of content-based approaches is generally good, although it varies across campaigns and over tests, as shown in Table 1, which reports mean F1 scores for monthly classifiers by experiment. F1 is the harmonic mean of precision and recall. It is a standard metric for binary classification tasks. Precision is the fraction of true positives over the sum of true positives and false positives. Recall is the fraction of true positives over the sum of true positives and false negatives.

Task 1: Cross-section train and test on month t

This task assesses how useful a sample of troll content at one point in time is at finding other such content, essentially a measure of whether their activity is systematically different from that of normal users in terms of content alone. Troll activity is quite distinct; the average monthly F1 score is 0.85 for the Russian Twitter campaign from 1 December 2015 to 1 September 2018 (Table 1). Precision is almost always greater than recall, i.e., while most of the tweets labeled as troll activity by the classifier belong to trolls, the approach always misses a portion of troll tweets. Performance for Russian operation on Reddit from July 2015 to November 2016 is a bit lower (average monthly F1 score of 0.82 in test 1). Similar to the Russian campaign on Twitter, precision is almost always equal or greater than recall. For a more detailed view of performance, we plot the monthly precision, recall, and F1 scores of tasks 1 to 5, along with the monthly number of positive (i.e., trolls) cases in train and test sets for all campaigns in figs. S1 to S5.

Chinese troll activity is easier to distinguish than Russian, with an average monthly F1 score of 0.89 on task 1 for the 47 months between January 2015 and December 2018 (Table 1). Once again, precision is greater than recall in most of the months. The consistency between Chinese and Russian efforts in terms of performance on these metrics suggests that content-based classifiers will generally be very selective. The Venezuelan English-language influence operation on Twitter is easy to detect with content-based features. This is because of their excessive use of distinct and fake websites, focus only on political issues, and simple organization of users, as we discuss at length in the "What makes Venezuelan campaigns easy to detect?"

Table 1. Mean and SD of monthly macro-averaged F1 scores.

Country	Platform	Task 1: Within-month train/test*	Task 2: Train on *t* – *1*, test on *t*†	Task 3: Train on *t* – *1*, test on new users in *t*‡	Task 4: Within-month cross-release	Task 5: Within-month cross-platform
China	Twitter	0.89	0.93	0.89	NA§	NA‖
		(0.08)	(0.04)	(0.12)	–	–
Russia	Twitter	0.85	0.81	0.81	0.75	0.60‡
		(0.13)	(0.07)	(0.13)	(0.11)	(0.03)
Russia	Reddit	0.82	0.82	0.74	NA§	0.37
		(0.07)	(0.09)	(0.15)	–	(0.03)
Venezuela	Twitter	0.99	0.99	0.92	0.49	NA‖
		(0.03)	(0.002)	(0.15)	(0.07)	–

*Training data are all tweets from a 50% random sample of troll users combined with independent random samples from each of our two control groups. Test data use all tweets by the other 50% of troll users and a stratified random sample of 50% of tweets by nontroll users. †Because this test includes the same troll accounts in both train and test sets, we exclude features related to account creation date. ‡We calculate mean and SD in F1 over months in which there are at least 1000 troll tweets or 500 troll Reddit posts in the test month. §Not applicable. There was only one official data release for the Chinese campaign on Twitter and the Russian campaign on Reddit as of 1 December 2019. ‖Not applicable. Cross-platform data are only available for Russian campaign.

section. From October 2016 to February 2019, our classifier yields average monthly F1 score of 0.99 in test 1.

Task 2: Train on t – 1, test on all users in t

This task effectively assesses how consistent troll content is over time by seeing whether troll activity in the previous month distinguishes such activity in the current month. Unlike our other tasks, the same troll accounts can be present in both training and test data. We therefore remove the only user-level information that we used in feature engineering, account creation date, and features related to it (e.g., days since creation, creation date before 2013, and creation date less than 90 days). We obtained fairly stable prediction performance across campaigns and months (fig. S2), with a minimum average monthly F1 score of 0.81 for Russian operations on Reddit and a maximum of 0.99 for Venezuelan operation on Twitter (Table 1). Troll activity appears to be fairly predictable month to month.

Task 3: Train on t – 1, test on new users in t

Probably the hardest test for a supervised classifier in this space is identifying activity by previously unseen accounts that are part of a previously observed effort. We simulate this challenge by training classifiers on all available troll social media posts in month $t - 1$ and testing on social media posts in month t by trolls who were not active in $t - 1$ (fig. S3). The duration of analyses in these tests is shorter than for task 1 or 2 because of the low number of new users in some periods. In reporting average results, we restrict attention to months with at least 1000 troll tweets, or 500 Reddit posts, in the test set (which drops 1 month from the Russian Twitter campaign and reduces the sample to 11 months for the Venezuelan one).

For the 36-month period from January 2015 to January 2018, we obtained an average monthly F1 score of 0.81 for Russian IRA operation on Twitter (fig. S3A). Recall is higher than precision in almost all months with less than 500 troll tweets in the test set. The classifier is still detecting the vast majority of troll tweets in these months, but it is misclassifying a relatively large set of nontrolls as trolls. Content-based features provide an average F1 score of 0.74 for finding new IRA activity on Reddit from July 2015 to November 2016 (fig. S3B).

Even in this test, detecting Venezuelan troll activity is relatively easy. Our approach produces an average monthly F1 score of 0.92 for the 10-month period between October 2016 and January 2018, with new users in this campaign. The performance drop in October 2016 is due to a sudden increase in the number of newly created accounts. Similarly, the drop in July 2017 is a function in part of the addition of 190 new accounts in that month. When a large number of new accounts become active, that likely represents a shift into topically new content, making the classification task harder than if a small number of new accounts are being activated to comment on previously discussed topics.

Turning to Chinese operations, content-based features provide an average monthly F1 score of 0.89 for identifying activity by new Chinese trolls over the 36-month period from January 2016 to December 2018 (fig. S3D). The predictive performance of the classifiers shows cycles of gradually decreasing over 6-month intervals to approximately 0.7 and then increasing to greater than 0.9. This pattern matches the regular cycles of new account creation in the Chinese influence operations evident in the lower panel of fig. S3D.

Overall, influence operation content appears to be quite standardized month to month. New accounts introduced in any given month are producing posts that look a great deal like those of existing accounts, especially for Chinese and Venezuelan activity.

Task 4: Train on first data release, test on second data release

Twitter released two major datasets related to Russian IRA on October 2018 and January 2019 (the third release on June 2019 only includes four IRA accounts), as discussed in greater detail in Materials and Methods. It also released two datasets related to Venezuela, in January and June 2019. In both cases, many of the trolls identified in the second release were active at the same time as trolls identified in the first release. This implies that Twitter's detection either failed to identify the accounts in the second set before the first set was released or failed to do so with sufficient certainty. Either way, assessing how well a classifier trained on the first release would perform in detecting tweets from users in the second release provides evidence about whether content-based features provide information not found in the account-level features that Twitter initially relied on. As section S2 documents, publicly available information suggests that Twitter relied primarily on account-level features in their first release to attribute a coordinated behavior to a country or organization, such as whether the user logged in from any Russian IP address more than once.

Content-based approaches provide average recall of 0.71 and precision of 0.87 for the IRA campaign for the 36-month period between January 2015 and January 2018 (fig. S3A). The large gap between precision and recall in most months implies that the classifier trained only on the first release is selective: Tweets missed by Twitter's first release that the classifier identifies as belonging to trolls do so 87% of the time. But the classifier only identifies 71% of the missed troll activity. One interpretation is that while the majority of activity identified in Twitter's second data release was being done by accounts operating with the same goals as those in the first release, roughly 30% of accounts were focused on new issues or were operating with a different set of guidelines.

Cross-release classification performance on Venezuelan Twitter activity is the weakest of all the tests we conducted, with a mean precision of 0.60 and a mean recall of 0.51. This is because the content being promoted by accounts in each Twitter release was fundamentally different. Accounts in the first release were focused tightly on U.S. politics. In the second release, the accounts worked to affect English-language users' views on a wide range of political issues. Content-based features perform poorly across those two substantively distinct political campaigns.

Task 5: Cross-platform train and test on month t

This test explores the extent to which classifiers trained on Twitter can detect an ongoing coordinated influence operation from the same country/organization on Reddit and vice versa. Both Twitter and Reddit released data on influence operations attributed to IRA. Assessing how well a classifier trained on either of these in a given month would perform in detecting the content of the other in the same month provides evidence about whether their simultaneous operations across different campaigns shares similar content-based characteristics.

For the 17-month period between July 2015 and November 2016, we obtained average monthly F1 scores of 0.60 for training on Twitter and testing on Reddit, and 0.38 for training on Reddit and testing on Twitter. Classifiers trained on Reddit and tested on Twitter always perform poorly (fig. S5A). This may be due to having too few positive cases in the training data compared to the test data (and thus an inability of the classifier to capture the diversity present in the test data) or to the fact that Reddit post titles are much shorter and therefore less informative than tweets. However, classifiers trained on Twitter and tested on Reddit do yield reasonable F1 scores of

approximately 0.70 between May and October 2016 (fig. S5B). Looking at the timeline and characteristics of troll activities on Reddit helps makes sense of this second pattern.

Russian trolls activity on Reddit had three phases (see fig. S9A): (i) karma farming: between July and November 2015, trolls posted in popular culture subreddits such as "cute" and "humor" subreddit categories as a cheap way of earning karma (see fig. S9B for variations in category of subreddits targeted by Russian trolls); (ii) inactivity: between December 2015 and March 2016, the troll activity decreased substantially to few dozen posts per month; and (iii) political engagement: between April and October 2016, trolls increased their activity again, this time posting mostly political content in subreddit categories such as "politics," "alt-right," "conspiracy," and "angering." We suspect that classifiers trained on Twitter worked poorly on Reddit in phases 1 and 2 because of topical differences between the mission of the trolls in each platform. However, once Reddit trolls began engaging in political conversations, as they were on Twitter, it became possible to detect them using classifiers trained on Twitter data.

Our finding of poor cross-platform prediction performance and the previous finding that the volume of IRA Reddit activity Granger causes IRA Twitter activity (*23*) are both consistent with qualitative work, suggesting that a portion of the ideas used on Twitter were tested on Reddit (*24*). But they suggests further that a large share of Twitter content was generated without testing on Reddit and deployed without coordination with the Reddit effort.

Effect of train and test periods

For prediction tasks 2 and 3 (train on month $t-1$ and test on t), an important question is how dynamic the difference between troll and normal user activity is. If the language used by one or both parties varies tremendously week to week in response to events and news, then (i) reducing the test period to weekly should improve performance and (ii) extending the training period back should lower performance. With weekly models, the shorter testing period would allow the classifiers to catch up with most recent trends in troll activity more often than monthly classifiers. Extending the training period back in time would mean learning the differences between troll and normal users in a period very different than the test period.

To check (i), we estimate a weekly version of task 3 (our hardest task) on the Russian Twitter activity (result plotted in fig. S8A). The weekly F1 score is 0.85 for weeks, with 1000 or more troll tweets in the test period, approximately 4 percentage points greater than the corresponding monthly performance reported in Table 1. To check (ii), whether training on longer periods reduces prediction performance, we use the Russian Twitter data and increase the training period for task 3 to 3 months. Figure S8B compares the F1 scores with the longer training period to the baseline reported above. While training on a 3-month period improves performance in few months, it leads to 9 percentage point reduction in average monthly F1 scores over the full period.

Although the differences between normal users and trolls are quite dynamic, we focus on monthly results for several reasons. First, we want to set a lower bound on content-based detection. Second, there are relatively few weeks (of 175 weeks, only 27 had more than 1000 and 38 had more than 500 troll tweets in the test set) with enough posts by new troll accounts to make task 3 meaningful. Third, in practice, weekly retraining would require weekly deliveries of annotated troll data from the platforms or other sources, which is unrealistic given the investigative process.

Effect of different control users

Our main tests combine samples of random and politically engaged American Twitter users as the control users (i.e., negative class). To measure how distinctive the activity of trolls is from each of these samples, we compare the performance of classifiers trained and tested against only random and politically engaged users, respectively.

Distinguishing between trolls tweets and tweets written by politically engaged users is almost always harder than those written by random users, as we see in table S7, which reports average monthly F1 scores for tasks 1 to 5 on Chinese and Russian Twitter trolls (the exception is tasks 1 and 2 for the Chinese influence operation). We left Venezuelan trolls out because our prediction performance is close to 1 for them and shows no difference between training and testing against random or politically engaged users.

Across prediction tasks 1 to 5, classifiers trained and tested against random perform an average of 7 percentage points higher in terms of F1 scores than those trained and tested against politically engaged users. In the case of Chinese activity, performance is similar across samples, with tasks 1 and 2 showing slightly better average F1 scores on politically engaged users and task 3 yielding higher scores for random users.

In practical terms, this means that Russian trolls produced content that looked more like that of politically engaged American users than of random American users. Chinese troll content was similarly distinctive from both random and politically engaged users' content.

Explaining performance over time and across experiments

Measuring the sources of variations in classifiers' prediction performance over time gives us a way of assessing which features explain the variation in performance and how much variation they explain. To assess the sources of performance, we regress monthly precision, recall, and F1 scores on campaign, prediction task, temporal trends, communication type of troll content, data characteristics (size of minority class in train and test and severity of class imbalance), and number and significance of political events in the United States in a given month. Comparing coefficients from this exercise also allows us to compare how much harder certain tasks or campaigns are to distinguish from organic activity relative to other ones.

Specifically, we estimate the following where i indexes experiments (i.e., campaign/task pairs) and t indexes months

$$y_{i,t} = \boldsymbol{\alpha}\mathbf{C_i} + \boldsymbol{\beta}\mathbf{T_i} + \gamma\mathbf{X_{i,t}} + \boldsymbol{\tau}\boldsymbol{\Theta}_\mathbf{t} + \boldsymbol{\epsilon}_\mathbf{i,t} \quad (1)$$

$y_{i,t}$ is the outcome (recall, precision, or F1), $\mathbf{C_i}$ are indicator variables for which campaign i is in, $\mathbf{T_i}$ are indicator variables for the prediction task, $\mathbf{X_{i,t}}$ are time-varying controls, Θ_t is a set of controls for temporal trends, and $\epsilon_{i,t}$ is an error term. We report heteroscedasticity-consistent SEs (clustering at the experiment level changes little).

We vary $X_{i,t}$ and the estimation sample in six different regressions: (i) controlling for different operations by country and platform (baseline); (ii) controlling for quarter fixed-effects and a quintic polynomial in time; (iii) excluding the Reddit data because they have many fewer observations than the three Twitter operations (and thus should not be included when controlling) for data characteristics; (iv) communication type of trolls, including share of troll posts in a given month that contain retweets, replies, hashtags, mentions, and links to local news URLs; (v) controlling for the log number of positive cases in train and test sets as well as class imbalance in a test set; and (vi) controlling for the number of major political events happened at each month and their average severity. Because of space

limitations, we plot coefficient estimates and SEs for model (vi) in Fig. 1, and place full results in table S1.

Five facts stand out from this analysis, in which the baseline task is cross-sectional prediction (task 1) and the baseline campaign is Russia's efforts on Twitter. First, Chinese and Venezuelan activities on Twitter are substantially more distinctive than Russian: F1 for them is approximately 5.7 and 9.5% greater than on Russian activity controlling for temporal trends (see table S1, column 3; mean F1 for Russian Twitter activity is 0.79). Most of that effect, however, is due to the communication type features of their activity. Once those are accounted for, Chinese activity is no more predictable than Russian and Venezuelan is actually less predictable (see table S1 column 6). Venezuelan operations are easy to find because they use retweets, replies, hashtags, mentions, and URLs in an unusual way. Second, conditional on timing and other factors, predicting forward in time (task 2) is challenging (F1 is 2.7% lower than for task 1), separating activity by previously unseen accounts that are part of an influence campaign from organic activity is harder still (F1 is 10% lower than for task 1), and finding missed accounts across data releases is hardest of all (F1 is 31% lower). Third, the communication type of troll content in any given month is very informative; adding variables that capture that communication type increases the share of variance in F1 explained by almost 20% (r^2 goes from 0.41 to 0.49). Fourth, data characteristics matter (adding them to the model increase r^2 by 20%), but in the intuitive manner. The number of positive cases (i.e., troll activity) in train and test sets is important; a 1% increase in the number of positive cases in the test set predicts a 7% increase in precision, and a 1% increase in the number in the training set predicts a 5% increase in recall. But class imbalance in the test set matters in the opposite of the intuitive direction. We suspect that this is because months in which there are very few troll posts (and thus considerable class imbalance) the trolls tend to be posting highly unusual content, which is therefore easy to find. Fifth, performance is lower for months with higher average event significance (as judged by our four raters), but the substantive magnitude of that relationship is small compared to other factors.

Important predictors

A key scientific question is how the content of coordinated influence campaigns is different from that of other users. The experiments above provide valuable evidence on what distinguishes this activity because random forest classification algorithms provide the importance of each feature in terms of a real number. These variable importance measures give us a way to assess the importance of different

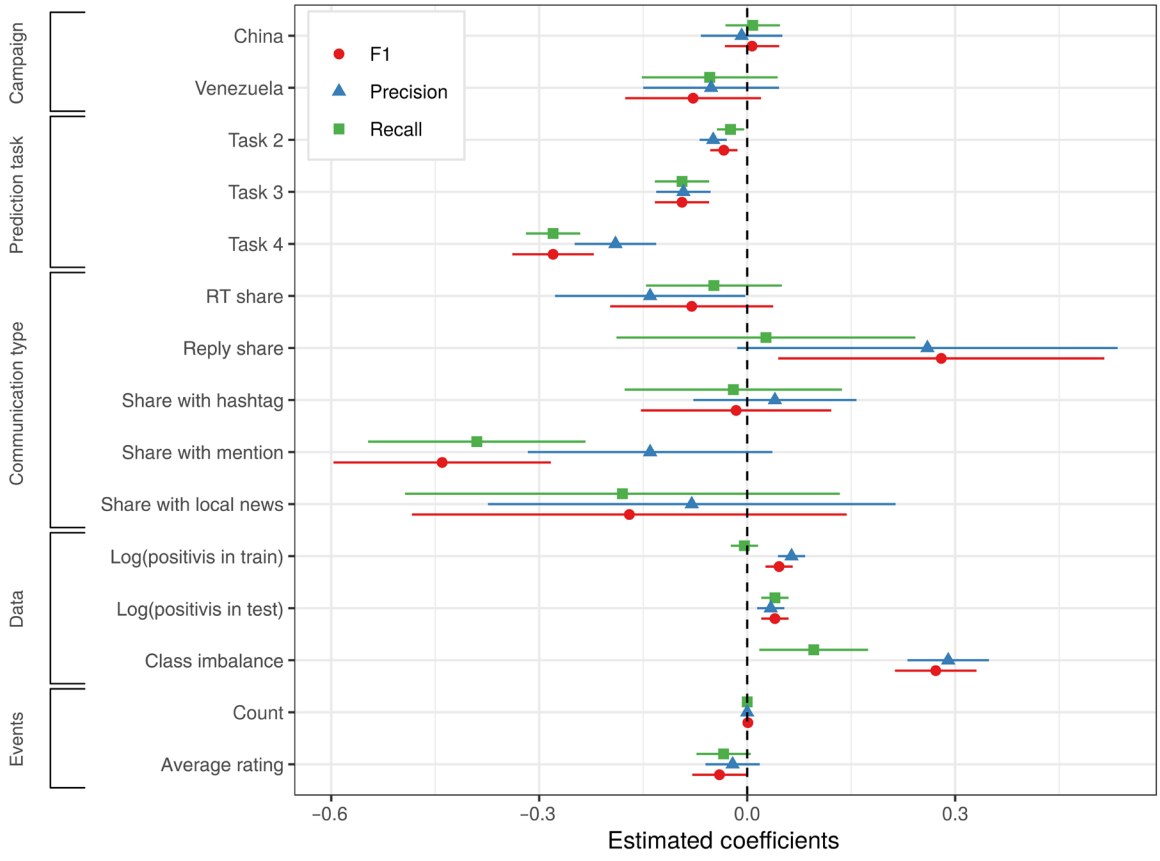

Fig. 1. Ordinary least squares (OLS) regression coefficients for variables explaining the predictive performance of classifiers across campaigns and tests. Points represent estimated coefficients, and bars represent the 95% confidence interval of the estimate. F1 scores for Chinese and Venezuelan operations are approximately 5.7 and 9.5% greater than on Russian activity controlling for temporal trends. Most of that effect, however, is due to their distinct communication features. Once those are accounted for, Chinese activity is no more predictable than Russian, and Venezuelan is actually less predictable. In other words, Venezuelan operations were the easiest to detect because of their unusual way of using retweets (RT), replies, hashtags, mentions, and URLs. Conditioning on timing and other factors, F1 for task 4 is 31% lower than for task 1, making it the hardest prediction task. A 1% increase in the number of positive cases in the test set predicts a 7% increase in precision, and a 1% increase in the number in the training set predicts a 5% increase in recall. Task 5 is excluded because of noncomparability issue.

features for detection of social media influence operations. Here, we review the key features across campaigns for task 3 (our hardest test) and report the top 10 features that most often have monthly variable importance of 0.1 or greater in table S5.

Important features vary by month, operation, and platform. For Twitter campaigns, features related to the age of accounts, users mentioned or replied to by trolls, top hashtags and words used by trolls, and combinations of mentioned users and URL domain types are frequently among the most important features, as table S5 shows. Days since creation and whether or not an account was created before 2013 are among the 10 most important features for all but the Russian campaign on Reddit. We suspect that this is because most troll accounts detected to date were created during a short time frame. Most IRA Twitter accounts, for example, were created between mid-2013 and early 2015.

Meta-content helps

To formally examine the relative importance of various types of features, we categorize our features into five groups: content, meta-content, content-level timing, account-level timing, and network (see table S3 for the full list of features in each group). We consider the model trained on content features alone as a baseline and compare prediction performances by adding each group of features across all combinations of platform, country, and test, as in Table 1.

Because of space limit, we only demonstrate the results for Russian Twitter campaign in this section (Table 2) and report the rest in table S6. We also excluded task 5 because it is based on a reduced set of features and therefore not comparable to others tasks. Compared to baseline, adding meta-content features on average increases the F1 score by 6.5 percentage points across our four tests. Content-level timing features are not effective and add little to the performance (after accounting for other aspects of the content they produced, the fact that many IRA trolls worked St. Petersburg hours in 2016 does not appear to be important). Account-level timing, however, increases the F1 score by 4.3 percentage points, on average, across various tests. Similar patterns can be observed in results from the other campaigns (table S6). Last, including network features (e.g., various attributes of the co-shared and co-occurring hashtags network) has mixed effects on the prediction performance. In some cases, it leads to better performance, but in most of the cases, it has zero or negative effects; hence, their exclusion from the results above (see tables S3 and S4 for complete lists of the considered network features).

Feature importance trends reveal tactical changes

Analyzing the dynamics of feature importance over time can provide insight into troll's tactics. Doing so requires care as it is ambiguous whether changes in feature importance comes from control users or trolls. However, by checking trends in both groups to see which one moved on a given feature at a particular time, we can still sort out that ambiguity. Doing so provides insights into how Russian trolls built their audience and tried to evade detection.

On Twitter, one can deploy various audience-building or campaigning tactics by using a combination of hashtags, retweets, replies, mentioning other users, posting at specific times, and sharing URLs. To illustrate how their relative importance can help us understand tactical changes, Fig. 2 plots importance trend on Russian activity for task 3 (train on $t-1$ and test on new users in t) for nine key features divided as features whose importance was higher (i) before the November 2016 U.S. elections (top row), (ii) after November 2016 (middle row), and (iii) outside of election times.

Before 2016, features relate to mentioned users or a combination of mentioned users and shared URLs. The increase in the importance of top users mentioned by IRA trolls through 2015 correlates with an increase in the share of troll tweets with at least one mentioned user (fig. S7C). However, the subsequent importance decrease cannot be explained by changes in the volume or type of mentioned users (see fig. S6 for a timeline of six types of users mentioned by IRA trolls). This pattern suggests that IRA trolls began behaving like organic American users in their user mentioning behavior through 2016 and into 2017.

Features combining mentioned users and the types of URLs being shared also reveal tactical shifts. In 2015, trolls stood out for the users they mentioned in posts with links to local news sites, while in 2016, the users they mentioned in posts with links to far-right websites were more distinctive. Both features became important again in 2018. The combination of mentioned users with specific types of co-occurring URLs is distinctive, and its explanatory power cannot be explained by changes in the share of local/far-right URLs (see fig. S7F for the share of tweets with a link to local news websites) or by the share of mentions.

During election months in 2016, the way IRA trolls were retweeting, replying to others, and mentioning users became more distinctive, as we see in the second row of Fig. 2. To better characterize these changes, we plot the monthly share of retweets, replies, and mentions

Table 2. Monthly mean of macro-averaged F1 scores for detection of Russian troll tweets, with varying predictor sets. User-timing features were removed for task 3. Task 5 is excluded because it is based on a reduced set of features and therefore not comparable to other tasks.

	Only content	(1) + meta-content	(2) + content timing	(3) + user timing	(4) + network features
Model number	(1)	(2)	(3)	(4)	(5)
Experiments					
Within-month train/test (task 1)	0.76	0.81	0.82	0.85	0.84
Train on $t-1$ test on t (task 2)	0.74	0.82	0.82	NA	0.85
Train on $t-1$ test on new users in t (task 3)	0.66	0.75	0.75	0.81	0.82
Within-month cross-release (task 4)	0.66	0.70	0.70	0.74	0.75

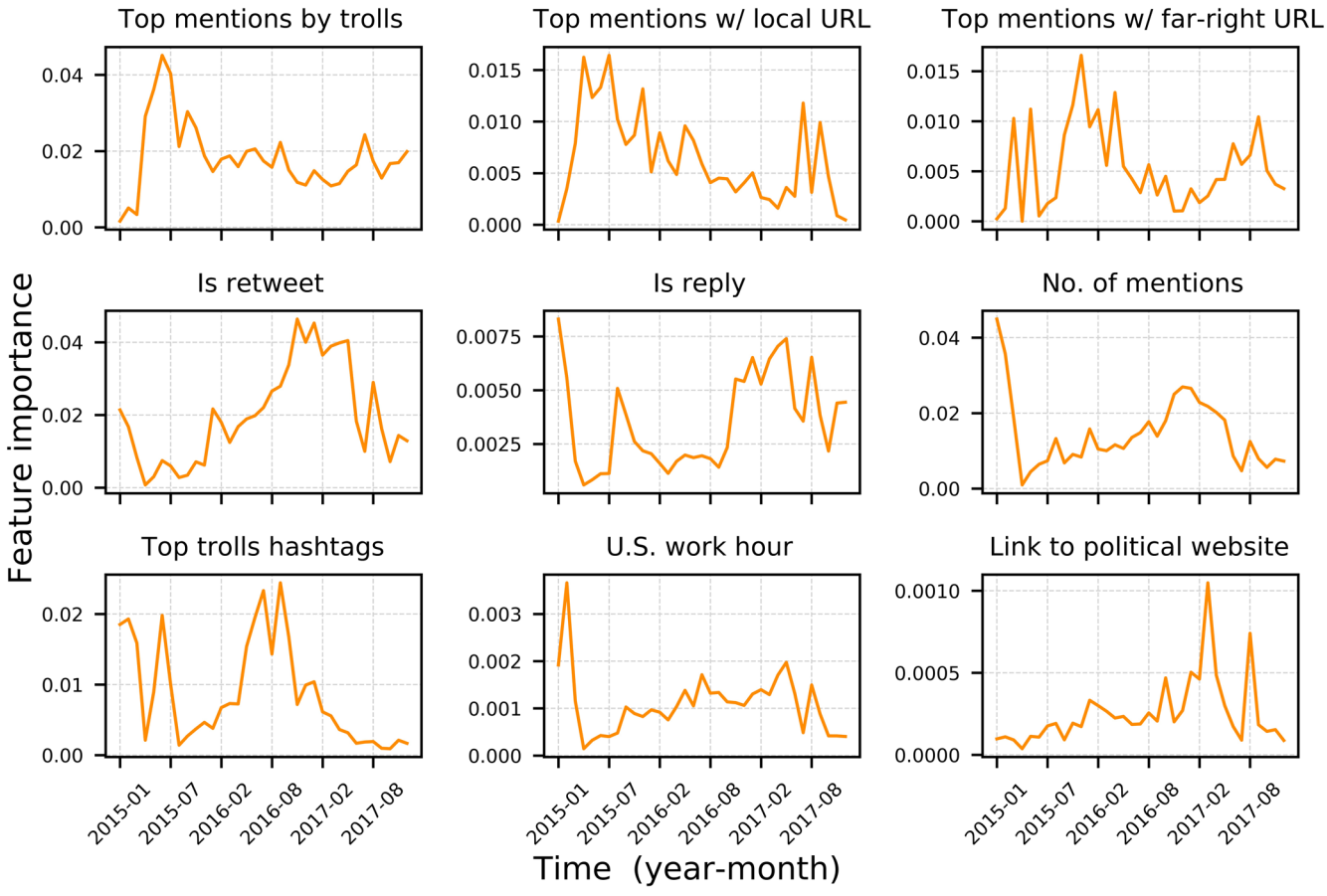

Fig. 2. Feature importance trends of a set of selected predictors. Analyzing changes of feature importance trends over time can reveal tactical changes in the Russian IRA influence operation on Twitter. The top row shows features whose importance was higher before the November 2016 U.S. elections. The second row focuses on features whose importance was relatively higher in the election month of November 2016. The last row illustrates importance trend for features that were most important at other times.

in troll's tweets (see fig. S7). Both mention and retweet shares were higher in November 2016 than at any point before and decreased after the election, but the share of replies does not show any related surge or decline (fig. S7B). The increase in the importance of being a reply on and around the election month is suggestive of changes in control user's behavior, and not a change in troll tactics.

Last, we see clear evidence of tactical adaptation in the bottom row of Fig. 2. While top hashtags used by trolls were highly important during the spring and fall of 2016, their importance declined considerably in November 2016 and continues to decrease over 2017 and 2018. This is suggestive of tactic change in the usage of hashtags by IRA trolls. That interpretation is consistent with the fact that the monthly share of IRA tweets with at least one hashtag peaked in late 2015 and early 2016, but then continuously decreased until 2017 (fig. S7D). We suspect that this change was due to IRA operators realizing that hashtags are a powerful clue to identify their activities, as analysts and platforms can simply query distinctive hashtags to detect coordinated communities.

What makes Venezuelan campaigns easy to detect?

Content-based features provide near-perfect prediction performance for the Venezuelan influence operation on Twitter in most of the months for the first three tests in Table 1. Our analysis suggests that three factors can explain the large differences between the Venezuelan campaign and the others.

First, for much of their activity, the Venezuelan accounts barely used hashtags and rarely retweeted or replied to others. In addition, except for a few months, they always shared at least one URL in their tweets (see fig. S10 for timelines of the share of troll's tweets with hashtag, retweets, mention, and URL). This is in agreement with the finding in the "Explaining performance over time and across experiments" section that most of the differences between Venezuelan prediction performance and others are due to their distinct types of communication features. Once those are accounted for, the Venezuelan campaign is actually less predictable than others (see table S1, column 6). Second, when they began to retweet other accounts in mid-2018, the Venezuelan trolls mostly retweeted only one account. Ninety-five percent of their retweets were from @TrumpNewsz, which is one of their own accounts (see fig. S10E for the top 20 users retweeted by trolls). Third, the Venezuelan trolls were sharing a lot of distinct fake websites, such as trumpnewss.com and trumpservativenews.club. They also used two nonmainstream URL shorteners: viid.me and zpr.io (see fig. S10F for the top websites shared by Venezuelan trolls). Together, these three tactical choices made Venezuelan activity easily recognizable.

Last but not least, the organization and division of labor within the Venezuelan operation were very simple. If we construct the retweet network of the accounts in the Venezuela's Twitter campaign (Fig. 3A), we see few central accounts and many side accounts (most likely bots) around each, which is the simplest form of running a campaign on Twitter. In comparison, we see half a dozen distinct communities within the retweet network of the IRA trolls (Fig. 3B), which reflects a clear division of labor among them [see (*25*) for a detailed discussion of categories of trolls in Russian IRA accounts]. For example, while accounts in the blue community of the IRA trolls were mostly engaging in hashtag gaming or sharing commercial and diet links, users in the green and purple communities were in charge of targeting Republican and Democrat supporters, respectively. However, the central accounts in Venezuela's campaign were all related to Trump and Trump-related issues.

DISCUSSION

Domestic and foreign agents no longer need to physically participate in street riots or student protests to polarize the population or invest in television advertisement or movies to manipulate public opinion. Although social media platforms make advertisement cheaper and make it easier for non-incumbents to get public and media attention, they also make it easier and cheaper to conduct influence operations shape politics at home or in a foreign state. While the research community has yet to measure the extent to which social media manipulation efforts affected voter preferences (*26*), there is a strong consensus within academic scholars and policy makers that action should be taken to address this malicious behavior.

To better understand how easy it is to distinguish such activity from that of normal users, we developed a platform-agnostic supervised learning approach to classifying posts as being part of a coordinated influence operation or not. To assess variation in the predictability of industrialized influence operations, we evaluate the system's performance on a monthly basis across four different influence campaigns on two platforms in four distinct tests (for a total of 16 experiments and 463 observations).

Overall, the results show that content-based features distinguish coordinated influence campaigns on social media. They also provide some key insights about what makes these campaigns distinctive. First, content-based classifiers do a pretty good job of identifying posts in most campaigns. This is likely because, to achieve impact, the campaigns need to produce a lot of content, which requires a substantial workforce using templates and standard operating procedures. Second, meta-content, how a given piece of content relates to what others are saying at that time, is an important complement to primary content. Third, as troll tactics change, the features that distinguish their activity do as well. This should make us cautious about the promise of generic unsupervised solution to the challenge of detecting coordinated political influence operations. Fourth, there is massive variation in the level of skill across campaigns.

Our results also have practical implications. An important policy challenge in combating coordinated influence operations is to estimate the size of their operations in real time, which, in turn, requires distinguishing participating accounts or content from that of normal users. Fortunately, the research community has made a great progress in detecting accounts controlled by automated approaches (i.e., bots) through developing machine learning–based tools such as Botometer (*18*). This makes it easy to identify less complex influence efforts or promotion campaigns [e.g., (*19*)], in which there are few central human-operated accounts and lots of bots surrounding them to spread their content and amplify their visibility. But much more needs to be done. Detecting more complex influence operations

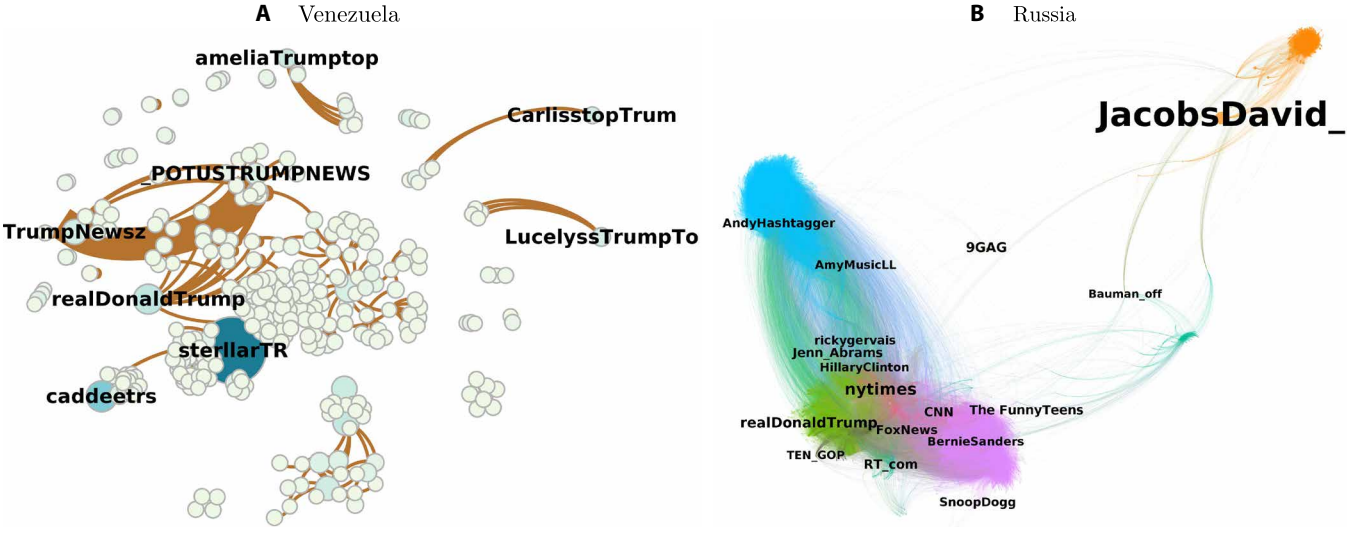

Fig. 3. Characterizing retweet networks of Venezuelan and Russian influence operations. Each node represents a user, and there is an edge between two nodes if one of them retweeted the other. Node label size represents Page Rank score. Color reflects different concepts in each graph. (**A**) Venezuelan trolls. Node color and edge size represent the number of retweets. Venezuelan trolls were mostly interested in tweets from or about Trump. We can see that a considerable portion of their campaign was a single-issue Trump-related campaign. Structurally, we see few central accounts and many side accounts around each, which is the simplest form of running a campaign on Twitter. (**B**) Russian IRA trolls. Node color represents communities derived by applying the Louvain algorithm. Edges are colored by source. Russian operations were quite diverse in terms of their topics and audience of interest. They were targeting right-leaning (green), left-leaning (purple), and African-American left-leaning (green) individuals and hashtag gamers (blue). Structurally, we see target-specific clusters reflecting division of labor with frequent communication between them.

composed of many human- or hybrid-operated accounts working in coordination, which sometimes include multiple teams targeting different types of audiences, is substantially harder than finding automation. Because foreign agents are active on multiple social media platforms—including Twitter, Facebook, Instagram, and Reddit—it is important to build detection tools that are not heavily dependent on platform-specific features.

Content-based approaches can help. Our machine learning approach achieves strong performance on a platform-agnostic unit of analysis (the post-URL pair), with modest amounts of labeled data, short training periods, out-of-the-box libraries, and a limited set of human-interpretable features. Considerably higher classifier performance is surely possible through hyperparameter tuning, dynamic features selection, varying training periods, and other enhancements.

The fact that our simple baseline works well across multiple tasks and campaigns is strong evidence that content-based approaches could (i) support public-facing dashboards alerting polities to the extent of foreign disinformation campaigns, (ii) drive recommender systems to alert users when they are seeing promoted content or inadvertently spreading it themselves, (iii) cue investigations on platforms where user data are hidden for privacy reasons (e.g., anonymous messaging apps), and (iv) help identifying the coordination pattern more quickly.

There are at least two important concerns with a system based on our approach. First, it could benefit those running influence campaigns by, for example, alerting managers trying to hide their activity to posts that stand out. While that is a possibility, we think that it misses the costs of acting on that information. Getting around content-based detection requires more variety in content (e.g., linking to a wider variety of URLs) and less concentrated messaging (i.e., more accounts). Those changes reduce the marginal product of labor for influence operations, which must either increase their costs or decrease their influence. This is the same argument others have made in the context of "adversarial design" (*27*).

Second, we should not expect the approach to do a good job detecting activity that is not yet known if the content is quite different from what has previously been observed. We saw exactly this in the results for task 4, where training on the first release of the Venezuela data did a poor job of predicting content in the second release. This limitation of supervised content-based detection, which can only find new activity that is similar to previous activity, is important. In addition, real-work application of our approach faces the challenge of highly imbalanced positive and negative classes, as we expect the share of troll activity to be very small compared to organic content. A pipeline of using unsupervised approaches [e.g., (*20*)] to uncover new coordinated activities and feeding its output to a supervised classifier seems to be a plausible solution to mitigate these limitations.

Last, there is promise in using experimental design to identify the sources of classifier performance. This analysis is rare in the applied machine learning literature, but when panel data are available on factors that could affect classifier performance, it can enable a richer assessment of the underlying data generating process.

We see two main avenues for future work. First, developing approaches to efficiently follow influence campaigns over time should be a priority. Second, one could likely improve prediction performance by using features extracted from images and videos shared by trolls, developing richer features sets based on the content linked to in posts, and implementing classification approaches that leverage longer histories at the account level.

MATERIALS AND METHODS

Here, we describe a series of experiments using classifiers trained on human-interpretable features to assess whether posts on a given social media platform are part of a previously observed coordinated influence operation. Our unit of analysis is the post-URL pair, making our approach platform agnostic. This, however, does not mean that we ignore platform-specific attributes such as retweets for Twitter or subreddit topics for Reddit. Rather, it means that our approach is generalizable to any platform with post-URL format. We extend and complement the previous works in five ways. First, we train classifiers on a monthly basis. Second, we test our classifier on influence operations conducted by three different countries. Third, in addition to Twitter data, we test our classifier on Reddit posts published by Russian IRA trolls. Fourth, we examine performance on four different out-of-sample tests. Fifth, we introduce new human-interpretable features that are specifically crafted to capture the "coordinated" nature of influence operations. We also use a set of content-based and URL domain–based features to extract more information from each post-URL pair.

Data

Our classifier requires data on two kinds of social media activity: (i) posts by the accounts of a given coordinated influence operation (i.e., positive class) and (ii) a principled sample of organic user's activity (i.e., negative class).

Twitter and Reddit data of coordinated influence operations

We identify post-URL pairs from influence campaigns attributed to China, Russia, and Venezuela using data that the company released in 2018 and 2019 (*6*). The datasets include information from 2660 Chinese, 3722 Russian, and 594 Venezuelan accounts, who have at least one tweet posted in English, many of which targeted U.S. politics (Table 3). Because of the importance of the 2016 U.S. presidential election and its aftermath, we focus on the time period from 1 December 2015 to 1 December 2018. While Twitter has released datasets related to other countries, those that contain mostly non-English content focused either on non-U.S. audiences. We focus on English-language activity because applying our machine learning framework to those data would require both properly collected control data from those languages and countries, which do not currently exist. Moreover, some of the content- and URL-based features we use in the model are not available for other languages and countries, e.g., classification of press sources according to political slant and typical content.

Following Facebook and Twitter's efforts to take down IRA-related accounts, Reddit also found 944 accounts on the platform with ties to the IRA and released the entire list of accounts on April 2018. While 70% of these accounts had zero karma, 6% of them had 1000 to 9999 karmas, and only 1% of them (13 accounts) had a karma score greater than 10,000. The IRA trolls were active on various types of subreddits, including alternative media, humor, angering, lifestyle, cute, and news. Although the IRA accounts were active from early 2014 to mid-2018, the volume of their activity was too low for almost half of this period. Therefore, we only focus on the period between July 2015 and December 2016.

Control data

For a Twitter comparison group, we combine two sources: (i) 5000 random U.S. accounts, sampled by generating random numeric user IDs, validating their existence, and checking the location to be in the United States, and (ii) 5000 politically engaged Twitter users, defined as those who follow at least five American politicians, generated through a similar approach. We collect all tweets published by these 10,000 users between 1 December 2015 and 1 December 2018

Table 3. Summary of Twitter and Reddit data for troll and control accounts.

Platform	Type	Category	No. of accounts	No. of posts
Twitter	Troll	China	2,660	1,940,180
	Troll	Russia	3,722	3,738,750
	Troll	Venezuela	594	1,488,142
	Control	U.S. political	5,000	22,977,929
	Control	U.S. random	5,000	20,935,038
Reddit	Troll	Russia	944	14,471
	Control	Political subreddits	107,052	713,236
	Control	Top 30 Russian-targeted subreddits	784,711	5,475,687

(see Table 3 for statistics). Last, for each month, we randomly sample the size of IRA troll tweets in that month from both random and political users' tweets to create our negative class in training set (i.e., the ratio of troll to nontroll posts in training is 1 to 2). For the test set, we randomly sample half of tweets by random and politically engaged users in each month. This results in varying class imbalance in the test set, with an average ratio of 1 to 5 between positive and negative class. While most similar methods only include users whose time zones or self-reported locations reside in the United States, our data contain a variety of American users by including those users who have not declared their location, but whose location can be estimated through a network of their friends and followers of their friends [details in (*28*)].

For a Reddit comparison group, we combine (i) all posts on 174 political subreddits (*29*) and (ii) all posts on the top 30 subreddits in which the IRA trolls were most active (out of a total of 934 subreddits). For these subreddits, first, we get the list of users who posted at least one post between 1 December 2015 and 31 December 2017 (no IRA activity on Reddit has been officially reported after this period), then uniformly sample 10% of users at random, and then collect all of their Reddit posts published between 1 June 2015 and 31 December 2017 (Table 3). For both samples, we only draw on non-IRA users. Last, for each month, we randomly sample the size of IRA troll posts in that month from both top 30 and political subreddits samples to create our negative class in training set. For the test set, we randomly sample half of Reddit posts by nontroll users in each month, which results in average class imbalance of 1 to 25 between positive and negative classes. All Reddit control data are provided by pushshift.io.

Political events

Major political events in month t could negatively affect classifier performance. If influence campaign activity pivots to discuss the event, then the shifts in word choice and linked to URLs could lead to poor predictive performance. To account for the potential role of political events in shaping classifier performance, we developed a dataset of all major political events in the United States from January 2015 to 2018. Our list of major political events was sourced from the annual lists of major political events occurring in and related to the United States from ABC News, CBS News, CNBC, and the Council on Foreign Relations. We collected specific details of each event by following internal links and citations provided by these sources, including mentioned personae, location, event description, and date.

We then down-selected to events that fit certain categories: political news (related to electoral and domestic politics); foreign relations that affect the United States (such as international accords and nuclear weapon tests); mass shootings; protests; important legislation passed; political scandals (such as sexual harassment/assault allegations against politicians and lawmakers); Donald Trump's political or judicial nominations, hirings, and firings; and candidate or President Trump's notable statements. We dropped events such as natural disasters, death of famous individuals, sports-related news, and entertainment news, unless it had political valence, such as the Camp Fire.

The resulting data included 105 total events. Of these, 25 events occurred in 2015, 34 events in 2016, 26 events in 2017, and 20 events in 2018. Each event was then coded independently by three American RAs and one of the authors. Of 105 identified events, 36 were judged to be major by all four judges, 18 were judged to not be major by all four, and the remainder were evenly spread between having one, two, and three judges determine the event to be major.

Feature engineering

We apply standard text preprocessing steps on our text data and expand shortened URLs. Then, we calculate five types of human-interpretable features for each post-URL pair: (i) content: e.g., word count of a post and URL, topic of a post, sentiment, and Linguistic Inquiry and Word Count (LIWC) scores; (ii) meta-content: e.g., ratio of top 25 words or bigrams used by trolls in a post to all words or bigrams; (iii) URL domain: e.g., whether a URL domain is a news, political, commercial, or sport website; (iv) meta URL domain: e.g., whether a URL domain is in the top 25 political, left, right, national, or local domains shared by trolls; and (v) timing: e.g., day of the week or hour of day on which a post is posted. See tables S3 and S4 for complete lists of extracted features from Twitter and Reddit data. In total, we represent each tweet-URL and Reddit title–URL pairs as a vector of 1300 and 986 features, respectively.

It should be noted that, except for account age–related features, we are not using any user-level features (e.g., number of followers, friends, and posts) or any post popularity or engagement features such as number of favorites, retweets, and mentions. These design choices enable fuller anonymization of potential input data (important, e.g., if one wanted to extend the approach to Facebook data) and minimize compute requirements. In addition, as much as we would like to use Botometer score as a feature, unfortunately, it is not possible to obtain bot score for troll accounts. The reason is that Botometer uses Twitter API (application programming interface) to fetch public profile and tweets of an account, and because these troll accounts have already been removed by Twitter, Botometer cannot work and yields error message.

Modeling and evaluation

Our primary goal here is to test whether a simple out-of-the-box classification model can help us to detect the content of coordinated influence operations on social media. Hence, we avoid any model comparison or grid search on hyperparameters and simply train a random forests classifier using the scikit-learn library for Python (*30*). For all classifiers, we set the number of trees in the forest at 1000 (i.e., n_estimators = 1000) and the number of features to consider

when looking for the best split as the square root of the number of potential features (i.e., max_features = "sqrt"). We use scikit-learn's default setting for the rest of hyperparameters. We report macro-weighted precision, recall, and F1 scores using the default classification threshold of 0.5. Our main evaluation metric of interest is the F1 score.

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/6/30/eabb5824/DC1

REFERENCES AND NOTES

1. Z. Papacharissi, Affective publics and structures of storytelling: Sentiment, events and mediality. *Inf. Commun. Soc.* **19**, 307–324 (2016).
2. B. F. Welles, S. J. Jackson, The battle for #Baltimore: Networked counterpublics and the contested framing of urban unrest. *Int. J. Commun.* **13**, 21 (2019).
3. D. A. Martin, J. N. Shapiro, *Trends in Online Foreign Influence Efforts* (Princeton Univ., 2019).
4. F. B. Keller, D. Schoch, S. Stier, J. Yang, Political astroturfing on twitter: How to coordinate a disinformation campaign. *Polit. Commun.* **37**, 256–280 (2020).
5. E. Walker, *Grassroots for Hire: Public Affairs Consultants in American Democracy* (Cambridge Univ. Press, 2014).
6. Elections integrity; https://about.twitter.com/en_us/values/elections-integrity.html#data.
7. D. A. Broniatowski, A. M. Jamison, S. Qi, L. AlKulaib, T. Chen, A. Benton, S. C. Quinn, M. Dredze, Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *Am. J. Public Health* **108**, 1378–1384 (2018).
8. Y. Xia, J. Lukito, Y. Zhang, C. Wells, S. J. Kim, C. Tong, Disinformation, performed: Self-presentation of a Russian IRA account on Twitter. *Inf. Commun. Soc.* **22**, 1646–1664 (2019).
9. A. Arif, L. G. Stewart, K. Starbird, Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proc. ACM Hum. Comput. Interact.* **2**, 20 (2018).
10. R. L. Boyd, A. Spangher, A. Fourney, B. Nushi, G. Ranade, J. Pennebaker, E. Horvitz, Characterizing the internet research agency's social media operations during the 2016 U.S. presidential election using linguistic analyses. *PsyArxiv* 1 October 2018. https://doi.org/10.31234/osf.io/ajh2q.
11. Y. Golovchenko, C. Buntain, G. Eady, M. A. Brown, J. A. Tucker, Cross-platform state propaganda: Russian trolls on twitter and youtube during the 2016 U.S. presidential election. *Intl. J. Press Polit.* , (2020).
12. Y. M. Kim, J. Hsu, D. Neiman, C. Kou, L. Bankston, S. Y. Kim, R. Heinrich, R. Baragwanath, G. Raskutti, The stealth media? Groups and targets behind divisive issue campaigns on Facebook. *Polit. Commun.* **35**, 515–541 (2018).
13. C. A. Bail, B. Guay, E. Malony, A. Combs, D. S. Hillygus, F. Merhout, D. Freelon, A. Vlofovsky, Assessing the Russian internet research agency's impact on the political attitudes and behaviors of American twitter users in late 2017. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 243–250 (2019).
14. A. Badawy, E. Ferrara, K. Lerman, Analyzing the digital traces of political manipulation: The 2016 Russian interference Twitter campaign, in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (IEEE, 2018), pp.58–265.
15. S. Zannettou, B. Bradlyn, E. Deristofaro, G. Stringhini, J. Blackburn, Characterizing the use of images by state-sponsored troll accounts on Twitter. arXiv:1901.05997 (2019).
16. A. Badawy, K. Lerman, E. Ferrara, Who falls for online political manipulation?, in *Companion Proceedings of the 2019 World Wide Web Conference* (ACM, 2019), pp. 62–168.
17. D. Stukal, S. Sanovich, R. Bonneau, J. A. Tucker, Detecting bots on Russian political twitter. *Big Data* **5**, 310–324 (2017).
18. O. Varol, E. Ferrara, C. A. Davis, F. Menczer, A. Flammini, Online human-bot interactions: Detection, estimation, and characterization, in *Eleventh International AAAI Conference on Web and Social Media* (ICWSM, 2017).
19. O. Varol, E. Ferrara, F. Menczer, A. Flammini, Early detection of promoted campaigns on social media. *EPJ Data Sci.* **6**, 13 (2017).
20. D. Pacheco, P.-M. Hui, C. Torres-Lugo, B. T. Truong, A. Flammini, F. Menczer, Uncovering coordinated networks on social media. arXiv:2001.05658 [cs.SI] (16 January 2020).
21. B. Ghanem, D. Buscaldi, P. Rosso, TexTrolls: Identifying Russian trolls on Twitter from a textual perspective. arXiv:1910.01340 [cs.CL] (3 October 2019).
22. J. Im, E. Chandrasekharan, J. Sargent, P. Lighthammer, T. Denby, A. Bhargava, L. Hemphill, D. Jurgens, E. Gilbert, Still out there: Modeling and identifying Russian troll accounts on Twitter. arXiv:1901.11162 [cs.SI] (31 January 2019).
23. J. Lukito, Coordinating a Multi-Platform Disinformation Campaign: Internet Research Agency Activity on Three US Social Media Platforms, 2015 to 2017. *Pol. Commun.* **37**, 238–255 (2020).
24. Report of the Select Committee on Intelligence United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 2; Russia's Use of Social Media; https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf [accessed 3 March 2020].
25. D. L. Linvill, P. L. Warren, Troll factories: Manufacturing specialized disinformation on Twitter. *Polit. Commun.* **2020**, 1–21 (2020).
26. S. Aral, D. Eckles, Protecting elections from social media manipulation. *Science* **365**, 858–861 (2019).
27. N. Gleicher, Adverarial design (2019); https://www.youtube.com/watch?time_continue=20&v=EGlxgxvzPqg.
28. P. Barberá, A. Casas, J. Nagler, P. J. Egan, R. Bonneau, J. T. Jost, J. A. Tucker, Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data. *Am. Pol. Sci. Rev.* **113**, 883–901 (2019).
29. List of political subreddits; https://www.reddit.com/r/redditlists/comments/josdr/list_of_political_subreddits/ [accessed 27 February 2020].
30. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
31. M. Cataldi, L. Di Caro, C. Schifanella, Emerging topic detection on twitter based on temporal and social terms evaluation, in *Proceedings of the Tenth International Workshop on Multimedia Data Mining* (ACM, 2010), p. 4.
32. J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, J. Leskovec, Can cascades be predicted?, *in Proceedings of the 23rd International Conference on World Wide Web* (ACM, 2014), pp. 925–936
33. D. Freelon, M. Bossetta, C. Wells, J. Lukito, Y. Xia, K. Adams, Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Soc. Sci. Comp. Rev.*, 0894439320914853 (2020).
34. H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, S. Mishra, Analyzing labeled cyberbullying incidents on the instagram social network, in *International Conference on Social Informatics* (Springer, 2015), pp. 49–66.
35. A. Massanari, #Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media Soc.* **19**, 329–346 (2017).
36. Open Hearing: Social Media Inuence in the 2016 U.S. Election, United States Senate Select Committee on Intelligence Testimony of Sean J. Edgett Acting General Counsel, Twitter, Inc. (2017); https://www.intelligence.senate.gov/sites/default/files/documents/os-sedgett-110117.pdf.
37. Questions for the Record, Senate Select Committee on Intelligence Hearing on Social Media Inuence in the 2016 U.S. Elections (2017); https://www.intelligence.senate.gov/sites/default/files/documents/Twitter%20Response%20to%20Committee%20QFRs.pdf.
38. O. Tsur, A. Rappoport, What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities, in *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (ACM, 2012), pp. 643–652.
39. Twitter, Update on Twitter's Review of the 2016 U.S. Election (2018); https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html.
40. Twitter, Retrospective Review Twitter, Inc. and the 2018 Midterm Elections in the United States (2019); https://cdn.cms-twdigitalassets.com/content/dam/blog-twitter/official/en_us/company/2019/2018-retrospective-review.pdf.
41. L. Yin, F. Roscher, R. Bonneau, J. Nagler, J. A. Tucker, "Your friendly neighborhood troll: The Internet Research Agency's use of local and fake news in the 2016 U.S. presidential campaign" (SMaPP Data Report, Social Media and Political Participation Lab, New York University, 2018).

Acknowledgments: We thank R3 for suggesting task 5. We thank seminar participants at NYU and Princeton and the Data Science and Conflict workshop at Columbia University for their helpful feedback. P. Borysov, N. Evans, R. Iyengar, B. Kettler, J. Kedziora, J. Mayer, M. Nguyen, M. Salganik, G. Shiffman, B. Stewart, C. White, and three anonymous reviewers all shared specific ideas that we have tried to incorporate. We thank B. O'Hara, J. Oledan, J. Shipley, J. Tait, and K. Yadav for outstanding research assistance. All errors are our own. **Funding:** This work was possible, thanks, in part, to generous funding from the Bertelsmann Foundation and Microsoft. The NYU Center for Social Media and Politics—where J.A.T. is co-director and C.B. is a faculty research affiliated—is supported by the Knight Foundation, the Charles Koch Foundation, Craig Newmark Philanthropies, the Hewlett Foundation, NYU's Office of the Provost and the Global Institute for Advanced Study, and the Siegel Family Endowment. Publication fees were partially covered by the Princeton Open Access Publication Fund Program. **Author contributions:** J.N.S. and J.A.T. secured funding. M.A. formulated the problem, designed and developed the classifier, conducted all analyses, and wrote the initial draft of the manuscript. M.A. and J.N.S. contributed to the experimental design and revised the manuscript. All authors contributed in data collection and interpretation of the results.

Competing interests: M.A. and J.N.S. are inventors on a patent application related to this work filed by Princeton University (no. 62/992,551, filed on 20 March 2020). The authors declare no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Influence operations datasets are publicly accessible. Replication data for this study are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PMY0KF.

Submitted 4 March 2020
Accepted 9 June 2020
Published 22 July 2020
10.1126/sciadv.abb5824

Citation: M. Alizadeh, J. N. Shapiro, C. Buntain, J. A. Tucker, Content-based features predict social media influence operations. *Sci. Adv.* **6**, eabb5824 (2020).