Coordinated Reply Attacks in Influence Operations: Characterization and Detection

Manita Pote*[1]*, Tuğrulcan Elmas[2], Alessandro Flammini[1], Filippo Menczer[1]

[1] Indiana University Bloomington
[2] University of Edinburgh

Abstract

Coordinated reply attacks are a tactic observed in online influence operations and other coordinated campaigns to support or harass targeted individuals, or influence them or their followers. Despite its potential to influence the public, past studies have yet to analyze or provide a methodology to detect this tactic. In this study, we characterize coordinated reply attacks in the context of influence operations on Twitter. Our analysis reveals that the primary targets of these attacks are influential people such as journalists, news media, state officials, and politicians.

We propose two supervised machine-learning models, one to classify tweets to determine whether they are targeted by a reply attack, and one to classify accounts that reply to a targeted tweet to determine whether they are part of a coordinated attack. The classifiers achieve AUC scores of 0.88 and 0.97, respectively. These results indicate that accounts involved in reply attacks can be detected, and the targeted accounts themselves can serve as sensors for influence operation detection.

Introduction

Social media platforms are the primary environments in which civic engagement takes place. They play an important role in the exchange of ideas, discussion of political agendas, and development of political identities thanks to the ease with which one can access and consume information and build influence. However, social media platforms are also exploited by coordinated groups to purposefully distribute misleading information (Weedon, Nuland, and Stamos 2017), artificially amplify certain content (Elmas, Overdorf, and Aberer 2022), or interfere with elections (Ferrara et al. 2020; Office of the Director of National Intelligence 2017; Neudert, Howard, and Kollanyi 2019). These types of social media exploitation are referred to as *information operations* or *influence operations* (IOs).

Influence operations are organized attempts to achieve a specific effect, such as manipulating public opinion, usually through coordinated tactics (Pamment and Smith 2022). IO tactics include public relations via advertising or paid digital influencers (Ong and Cabañes 2018); hashtag hijacking to distort trends and attract or distract the attention of mainstream media (Ong and Cabañes 2018); use of inauthentic and automated accounts to create the appearance of popularity (Elmas 2023; Woolley and Howard 2018); deletion of content violating terms of service to avoid detection by platforms (Torres-Lugo et al. 2022); troll accounts (Zannettou et al. 2019a); the spread of disinformation and propaganda (Woolley and Howard 2018); political memes (Rowett 2018; Zannettou et al. 2020; Ng, Moffitt, and Carley 2022); and 'kompromat' strategies to influence political events (Woolley and Howard 2018).

IOs can be state-sponsored, and originate domestically or in a foreign state (Bradshaw and Howard 2017). A prime example of foreign-initiated campaign was the effort to interfere in the 2016 US Presidential Election by the Russian Internet Research Agency (IRA) (Senate Select Committee on Intelligence 2019). Reports on IOs from different countries like China, Brazil, and Nigeria show that such campaigns have emerged as a global threat (Bradshaw and Howard 2017; Woolley and Howard 2018; Bush 2020).

Here we focus on *coordinated reply attacks,* where a group of accounts work together to target specific individuals or entities by flooding their posts with replies. This can be done to overwhelm the target, push a particular narrative, or generate engagement. Coordinated reply attacks are actively employed in influence operations (Matthews and Goerzen 2019; Bush 2020). Such a tactic has been used for harassment, as observed for example in a hate-speech campaign against Mehreen Faruqi, Australia's first female Muslim senator (Thomas, Thompson, and Wanless 2020); amplification by inauthentic accounts (Weedon, Nuland, and Stamos 2017); and spamming, trolling, and incitement.

In this paper, we provide the first quantitative, large-scale study of coordinated reply attacks in influence operations reported by Twitter.[1] We explore the targeting patterns of IO actors employing this tactic and introduce methods to detect the targets of these attacks and the actors involved. We pose the following research questions:

- **RQ1**: Who are the targets of coordinated replies, and what specific topics characterize the tweets that attract such coordinated responses?

---

[1] Although Twitter is now called X, we use the previous name because the data analyzed here predates the name change.

- **RQ2**: Among a set of tweets by potential targets, how can we identify those that receive coordinated replies?
- **RQ3**: Given a set of targeted tweets, how can we detect the accounts that participate in coordinated replies?

We make the following contributions:

- We find that primary targets of coordinated reply attacks are mostly influential people such as journalists, news media, state officials, and politicians. Most of the targets are attacked only once. The attacks for most of the targets are sporadic and tend to focus on specific contexts, such as politics. The attackers can originate from within the target's country or from foreign states.
- We present a classifier model to identify tweets that are targeted by coordinated replies. This model is generalizable to other contexts, as it does not use any features specific to IOs. It can also be developed into a tool for monitoring and safety.
- We present a second model that performs well on the task of detecting accounts that are involved in reply attacks.

Related Work

Given the dearth of prior research on coordinated reply attacks, we review the literature on IOs in general.

Characterization of Influence Operations

Influence operations present novel challenges to content moderation on social media. A crucial initial step to address these challenges is to characterize how IO actors operate: their tactics, motivations, and engagement patterns. Matthews and Goerzen (2019) present different trolling techniques used in social media, from dogpiling to sock puppetry, along with interventions. Zannettou et al. (2019a) observe that Russian trolls displayed different behavior in the use of Twitter compared to random users. The same authors also find that Russian trolls on Twitter and Reddit were pro-Trump, while Iranian trolls were anti-Trump (Zannettou et al. 2019b). The images shared by Russian trolls appeared in many popular social networks as well as mainstream and alternative news outlets, and focused on Russia, Ukraine, and the USA (Zannettou et al. 2020). Dutt, Deb, and Ferrara (2018) analyze the advertisements purchased by IRA accounts on Facebook and identify their changing campaign targets over time by performing clustering and semantic analysis. Stewart, Arif, and Starbird (2018) investigate the behavior of Russian trolls around the #BlackLivesMatter movement and find that the trolls infiltrated both right- and left-leaning political communities to participate in both sides of the discussion. Farkas and Bastos (2018) manually annotate IRA-linked tweets into 19 different categories to study whether IRA operations are consistent with classic propaganda models. Merhi, Rajtmajer, and Lee (2023) find that the accounts involved in an IO in Turkey were resilient to large-scale shutdown. Elmas, Overdorf, and Aberer (2023) discover that IO actors and other adversarial accounts often change their names and assume new identities. The Stanford Internet Observatory (2021) produced several reports describing influence campaigns and a range of tactics used by IO actors, including coordinated reply attacks (Bush 2020). While that work provides a qualitative description of the tactic, here we focus on methods to detect it.

Coordinated reply attacks are also carried out by automated accounts; social bots have been reported to target influential users in an attempt to direct their attention toward fake news (Shao et al. 2018). Financial rather than political incentives may be the drivers of such tactics, as in the case of cryptocurrency manipulation (Yang and Menczer 2024). The methods presented here are context-independent and therefore could be applied to these kinds of campaigns.

Detection of Influence Operations

Many supervised machine-learning models have been proposed in the literature to detect IO actors, especially IRA trolls on Twitter, using deceptive linguistic cues (Addawood et al. 2019) and behavioral and linguistic features (Im et al. 2020). Luceri, Giordano, and Ferrara (2020) propose an inverse reinforcement learning model for this task. Alizadeh et al. (2020) build a content-based classifier to detect tweets from troll accounts in Russia, China, and Venezuela IO campaigns. Work from Sharma et al. (2021) uses a generative model to learn hidden group behavior to identify coordinated accounts. Ezzeddine et al. (2023) present an LSTM-based approach that identifies troll accounts based on behavioral cues. Kong et al. (2023) propose an interval-censored transformer Hawkes architecture to identify IO operators.

Our work is similar to the above-mentioned efforts in the use of a supervised learning approach to identify coordinated accounts. However, we design features that leverage the targeting behaviors of the IO actors, specifically focusing on reply/comment engagements. Our method does not use any IO-specific features or sentiment cues, therefore it can be generalized to different social media platforms that have similar engagement functionalities.

Influence operations are one kind of coordinated campaign. A body of work has explored unsupervised methods to detect coordinated behaviors in general. Pacheco et al. (2021) introduced a network-based framework for coordination detection. As campaigns use more than one tactic at a time, Uyheng, Cruickshank, and Carley (2022) present a multi-view modularity clustering method. A Bayesian model by Hudson Smith, Ehrett, and Warren (2024) leverages similarities in narrative and account characteristics. Nwala, Flammini, and Menczer (2023) propose a language framework that represents user actions and content as sequences of symbols to find coordinated accounts. Unlike the above methods, we do not cluster accounts based on similar behaviors. We classify individual posts based on aggregate features of their replies, and individual accounts based on their metadata and reply activity.

Data Collection

For the present study of coordinated reply attacks, we use 43 different state-sponsored IO datasets released by the Twitter Moderation Research Consortium from October 2018 to

December 2021.[2] These datasets are archives of suspended accounts that Twitter claims to have been involved in foreign influence operations. Along with the account metadata, the datasets provide all the tweets generated by the accounts.

Since according to Twitter the campaigns in these datasets are coordinated by a single entity, they provide us with ground truth for our study. Indeed, if we observe mass replies to a single target from multiple accounts labeled by Twitter as coordinated, we can establish that the coordinated reply attack tactic has been used.

Target Dataset

First, we merge the datasets of all the IOs and keep all the replies by IO accounts to tweets by non-IO accounts. We refer to the latter accounts as *targets* and to the replies by IO accounts as *IO replies*. There are in total 17,873,714 IO replies from 44,425 IO accounts targeting 15,256,547 tweets by 1,763,084 distinct targets. From this data, we extract 15,016 targets and 96,041 tweets that received five or more direct replies from IO accounts. We assume these tweets have been targeted by coordinated reply attacks, and we label them as *targeted tweets*. The threshold of five or more replies is arbitrary; a robustness analysis shows that the detection of targeted tweets does not seem to be affected by this parameter, as discussed later (Fig. 7).

The targeted tweets can still be publicly available at the time of our analysis, allowing us to collect all of their replies. Some of these replies may have originated from non-IO repliers, both before and after the IO accounts were taken down by Twitter. We refer to these replies as *normal replies* and to their authors as *normal repliers*. Since we only have direct replies by IO accounts, we only consider direct replies by normal repliers as well; replies to replies are discarded. In addition to the metadata about the IO replies that are present in the initial data, we query the `/users/:id`, `/search/all`, and `/tweets/?ids=` endpoints of the Twitter API[3] to collect metadata about the targets, the targeted tweets, their normal replies, and the normal repliers.

Among the 15,016 targets, 5,041 were suspended, 3,992 could not be found (possibly deleted accounts), and 5,983 were alive at analysis time (2,031 verified and 3,952 non-verified accounts). Of the total 96,041 targeted tweets, 43,048 could not be found, which means these tweets could have originated from deleted or suspended accounts; 18,808 had unauthorized access; and 34,185 were accessible. For our influence operation case studies (**RQ1**), we consider this *target dataset* of 34,185 tweets by 5,983 targets.

Classification Dataset

To identify tweets targeted by coordinated replies (**RQ2**) and accounts that participate in such activity (**RQ3**), we consider targeted tweets as our positive examples. For corresponding negative examples, we collect *control tweets* posted by the

---

Figure 1: Data collection for the classifiers. The dashed line separated the last IO reply and the first successive tweet by the target.

same targets after the last IO reply. This ensures that the tweets in our control data did not receive any coordinated replies by IO accounts. Fig. 1 illustrates the data collection.

As in the case of positive examples, we only retain control tweets with five or more replies. In addition, to avoid bias due to the diverse activity of the targets, we collect from each target as many control tweets as targeted tweets. Specifically, we select control tweets that were posted immediately after the last IO reply, subject to the five-reply minimum. In cases where we could not obtain as many control tweets as targeted tweets, we ensure a balanced dataset by keeping the most recent targeted tweets.

Similar to the targeted tweets, we fetch all the replies to the control tweets and all replier metadata. The resulting *classification dataset* includes 3,866 targeted tweets and the same number of control tweets by 1,507 targets. There are in total 881,918 and 323,378 repliers in the positive and negative examples, respectively. These include IO and normal repliers. While the full classification dataset is used for **RQ2**, for **RQ3** we only use the positive examples (targeted tweets): 7,670 IO repliers and 874,248 normal repliers.

RQ1: Targets and Topics

In this section, we present an exploratory analysis of the targets of coordinated reply attacks and two case studies of specific campaigns where we can analyze the targets as well as the topics of their targeted tweets and other tactics employed in the campaigns.

Exploratory analysis of target metadata (Fig. 2a) shows that targets tend to have more followers (median 22,540) than following (median 707). This suggests that targets can be influential people. Reply attacks tend to be selective (Fig. 2b): only a few tweets by each target were targeted (median one). The median number of coordinated replies received by targeted tweets was eight (Fig. 2c). However, 54 of the targeted tweets received more than 1,000 replies. Coordinated replies tend to occur quickly after a targeted tweet, with a median delay of 3 hours (Fig. 2d).

To better understand what kinds of accounts were targeted, we annotated some target profiles with the corresponding professions or organization types and country of origin. We used manual annotation by checking each Twitter profile, description, the profession metadata indicated by the
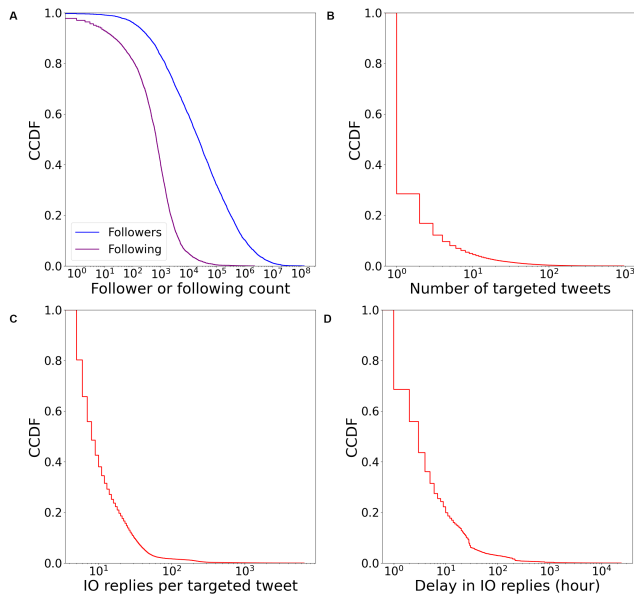
Figure 2: Complementary cumulative distribution functions (CCDF) of statistics describing target accounts and their tweets. (A) Numbers of followers and following (friends) of targets. (B) Number of targeted tweets per target. (C) Number of coordinated replies received by each targeted tweet. (D) Time delay between targeted tweets and their coordinated replies.

'briefcase icon,' and by searching Google for the accounts with more than a million followers. We grouped profession and organization types into broad categories, such as state officials, news media, and politicians. Accounts with insufficient information were labeled 'Not Available.'

As the annotation process was time-consuming, we focused on two cases, namely two of the five campaigns with the most targets: Serbia (the top campaign with 1,175 targets) and Egypt (the fifth campaign with 372 targets). In the next subsections, for each case, we report the top 10 target professions/types and countries. We also inspected the targeted tweets to understand the context of the attacks. In preprocessing, we translated the targeted tweets into English and removed stop words and emojis.

Case Study: Serbia. The majority of accounts targeted by the Serbia campaign, approximately 648, were from Serbia itself, with the remaining coming from the Balkan region (Fig. 3a). This suggests that the campaign focused its efforts on influencing public opinion within Serbia. Fig. 3b reveals that the coordinated reply attacks primarily targeted journalists (102), state officials (99), news media organizations (76), and politicians (43). A wordshift graph (Gallagher et al. 2021) highlighting the most prominent terms in the targeted tweets (Fig. 3c) shows that the campaign focused on President Vucic, the Serbian Progressive Party (SNS), the 2017 election, the "1 out of 5 Millions" protest, and the Serbia-Kosovo diplomatic crisis. These findings are consistent with analysis by Bush (2020), who reported that the primary objective of IO actors involved in the Serbia campaign was to rally support for President Alexander Vu-
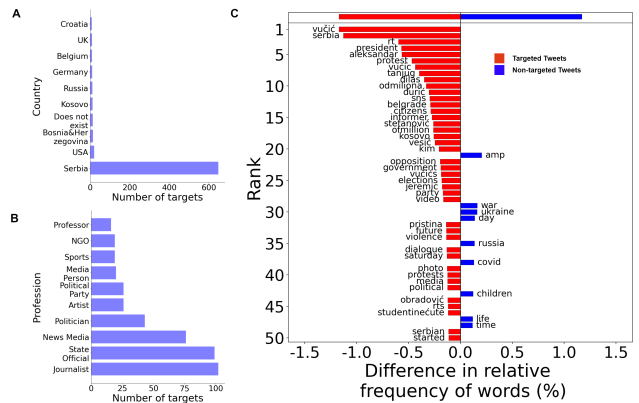

Figure 3: Characterization of the Serbia campaign. Distributions of (A) countries and (B) professions of the targets. (C) Wordshift graph comparing the most frequent words in targeted and non-targeted tweets.

cic and his party, the SNS. This was achieved by promoting the popularity and visibility of Vucic and the SNS through retweeting their content and replying to other accounts with supportive messages. The IO accounts also targeted opponent political parties with derisive tweets and attempted to discredit them by flooding their posts with negative comments. This tactic aimed to create a public perception that the opposition was unpopular.

Case Study: Egypt. Fig. 4a shows that the majority of accounts targeted by the Egypt campaign were from multiple Middle East and North Africa countries, primarily Saudi Arabia (74 targets), Egypt (39), UAE (36), Qatar (30), and Yemen (26). This suggests a potential interstate attack. News media organization (67), journalists (52), and state officials (29) were again the main targets of the coordinated replies (Fig. 4b). The analysis of common terms in the targeted tweets (Fig. 4c) and manual inspection reveal that the Egypt campaign primarily focused on religious themes, terrorism, and current affairs like the Iran Nuclear deal (2018), Yemen's Houthi movement, Sudan's military coup, and the Muslim Brotherhood. These observations are consistent with a report by DiResta, Kheradpir, and Miller (2020), describing an IO activity orchestrated by Egypt and the UAE, supporting the Saudi and Egyptian governments and criticizing Qatar, Turkey, Yemen, Iran.

Both case studies indicate that influential people like journalists, news media, state officials, and politicians, are the primary targets of coordinated reply attacks. These targets can be from different countries than the campaign's country of origin. The topics of the targeted tweets depend on the current affairs of the specific geographic region or country.

RQ2: Tweet Classification

Identifying the tweets that receive inauthentic coordinated replies is a necessary first step toward the detection of both the targets and the perpetrators of a coordinated attack. To address this challenge, we propose a campaign-independent classifier for identifying IO-targeted tweets.
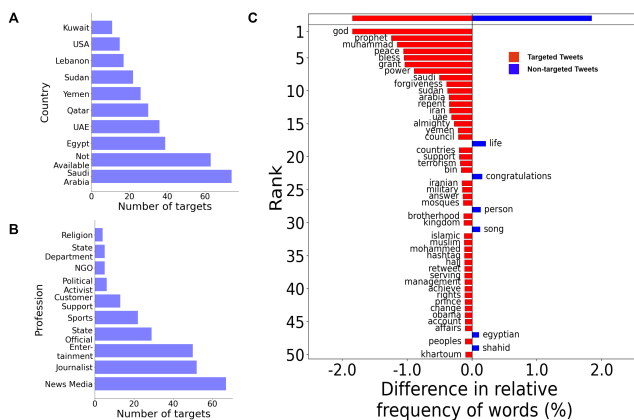
Figure 4: Characterization of the Egypt campaign. Distributions of (A) countries and (B) professions of the targets. (C) Wordshift graph comparing the most frequent words in targeted and non-targeted tweets.

Table 1: Reply-level attributes used to generate features for the tweet classifier.

Set	Attributes
Engagement	`like_count`
	`retweet_count`
	`reply_count`
Entities	`mention_count`
	`hashtag_count`
	`url_count`
Delay	`reply_time_diff`
Similarity	`cosine`

The same methodology could also be generalized to platforms other than Twitter.

Classifier Features

The tweet classifier leverages several features extracted from tweets and from the replies they receive. Let us first focus on tweet-level features, specifically tweet engagement. We find a few key differences between the engagement metrics of IO-targeted vs. control tweets. As illustrated in Fig. 5a, IO-targeted tweets receive more replies (median 31 vs. 22 for control tweets). On the other hand, control tweets receive slightly more retweets (median 84 vs. 75 for IO-targeted tweets, Fig. 5b) and more likes (median 420 vs. 250, Fig. 5c). This suggests that organic engagement generated more positive interactions and sharing, while inauthentic activity mainly focused on manipulating conversations through replies. Based on these observations, we use three tweet-level features: `reply_count`, `retweet_count`, and `like_count`.

Next, let us consider reply-level features. These are based on eight attributes, listed in Table 1. Engagement and entity attributes are defined for each reply. The delay is also defined, for each reply, as the difference between the timestamps of the tweet and the reply. The similarity is designed to capture the presence of similar narratives in replies, a common characteristic of inauthentic engagement. To this end, we first generate vector embeddings for the replies using the LaBSE model (Feng et al. 2020), which supports 109 languages. The `cosine` attribute is then computed for each pair of replies to the same tweet as the cosine similarity between the corresponding vectors.

Since targeted tweets can have many replies, this procedure yields many attribute values that must be aggregated to obtain a set of features for each tweet. In the case of engagement, entities, and delay attributes, we have one value per reply. For the similarity attribute, we have one value per a pair of replies. In all cases, we aggregate these values to obtain a single distribution of attribute values for each tweet. From these distributions we compute the following 12 summary statistic features: range, 25/50/75 quartiles, interquartile range, minimum, maximum, mean, standard deviation, skewness, kurtosis, and entropy. Since we do this for each of eight attributes, the total number of reply-level features used in the classifier is $8 \times 12 = 96$. Including the three tweet-level features, the classifier uses a total of 99 features.

Results

We compare different machine learning models: Logistic Regression, Random Forest, AdaBoost, Decision Tree, and Naive Bayes. Prior to training, we standardize the input features via z-scores. We conduct 10-fold cross-validation to mitigate over-fitting of the training data and report on the mean precision, recall, and F1 values across folds along with AUC in Table 2. Precision, recall, and F1 depend on a threshold to transform the model score into a binary classification label. We tune the threshold to maximize the mean F1 across folds. In the following, we focus on Random Forest (with 100 estimator trees), which yields the best scores overall.

To study the contributions of different features, we followed two approaches. First, we trained and tested Random Forest on individual tweet-level features and reply-level feature sets. The results using 10-fold cross-validation are given in Table 3. Second, we performed a permutation feature importance test, which measures the importance of features by computing the loss in accuracy when the values of those features are shuffled (permuted). To simplify the analysis, for each reply-level attribute we shuffled all the corresponding features rather than each feature individually. For example, for the `like_count` engagement attribute, we shuffled all 12 summary statistics features at once. We repeated this test 10 times and recorded the drop in mean F1 score from 10-fold cross-validation for each iteration. The distribution of these drop values is given in Fig. 6.

Both approaches consistently show that reply-level engagement features are the most important. A classifier using only those features achieves F1=0.77 and AUC=0.84 (Table 3), and removing those features causes significant drops in F1 (Fig. 6). In our classification dataset, the majority of targets are from the Serbia campaign. IO accounts in this campaign were not intended to generate engagement with other Twitter users; instead, they primarily boosted retweet and reply counts for other IO accounts to artificially amplify Vucic and his allies on Twitter (Bush 2020). In fact, we find that replies to targeted tweets from IO accounts have a
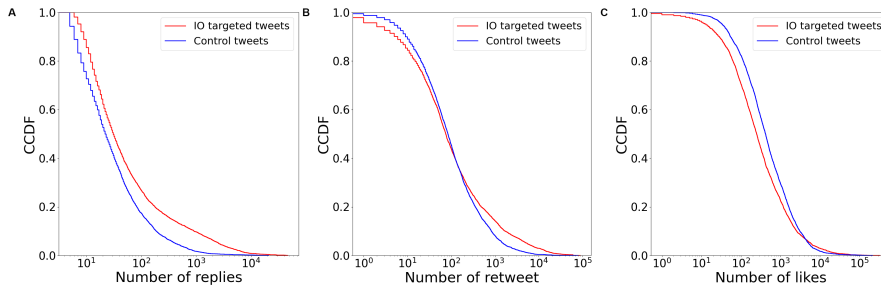
Figure 5: Engagement received by targeted and control tweets. (A) Replies, (B) retweets, and (C) likes.

Table 2: Results of different algorithms in the tweet classification task. We present standard errors rounded to the second decimal point.

Classifier	Prec.	Rec.	F1	AUC
Logistic Regression	0.65 ± 0.00	0.86 ± 0.00	0.74 ± 0.00	0.80 ± 0.00
Random Forest	0.73 ± 0.00	0.87 ± 0.00	0.80 ± 0.00	0.88 ± 0.00
AdaBoost	0.64 ± 0.00	0.89 ± 0.00	0.74 ± 0.00	0.81 ± 0.00
Decision Tree	0.52 ± 0.01	0.95 ± 0.02	0.66 ± 0.00	0.69 ± 0.00
Naive Bayes	0.49 ± 0.00	1.00 ± 0.00	0.66 ± 0.00	0.68 ± 0.00

Table 3: Contributions of different tweet-level features and reply-level feature sets to the Random Forest tweet classifier. The last row (using all features) corresponds to the results in Table 2.

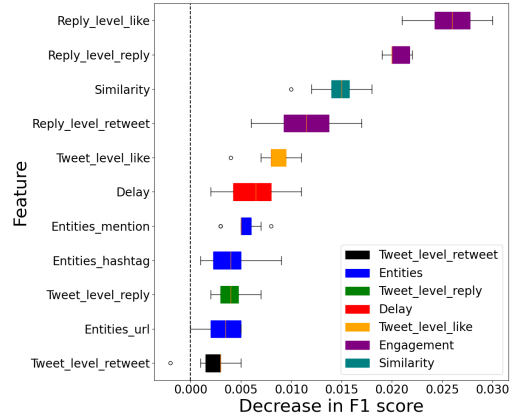| Features set	Prec.	Rec.	F1	AUC
`reply_count`	0.5	0.99	0.67	0.59
`retweet_count`	0.49	1	0.66	0.54
`like_count`	0.49	1	0.66	0.52
Engagement	0.69	0.86	0.77	0.84
Entities	0.52	0.95	0.67	0.65
Delay	0.51	0.96	0.67	0.66
Similarity	0.54	0.96	0.69	0.68
All features	0.73	0.87	0.80	0.88

Figure 6: Permutation feature importance for tweet classifier. We report the median (orange line), 50% confidence interval (box), and 99.3% confidence interval (whiskers) of the drop in F1 score when each feature/attribute is shuffled. Boxes with the same color indicate attributes in the same feature set. Larger values indicate higher importance.

higher mean retweet count than replies from normal repliers (0.63 vs 0.31) and also a higher mean like count (1.28 vs 0.81) and mean reply count (0.20 vs 0.13). However, our data does not allow us to determine if such engagement was mostly driven by IO accounts or organic.

Since reply-level engagement may be affected by the popularity of the targeted tweets, it is legitimate to ask whether tweet-level engagement features would provide sufficient signals to discriminate between targeted and control tweets. However, Table 3 indicates that tweet-level reply, like, and retweet counts do not provide very informative signals for tweet classification. To further explore this question, let us measure the correlation between tweet-level features (`reply_count`, `retweet_count`, and `like_count`) and the corresponding reply-level engagement counts. As each original tweet can have many replies, there are many more replies than original tweets. We therefore calculate the mean correlation between pairs of tweet/reply engagement features across 10 random samples of replies matching the number of original tweets. The correlations are all very small (around 0.001), confirming that reply engagement is not a mere reflection of tweet popularity.

We previously defined *targeted tweets* as those that re-

ceive five or more replies from IO accounts. Let us test the robustness of our classifier with respect to this definition by considering a range of threshold values between five and 20 replies from IO accounts. This filters down the set of targeted tweets and corresponding control tweets. We follow the same procedure described above to construct the classification dataset, extract the features, and train and evaluate the classifier. Fig. 7 reports the mean precision, recall, F1, and AUC from 10-fold cross-validation. While we observe slight increases as the criterion for defining targeted tweets becomes more stringent, the results appear to be robust with respect to this parameter.

Next, let us evaluate the generality of the classifier by testing how well a model trained on one campaign performs
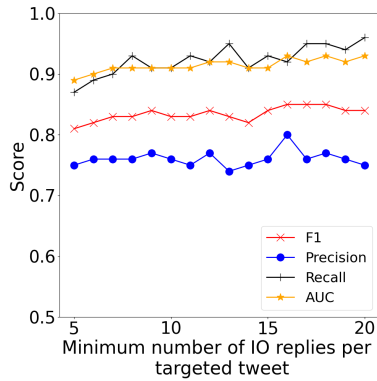
Figure 7: Scores of tweet classifiers based on different thresholds for the number of IO replies received by targeted tweets.

Table 4: F1 scores obtained from same-campaign (diagonal entries, in bold) and cross-campaign evaluations of the tweet classifier. RS=Serbia, SA=Saudi Arabia, TR=Turkey, EG=Egypt. SA/EG/AE is a campaign involving three countries.

Train	Test					
	RS	SA	TR	EG	SA/EG/AE	Other
RS	**0.85**	0.54	0.61	0.56	0.52	0.65
SA	0.55	**0.76**	0.53	0.63	0.74	0.68
TR	0.61	0.60	**0.74**	0.63	0.65	0.68
EG	0.43	0.37	0.36	**0.65**	0.38	0.39
SA/EG/AE	0.47	0.58	0.57	0.60	**0.73**	0.48
Other	0.62	0.59	0.63	0.56	0.52	**0.74**

when tested on other campaigns. First, we split the classification dataset into six subsets: one for each of the top five campaigns, based on the number of targeted tweets, and one with data aggregated from the remaining campaigns. Second, we train campaign-specific models on each of these datasets, as in the original tweet classification setup. Finally, we evaluate the models on test data from each dataset. In the diagonal of Table 4 we report F1 values when the model trained on one campaign is tested on the same campaign (mean across 10-fold cross-validation). The off-diagonal F1 values are obtained when the model trained on all data from one campaign (optimized to maximize F1) is tested on other campaigns. As expected, the models perform better when trained and tested on the same campaign. However, models can generalize, with F1 drops that depend on the specific campaigns. This suggests that at least some commonalities exist across coordinated reply campaigns.

RQ3: Replier Classification

Once we identify potentially targeted tweets, we can attempt to detect, among the accounts that reply to them, those that are engaged in coordinated activity. Distinguishing authentic replies from inauthentic ones poses a non-trivial challenge, given that inauthentic accounts attempt to create an impression of authenticity. For this task, we train a supervised replier classifier using the targeted tweet dataset.
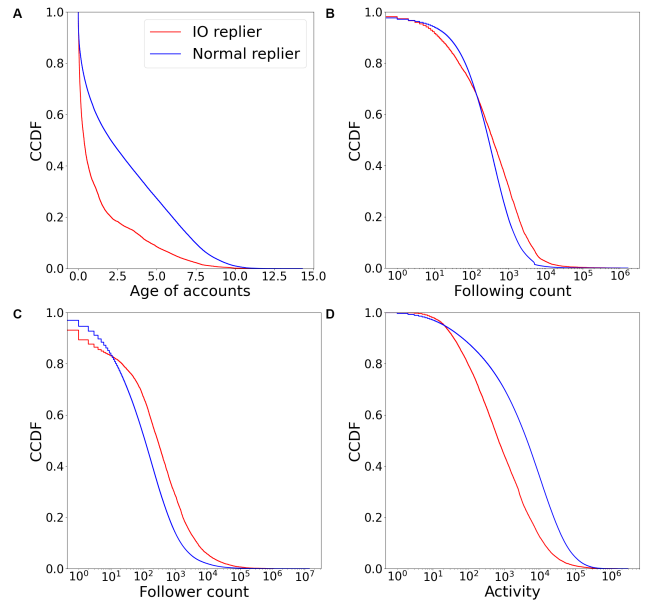

Figure 8: Differences between complementary cumulative distributions of IO and normal replier metadata: (A) age, (B) following count, (C) follower count, and (D) activity, as measured by the sum of the numbers of original tweets, replies, quotes, and retweets.

Classifier Features

We engineer features for each replier from their profile metadata and their replies to the targeted tweets. Starting with profile metadata, we calculate the age of repliers by subtracting the account creation date from the date of the last reply by the account. As illustrated in Fig. 8(A), most IO repliers are relatively new accounts, with a median age of 0.37 years compared to 2.08 years for normal repliers. Despite their relatively young age, IO repliers display a higher median number of followers (282) and followings (380) compared to normal repliers, whose medians are 114 and 292, respectively (Figure 8(B, C)). However, IO repliers exhibit lower activity levels (original tweets + replies + quotes + retweets), with a median of 699 compared to 4406 for normal repliers (Figure 8(D)).

To leverage these key differences between IO and normal repliers, we create four features specific to profile metadata: `age`, `follower_rate`, `following_rate`, and `activity_rate`. Since the numbers of followers/following and the activity are correlated with an account's age, we normalize the rate features by the age of the account.

Each replier may be involved in one or more replies to multiple targeted tweets. Therefore, we create a number of features that summarize the characteristics of the set of replies generated by each replier, including replies to multiple targeted tweets. These features are based on eight reply attributes, which we organize into four sets, just like those listed in Table 1. The only criterion that distinguishes how these features are calculated in the tweet versus the replier classification task is the reply set — all replies to a tweet in the former case and all replies by a replier in the latter.

Given a set of replies, the replier classifier features are calculated as for the tweet classifier, with two exceptions. First, the delay of each reply is computed with respect to the timestamp of the targeted tweet to which the reply was directed. Second, cosine similarity s for replier i is calculated for each pair (r_i^t, r_j^t) where r_i^t is a reply by i to a targeted tweet t and r_j^t is a reply by a different user j to the same targeted tweet t. We obtain a distribution of these similarities $\cup_{t \in T(i)} \cup_{j \in J(t)} s(r_i^t, r_j^t)$ across the set $J(t)$ of other users who reply to t and then across the set $T(i)$ of targeted tweets that receive a reply from i.

From the distribution of each attribute, we compute nine summary statistic features: range, 25/50/75 quartiles, interquartile range, maximum, minimum, mean, and entropy. We do not calculate standard deviation, skewness, and kurtosis because they are not defined for many repliers who are involved in a single reply to a single targeted tweet.

We end up with four profile metadata features and $8 \times 9 = 72$ reply-level features, for a total of 76 features.

Results

The targeted tweet dataset is highly imbalanced with 0.8% IO repliers (7,670 vs. 874,248 normal repliers). Such an imbalance leads to poor classification, which can be addressed in two ways. First, we downsampled the normal repliers by creating 10 different balanced datasets. Each includes all the IO repliers and an equal number (7,670) of normal repliers, sampled without replacement. We train and test the model on each balanced dataset using 10-fold cross-validation and report the average performance score. As a second approach, we over sampled the IO repliers by splitting the data into train and test sets, then replicating the minority class. Replication occurs only in the training data, to avoid data leakage. We run 10-fold cross-validation on the resulting dataset. The first approach might eliminate some potential false positives — normal repliers with similar reply behavior — potentially making the task easier. In the second approach, the model is tested on data that still maintains the class imbalance, potentially overfitting the training data. This approach is also more expensive due to the large dataset. Given these complementary disadvantages, below we report on both methods.

We standardize the features with z-scores and report the mean performance metrics obtained by different machine learning models: Logistic Regression, Random Forest, AdaBoost, Decision Tree, and Naive Bayes. As in the tweet classifier, we tune the threshold to maximize the mean F1 across folds. Table 5 shows that all classifiers perform better with downsampling, and Random Forest (with 100 estimator trees) performs the best with both downsampling and oversampling. Therefore, let us focus on this model — Random Forest trained with downsampling — for further analysis.

To test the contribution of each feature to the replier classifier, we follow the same procedure as for the tweet classifier. However, here we report the averages across the 10 balanced dataset. Table 6 reports on mean 10-fold cross-validation scores for Random Forest trained and tested on each profile metadata feature and reply feature set. Fig. 9 reports on the results of a permutation feature importance
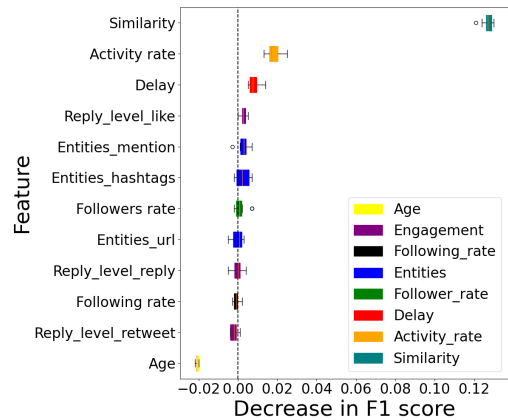

Figure 9: Permutation feature importance for replier classifier. We report the median (orange line), 50% confidence interval (box), and 99.3% confidence interval (whiskers) of the drop in F1 score when each feature/attribute is shuffled. Boxes with the same color indicate attributes in the same feature set. Larger values indicate higher importance.
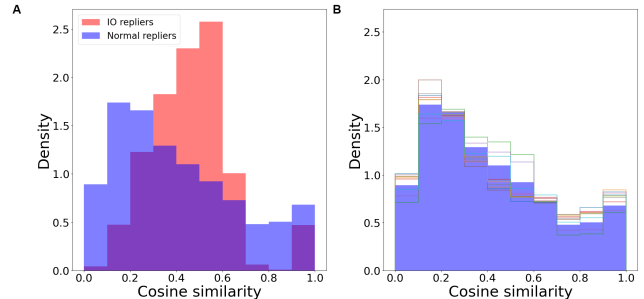

Figure 10: (A) Distributions of cosine similarity attributes for replies by IO and normal repliers. (B) Similarity distributions for normal repliers (blue, same as in panel (A)) and the 10 different samples (different colored outlines).

test. Both analyses consistently shows that the similarity among replies is the most important feature. To help interpret this finding, Fig. 10(A) compares the distributions of similarity attributes for replies by IO versus normal repliers. Replies by IO repliers are more similar to other replies to the same tweets, compared to those by normal repliers. This is a pattern that the classifier can exploit. We also observe in Fig. 10(B) that the downsampling process does not bias the similarity distributions.

So far, we have tested the replier classifier on balanced datasets. In more realistic scenarios, the data may be imbalanced with different ratios of coordinated and organic repliers. To test whether the classifier can generalize to such scenarios, we train and test with 10-fold cross-validation using different positive to negative data ratios ranging from 1:5 to 1:45. Fig. 11 shows that precision and AUC are robust to class imbalance, whereas recall (and consequently F1) drops as the class imbalance increases. While this result indicates that balancing the classes affects recall, we also found that training on the original imbalanced dataset leads to a high

Table 5: Results of different algorithms in the replier classification task. Top: downsampling of the majority class (normal repliers). Bottom: oversampling of the minority class (IO repliers). We present standard errors rounded to the second decimal point.

Downsampling	Prec.	Rec.	F1	AUC
Logistic Regression	0.89 ± 0.00	0.88 ± 0.00	0.88 ± 0.00	0.93 ± 0.00
Random Forest	0.93 ± 0.00	0.92 ± 0.00	0.92 ± 0.00	0.97 ± 0.00
AdaBoost	0.90 ± 0.00	0.90 ± 0.00	0.90 ± 0.00	0.96 ± 0.00
Decision Tree	0.88 ± 0.00	0.88 ± 0.00	0.88 ± 0.00	0.88 ± 0.00
Naive Bayes	0.62 ± 0.02	0.86 ± 0.02	0.68 ± 0.00	0.87 ± 0.00
Oversampling	**Precision**	**Recall**	**F1**	**AUC**
Logistic Regression	0.27 ± 0.00	0.48 ± 0.01	0.35 ± 0.00	0.94 ± 0.00
Random Forest	0.70 ± 0.00	0.72 ± 0.01	0.71 ± 0.00	0.96 ± 0.00
AdaBoost	0.47 ± 0.02	0.54 ± 0.02	0.50 ± 0.01	0.96 ± 0.00
Decision Tree	0.55 ± 0.01	0.51 ± 0.01	0.53 ± 0.01	0.75 ± 0.01
Naive Bayes	0.06 ± 0.00	0.50 ± 0.03	0.10 ± 0.00	0.86 ± 0.00

Table 6: Contributions of different profile features and reply-level feature sets to the Random Forest replier classifier. The last row corresponds to the results in Table 5 (top).

Features set	Prec.	Rec.	F1	AUC
`activity_rate`	0.60	0.63	0.61	0.65
`following_rate`	0.54	0.52	0.53	0.56
`follower_rate`	0.55	0.51	0.53	0.56
`age`	0.58	0.66	0.61	0.66
Delay	0.57	0.60	0.58	0.62
Engagement	0.57	0.57	0.53	0.63
Entities	0.63	0.50	0.53	0.63
Similarity	0.85	0.84	0.84	0.92
All features	0.93	0.92	0.92	0.97

false-positive rate and deteriorated precision.

Next, we test the generalizability of the replier classifier by training and testing the model across different campaigns. Similarly to the tweet classifier, we prepare six campaign datasets: five selected based on the highest number of IO repliers and one by aggregating the remaining campaigns. In the diagonal of Table 7 we report F1 values when the model trained on one campaign is tested on the same campaign (mean across 10-fold cross-validation and 10 different balanced datasets). The off-diagonal F1 values are obtained when the model trained on one of the balanced datasets (selected to maximize F1) is tested on other campaigns. We observe that the campaign-specific models tend to perform well both on their own campaigns and across campaigns. The Serbia campaign is an exception, where we observe a sizable deterioration in cross-campaign evaluation. These results suggest that the features of the replier classifier are relevant across different campaigns.

Discussion

Understanding the engagement patterns of IO operators is crucial for identifying vulnerable individuals and devising effective countermeasures. Our Serbia and Egypt campaign case studies reveal that journalists, state officials, news media, and politicians are primary targets for coordinated IO attacks. These attacks can originate either from within the targeted country or from other nation-states. Our findings further suggest that influential individuals may serve as po-
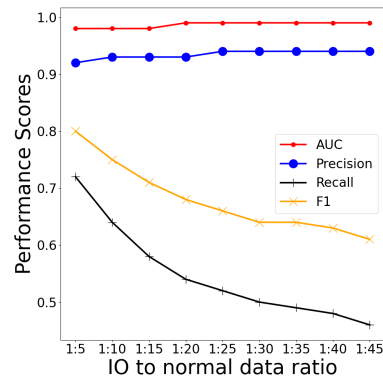

Figure 11: Replier classification scores for different data imbalance ratios.

Table 7: F1 scores obtained from same-campaign (diagonal entries, in bold) and cross-campaign evaluations of the replier classifier. HN=Honduras; see Table 4 for other country codes.

Train	Test					
	SA	RS	TR	EG	HN	Other
SA	**0.96**	0.79	0.86	0.86	0.55	0.90
RS	0.53	**0.89**	0.76	0.53	0.48	0.49
TR	0.90	0.91	**0.92**	0.84	0.84	0.85
EG	0.92	0.90	0.88	**0.92**	0.74	0.90
HN	0.93	0.97	0.98	0.89	**0.95**	0.91
Other	0.95	0.89	0.92	0.89	0.72	**0.94**

tential sensors for identifying IO campaigns.

To detect coordinated reply attacks, we propose a campaign-independent and general machine learning framework consisting of a tweet classifier and a replier classifier. First, the tweet classifier identifies tweets that receive coordinated replies, narrowing the scope for further investigation. This classifier is robust to variations in targeted tweet popularity and general across various campaigns. An analysis of the features used by the classifier indicates that the level of engagement received by the replies is the most distinguishing factor.

Second, the replier classifier identifies operators engaged in coordinated reply attacks. In addition to generalizing across campaigns like the tweet classifier, the replier classifier is also capable of handling different levels of class imbalance. The most significant features for replier classification are those describing the distribution of similarity between replies from a replier and those from other repliers to the same tweets.

This study presents a proof of concept for the proposed classifiers, evaluated on a number of IO campaigns that have been detected and taken down by Twitter. Our experiments were carried out on compute nodes equipped with two 64-core AMD EPYC 7742 2.25 GHz CPUs and 512 GB of RAM. The efficacy, efficiency, and robustness of our classifiers would ideally be validated in the wild. Unfortunately, due to changes in data-sharing policies from X/Twitter, we are unable to conduct such tests.

The proposed framework can be extended to other platforms with similar reply functionalities, including Facebook, Threads, Mastodon, and Bluesky. Additionally, our models could be developed into products or extensions that users could employ for enhanced online safety.

Our study has potential impacts on platform integrity and public dialogue. First, it reveals that what looks like public reactions may in fact be efforts to manipulate a target and other participants of a genuine conversation. For instance, politicians may be posting on social media to solicit public opinions. Coordinated replies may skew their perception of public sentiment (Stewart et al. 2019). Such replies can further distort genuine discourse if portrayed by the media as reflective of public opinion. Secondly, coordinated replies enable malicious actors to maximize the public exposure of their posts by exploiting the popularity of their targets, thereby amplifying their influence. This pollutes online dialogues with spam, influence campaigns, and divisive messages that harass the targeted individuals or provoke the public. Such behavior may also alienate the targets and prevent them from sharing their opinion on social media. For all these reasons, platforms should protect the targets from coordinated reply attacks. The research community may use our methodology to detect coordinated reply attacks and study the campaigns, perpetrators, and their potential effects on the individuals. Our methodology may also guide the development of countermeasures by social media platforms.

Ethical Impact. This study has been granted exemption from Institutional Review Board review (Indiana University protocols 12410 and 1102004860). Our results can be reproduced using code available at github.com/osome-iu/io-coordinated-replies and data available at doi.org/10.5281/zenodo.13896309 . The collection and release of the dataset comply with the Twitter platform's terms of service. To mitigate the potential ethical risks of analyzing human subjects, we only rely on the data of public Twitter accounts do not include any raw data. We only manually inspect the profiles of the targets of the attacks, who are public figures and constitute the vulnerable group our study aims to protect. We provide our annotation data about these profiles for reproducibility. Our classification models do not use any personally identifiable information. We only report aggregated results. While our main objective is to detect coordinated replies, an attack that is frequently employed by information operations, we acknowledge that it may be also used by regular social media users organizing among themselves for activism. Thus, we suggest that our classifiers should complement human investigation when employed in the wild. Furthermore, they should not be misused to label users as information operation accounts without thorough human verification.

References

Addawood, A.; Badawy, A.; Lerman, K.; and Ferrara, E. 2019. Linguistic Cues to Deception: Identifying Political Trolls on Social Media. In *Proc. Intl. AAAI Conf. on Web and Social Media*, volume 13, 15–25.

Alizadeh, M.; Shapiro, J. N.; Buntain, C.; and Tucker, J. A. 2020. Content-based Features Predict Social Media Influence Operations. *Science Advances*, 6(30).

Bradshaw, S.; and Howard, P. 2017. Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. Technical report, University of Oxford.

Bush, D. 2020. Fighting Like a Lion for Serbia: An Analysis of Government-Linked Influence Operations in Serbia. https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/serbia_march_twitter.pdf. Accessed: 30 Oct 2023.

DiResta, R.; Kheradpir, T.; and Miller, C. 2020. "The World is Swimming in a Sea of Rumors": Influence Operations Associated with El Fagr Newspaper (Egypt). Technical report, Stanford Internet Observatory. https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/egypt_whitepaper.pdf. Accessed: 8 Nov 2023.

Dutt, R.; Deb, A.; and Ferrara, E. 2018. "Senator, We Sell Ads": Analysis of the 2016 Russian Facebook Ads Campaign. In *Intl. Conf. Intell. Info. Techn.*, 151–168.

Elmas, T. 2023. Analyzing Activity and Suspension Patterns of Twitter Bots Attacking Turkish Twitter Trends by a Longitudinal Dataset. In *Companion Proc. ACM Web Conf.*, 1404–1412.

Elmas, T.; Overdorf, R.; and Aberer, K. 2022. Characterizing Retweet Bots: The Case of Black Market Accounts. In *Proc. Intl. AAAI Conf. on Web and Social Media*, 171–182.

Elmas, T.; Overdorf, R.; and Aberer, K. 2023. Misleading Repurposing on Twitter. In *Proc. Intl. AAAI Conf. on Web and Social Media*, 209–220.

Ezzeddine, F.; Ayoub, O.; Giordano, S.; Nogara, G.; Sbeity, I.; Ferrara, E.; and Luceri, L. 2023. Exposing Influence Campaigns in the Age of LLMs: A Behavioral-based AI Approach to Detecting State-Sponsored Trolls. *EPJ Data Science*, 12(1): 46.

Farkas, J.; and Bastos, M. 2018. IRA Propaganda on Twitter: Stoking Antagonism and Tweeting Local News. In *Proc. 9th Intl. Conf. on Social Media and Society*, 281–285.

Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; and Wang, W. 2020. Language-agnostic BERT Sentence Embedding. *arXiv preprint arXiv:2007.01852*.

Ferrara, E.; Chang, H.; Chen, E.-P.; Muric, G.; and Patel, J. 2020. Characterizing Social Media Manipulation in the 2020 U.S. Presidential Election. *First Monday*, 25(11).

Gallagher, R. J.; Frank, M. R.; Mitchell, L.; Schwartz, A. J.; Reagan, A. J.; Danforth, C. M.; and Dodds, P. S. 2021. Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts. *EPJ Data Science*, 10(1): 4.

Hudson Smith, D.; Ehrett, C.; and Warren, P. L. 2024. Unsupervised Detection of Coordinated Information Operations in the Wild. Technical Report 2401.06205, arXiv.

Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; and Gilbert, E. 2020. Still Out There: Modeling and Identifying Russian Troll Accounts on Twitter. In *12th ACM Conf. on Web Science*.

Kong, Q.; Calderon, P.; Ram, R.; Boichak, O.; and Rizoiu, M.-A. 2023. Interval-censored Transformer Hawkes: Detecting Information Operations using the Reaction of Social Systems. In *Proc. ACM Web Conf.*, 1813—-1821.

Luceri, L.; Giordano, S.; and Ferrara, E. 2020. Detecting Troll Behavior via Inverse Reinforcement Learning: A Case Study of Russian Trolls in the 2016 US Election. In *Intl. AAAI Conf. on Web and Social Media*, volume 14, 417–427.

Matthews, J.; and Goerzen, M. 2019. Black Hat Trolling, White Hat Trolling, and Hacking the Attention Landscape. In *Companion Proc. WWW Conf.*, 523–528.

Merhi, M.; Rajtmajer, S.; and Lee, D. 2023. Information Operations in Turkey: Manufacturing Resilience with Free Twitter Accounts. In *Proc. Intl. AAAI Conf. on Web and Social Media*.

Neudert, L.-M.; Howard, P.; and Kollanyi, B. 2019. Sourcing and Automation of Political News and Information During Three European Elections. *Social Media + Society*, 5(3).

Ng, L. H. X.; Moffitt, J.; and Carley, K. M. 2022. Coordinated through a Web of Images: Analysis of Image-based Influence Operations from China, Iran, Russia, and Venezuela. *Preprint arXiv:2206.03576*.

Nwala, A. C.; Flammini, A.; and Menczer, F. 2023. A Language Framework for Modeling Social Media Account Behavior. *EPJ Data Science*, 12(1): 33.

Office of the Director of National Intelligence. 2017. Assessing Russian Activities and Intentions in Recent US Elections. Technical report, National Intelligence Council.

Ong, J. C.; and Cabañes, J. V. A. 2018. Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines. Tech. rep., UMass Amherst. https://doi.org/10.7275/2cq4-5396.

Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. In *Proc. Intl. AAAI Conf. on Web and Social Media*, volume 21, 455–466.

Pamment, J.; and Smith, V. 2022. Attributing Information Influence Operations: Identifying those Responsible for Malicious Behavior Online. Technical report, NATO Strategic Communications Center of Excellence.

Rowett, G. 2018. The Strategic Need to Understand Online Memes and Modern Information Warfare Theory. In *Proc. IEEE Big Data*.

Senate Select Committee on Intelligence. 2019. 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 with Additional Views. https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf. Accessed: 2024-01-16.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.; Flammini, A.; and Menczer, F. 2018. The Spread of Low-credibility Content by Social bots. *Nature Communications*, 9: 4787.

Sharma, K.; Zhang, Y.; Ferrara, E.; and Liu, Y. 2021. Identifying Coordinated Accounts on Social Media through Hidden Influence and Group Behaviours. In *27th ACM SIGKDD Conf. on Knowledge Discovery & Data Mining*, 1441–1451.

Stanford Internet Observatory. 2021. Published reports of the Stanford Internet Observatory. https://github.com/stanfordio/publications.

Stewart, A.; Mosleh, M.; Diakonova, M.; Arechar, A.; Rand, D.; and Plotkin, J. 2019. Information gerrymandering and undemocratic decisions. *Nature*, 573(7772): 117–121.

Stewart, L. G.; Arif, A.; and Starbird, K. 2018. Examining Trolls and Polarization with a Retweet Network. In *Proc. ACM WSDM Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*. https://api.semanticscholar.org/CorpusID:44033303.

Thomas, E.; Thompson, N.; and Wanless, A. 2020. The Challenges of Countering Influence Operations. Technical report, Carnegie Endowment for Intl. Peace. https://carnegieendowment.org/2020/06/10/challenges-of-countering-influence-operations-pub-82031.

Torres-Lugo, C.; Pote, M.; Nwala, A.; and Menczer, F. 2022. Manipulating Twitter through Deletions. In *Proc. Intl. AAAI Conf. on Web and Social Media*, 1029–1039.

Uyheng, J.; Cruickshank, I. J.; and Carley, K. M. 2022. Mapping State-sponsored Information Operations with Multi-view Modularity Clustering. *EPJ Data Science*, 11(1): 25.

Weedon, J.; Nuland, W.; and Stamos, A. 2017. Information Operations and Facebook. Technical report, Facebook. https://about.fb.com/wp-content/uploads/2017/04/facebook-and-information-operations-v1.pdf.

Woolley, S. C.; and Howard, P. N. 2018. Conclusion: Political Parties, Politicians, and Computational Propaganda. In *Computational propaganda: Political parties, politicians, and political manipulation on social media*, 241–248. Oxford University Press.

Yang, K.-C.; and Menczer, F. 2024. Anatomy of an AI-powered Malicious Social Botnet. *Journal of Quantitative Description: Digital Media*, 4.

Zannettou, S.; Caulfield, T.; Bradlyn, B.; De Cristofaro, E.; Stringhini, G.; and Blackburn, J. 2020. Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter. *Proc. Intl. AAAI Conf. on Web and Social Media*.

Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019a. Disinformation Warfare: Understanding State-sponsored Trolls on Twitter and their Influence on the Web. In *Companion Proc. WWW Conf.*, 218–226.

Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; and Blackburn, J. 2019b. Who Let the Trolls Out? Towards Understanding State-sponsored Trolls. In *Proc. 10th ACM Conf. on Web Science*, 353–362.

Ethics Checklist

1. For most authors...

   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes

   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes

   (e) Did you describe the limitations of your work? Yes

   (f) Did you discuss any potential negative societal impacts of your work? NA

   (g) Did you discuss any potential misuse of your work? NA

   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes

   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

2. Additionally, if your study involves hypotheses testing...

   (a) Did you clearly state the assumptions underlying all theoretical results? NA

   (b) Have you provided justifications for all theoretical results? NA

   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

   (e) Did you address potential biases or limitations in your theoretical framework? NA

   (f) Have you related your theoretical results to the existing literature in social science? NA

   (g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? Answer

3. Additionally, if you are including theoretical proofs...

   (a) Did you state the full set of assumptions of all theoretical results? NA

   (b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes

   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes

   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes. Our experiments report averages across 10-fold cross-validation as well as standard errors.

   (d) Did you include the total amount of computation and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes

   (e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes

   (f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. We recommend manual inspection to complement our classifiers.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

   (a) If your work uses existing assets, did you cite the creators? Yes.

   (b) Did you mention the license of the assets? NA

   (c) Did you include any new assets in the supplemental material or as a URL? Yes. We provide link to data and code as URL.

   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes; this is addressed by IRB exemption.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes. The Twitter profiles are public information.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR? Yes. The link to data is included in the paper so that anyone can access it, including the details of the metadata in the Datasheet to make it interoperable and reusable.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset? The Datasheet is available at ANON.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA