Unraveling the Web of Disinformation: Exploring the Larger Context of State-Sponsored Influence Campaigns on Twitter

Mohammad Hammas Saeed[♣], Shiza Ali[♣], Pujan Pauel[♣], Jeremy Blackburn[♦], and Gianluca Stringhini[♣]
[♣]Boston University, [♦]Binghamton University
{hammas,shiza,ppaudel,gian}@bu.edu, jblackbu@binghamton.edu

***Abstract*—Social media platforms offer unprecedented opportunities for connectivity and exchange of ideas; however, they also serve as fertile grounds for the dissemination of disinformation. Over the years, there has been a rise in state-sponsored campaigns aiming to spread disinformation and sway public opinion on sensitive topics through designated accounts, known as *troll accounts*. Past works on detecting accounts belonging to state-backed operations focus on a single campaign. While campaign-specific detection techniques are easier to build, there is no work done on developing systems that are campaign-agnostic and offer generalized detection of troll accounts unaffected by the biases of the specific campaign they belong to.**

In this paper, we identify several strategies adopted across different state actors and present a system that leverages them to detect accounts from previously unseen campaigns. We study 19 state-sponsored disinformation campaigns that took place on Twitter, originating from various countries. The strategies include sending automated messages through popular scheduling services, retweeting and sharing selective content and using fake versions of verified applications for pushing content. By translating these traits into a feature set, we build a machine-learning-based classifier that can correctly identify up to 94% of accounts from unseen campaigns. Additionally, we run our system in the wild and find more accounts that could potentially belong to state-backed operations. We also present case studies to highlight the similarity between the accounts found by our system and those identified by Twitter.

I. Introduction

Social media platforms have become prominent sources for accessing information and communication for millions of people worldwide. As these platforms are used more and more for information propagation, there has been an associated risk of *misinformation* and *disinformation*. *Misinformation* is the spread of false information without malicious intent (e.g., a regular Twitter user innocuously promotes a COVID-19 false narrative tweet), whereas *disinformation* is the intentional spread of false information by malicious actors [43], [45], [47], [61]. Recent years have seen a rise in state-sponsored disinformation campaigns, where governments and their affiliated entities exploit social media platforms to shape narratives, manipulate public opinion, and advance their strategic agendas [47], [48]. These campaigns are often conducted by a designated set of accounts, known as *troll accounts*, which are created by malicious actors and often manually controlled to post content and interact with each other and real users [5], [7], [20], [42], [57]. The actors, often operating covertly, exploit various channels and tactics to disseminate misleading information. For example, in 2020, a six-year-long Russian disinformation campaign named "Secondary Infektion" was found to be spreading pro-Russian narratives and interfering with the 2016 US presidential election across 300 social media platforms in seven different languages [2].

Motivation. Over the years, disinformation campaigns have grown in scale and are now operating globally. According to recent studies, over 200 operations targeting various countries were taken down by Meta in 2022 alone [32]. In fact, many campaigns are now being outsourced to third-party agencies (e.g., troll-farms [31] and PR firms [35]) by political actors [24]. However, unlike social media bots and spam, these campaigns are often more sophisticated and deliberate in nature [38], making the threat model more intricate. The bare-bones attack model is often two-fold: 1) the campaign is geared towards achieving a certain goal (e.g., spreading an ideology, causing conflict and strife) on a target platform, while 2) slipping under the radar of detection systems, blending into the community, and appearing "legitimate." Since these campaigns exhibit group activity and human-like behavior [17], [38], it is important to develop systems that are specifically tailored towards them and more sophisticated than systems designed to detect automated activity.

With the passage of time, various evasion tactics are being used by troll accounts to protect from detection systems and human moderation [27], [41], [75]. Despite this, little work has been done to understand the shared characteristics of these campaigns to develop systems that can offer broad-scale generalized detection. Some recent research has leaned towards understanding the common traits of these campaigns; e.g., miscreants purposely pump a large quantity of "spammy" comments on a target platform to divert attention from the original narratives being pushed, or designated fake accounts within a campaign are created for specific tasks [32]. Other works have looked into specific state-sponsored campaigns or events on Twitter (e.g., the 2018 Brazilian presidential campaign [46], Italian disinformation during the 2019 European elections [36], and Internet Research Agency (IRA)

This paper is accepted for publication in the Proceedings of the 2024 International Symposium on Research in Attacks, Intrusions and Defenses (RAID). Please cite accordingly.

accounts [70]).

Technical Roadmap. Unlike past research that is campaign-specific, our research focuses on uncovering the common themes among disinformation campaigns to perform campaign-agnostic detection. In this paper, we identify several universal traits that are shared among different campaigns. We analyze Twitter data for state-sponsored campaigns that spans 19 countries and includes more than 200 million tweets from the years 2018 to 2022. We find that these campaigns use a variety of techniques to perform their operations, e.g., using scheduling services to delegate their posting tasks, utilizing fake third-party versions of popular applications (e.g., "Twitter for Android") to post messages, extensively retweeting to push certain agendas, and posting innocuous messages (e.g., motivational quotes) to potentially avoid detection. We also highlight potential coordination patterns where accounts from different campaigns exhibit similar characteristics (e.g., using the same Twitter sources to post their messages) around the same timeframe, making our findings in line with some recent works pointing towards potential inter- and intra-state coordination in campaigns [64].

Overall, we identify several universal traits and create a cross-campaign detection system that can detect upto 94% accounts from unseen campaigns. We demonstrate the efficacy of our system by training the classifier on each campaign one by one and detecting accounts from all the other campaigns. Lastly, we find potential malicious accounts in the wild and highlight some case studies that showcase their involvement in disinformation campaigns.

Contributions. This paper makes the following contributions:

- We uncover salient features used by state-sponsored disinformation operations (e.g., using fake third-party applications). These findings shed light on ways to identify this activity and detect potential malicious accounts.
- We develop a machine learning classifier to detect malicious accounts from previously unseen campaigns. Our most successful implementation uses a Random Forest model with an F1-Score of 97.8%. We also perform cross-campaign detection, flagging upto 94% accounts from unseen campaigns.
- We evaluate our system on 2,696 Twitter accounts using fake third-party applications. We identify 116 new malicious accounts potentially belonging to state-backed operations and present case studies to provide evidence that these accounts operate similarly to accounts from known campaigns.

Paper Organization. The rest of the paper is organized as follows. The next section describes our dataset. In Section 3, we analyze the campaigns in our dataset and highlight the common strategies they use, which informs the development of our detection model. Next, we translate our findings into features to build a machine learning classifier that can distinguish between real accounts and state-backed accounts in Section 4, while in Section 5, we show that our system can perform cross-campaign detection. In Section 6, we discuss several case studies of accounts that our system detects in the wild, while in Section 7, we discuss other works related to our research; finally, we discuss the implications of our results, future work, and conclude the paper in Section 8.

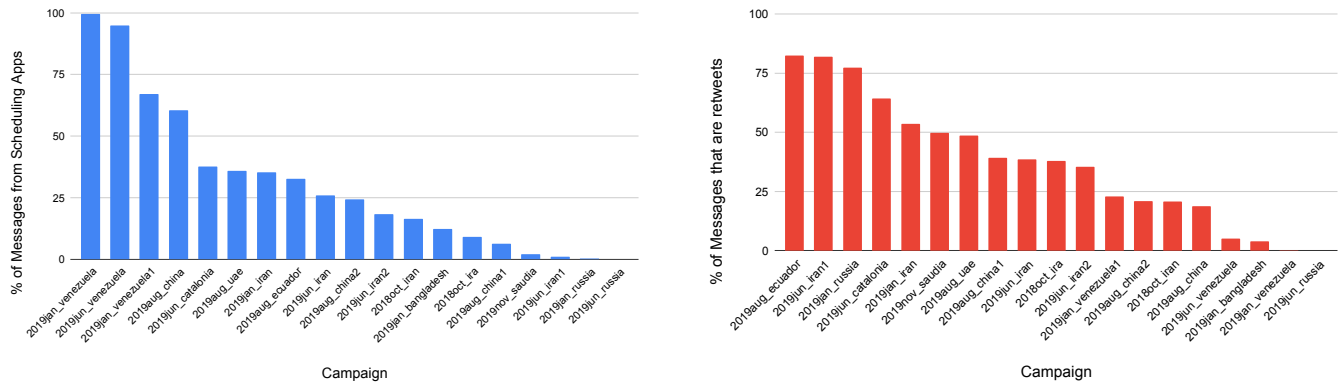
Campaign	Number of Accounts
2019nov_saudia	5,929
2019aug_china2	4,301
2019aug_uae	4,248
2018oct_ira	3,613
2019jun_iran2	2,865
2019jan_iran	2,320
2019jan_venezuela1	1,196
2019jun_iran	1,666
2019aug_ecuador	1,019
2018oct_iran	770
2019jan_venezuela	764
2019aug_china	744
2019jan_russia	416
2019jun_iran1	248
2019aug_china1	197
2019jun_catalonia	130
2019jun_venezuela	33
2019jan_bangladesh	15
2019jun_russia	4

TABLE I: Number of Accounts Per Campaign.

II. Data

We use Twitter for our analysis, as the platform released a publicly available dataset from campaigns active in diverse countries. Twitter is a popular social media platform that enables users to share and interact with short messages, known as tweets, in real-time. With its widespread adoption and influence, Twitter has attracted attention from researchers across various domains, with past works exploring different aspects of Twitter, including user behavior [11], information diffusion [4], and fake account detection [22], [51]. Later on, we use Twitter's API for collecting tweets; note that the Twitter free API was discontinued during the course of this project, and we discuss its implications in Section VIII. Broadly, we divide our dataset into two categories:

Campaign Data. The Twitter Transparency Dataset is a collection of publicly available information provided by Twitter related to various aspects of the platform's operations. It aims to promote transparency and enable researchers, journalists, and the public to analyze and understand Twitter's activities, policies, and content moderation efforts. We use the data from the Twitter Moderation Research Consortium [56] which contains information on platform manipulation campaigns attributed to various state-backed actors from 2018 to 2022. We study 19 campaigns, spanning more than 200 million Tweets and nine terabytes of media. It contains data from almost 80,000 accounts in total. Table I shows the campaigns we use in our dataset and the number of accounts in each campaign. Throughout the paper, the naming convention we use for campaigns is: {year}{month}_{country}{variant}.



(a) Scheduled Messages

(b) Retweeted Messages

Fig. 1: The graphs show the percentage of messages through (a) scheduling applications and (b) the percentage of messages that are retweets.

The month and year refer to the time when said campaign was *detected* and is officially designated by Twitter in the dataset. The variant portion of the name helps distinguish between campaigns because there are multiple active campaigns from the same country detected at similar time periods.

Twitter 1% Data. For this research, we use pre-crawled data from the Twitter 1% Data API for the years 2017-2022. The API used a 1% random sample of public Tweets for a given day. We use different parts of this data in later sections for finding regular accounts to compare with state-backed accounts.

Ethics. Our work is not categorized as human subjects research by our IRB since we do not interact with human subjects and solely use data that is already available to the public. Nonetheless, we adhere to ethical standards by removing any personally identifiable information when presenting example tweets in the paper and redacting usernames from tweets to prevent deanonymizing users.

III. Characteristics of Campaigns

In this section, we dive deeper into the different techniques used by state-sponsored campaigns to varying degrees. This section is divided into two parts: 1) Campaign-Level characteristics and 2) Account-Level characteristics. We first identify campaign-level characteristics by analyzing the tweets across various metrics like timing, coordination, content, and posting habits. We also uncover patterns that are shared across campaigns. We then map these campaign-level characteristics to account-level features, which can later aid us in developing a system to distinguish between real-world users and accounts from state-sponsored campaigns.

A. Campaign-Level Characteristics

Scheduled Messages. On Twitter, each tweet has a "source" field that refers to the application or platform from which the tweet was posted. We analyze the source field of all tweets posted by disinformation campaigns and pick the top 50 most commonly used applications. We identify seven popular scheduling applications (i.e., IFTTT, TweetDeck, dlvr.it, Hootsuite, Twibble, SocialOomph, and Zapier.com) among the top applications. These applications are used to send a total of 15,600,226 messages across 19 campaigns. In Figure 1a, we show that most coordinated campaigns use scheduling applications for a large percentage of their messages, with some campaigns posting almost all messages from scheduling applications, like the June 2019 and January 2019 Venezuelan campaigns. On average, 30% of the messages in a campaign belong to one of the scheduling applications.

Retweeting Similar Messages. A retweet allows users to share someone else's tweet with their own followers. When a user retweets a tweet, it appears on their own timeline and is visible to their followers. This enables users to amplify and redistribute content that they find interesting, informative, or worthy of sharing for other reasons. We find that campaigns have upto 78% of their entire tweets as retweets, as shown in Figure 1b. We also find that certain messages are retweeted both intra- and inter-campaign, which can be suggestive of a collaborative effort or a shared modus operandi (e.g., retweeting popular tweets to appear "legitimate"). For example, the following tweets are retweeted by multiple campaigns (i.e., 2019aug_china1, 2019nov_saudia, 2019aug_ecuador and 2019aug_china2) which are unrelated suggesting group-like behavior:

> RT [REDACTED]: follow everyone that retweets this. ✂
> RT [REDACTED]: Now that's a t-shirt cannon if I've ever seen one
> RT [REDACTED]: follow everyone who LIKES and RTs this 🐟
> RT [REDACTED]: Need to mass unfollow? Go to http://www.iunfollow.com There are no limits and it's free! No signup required!
> RT [REDACTED]: https://t.co/dTdDjRvXbs

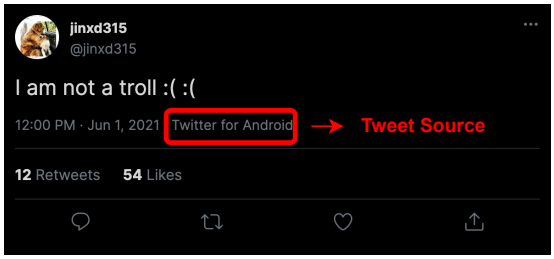


Fig. 2: Tweet from "Twitter for Android" source.

> RT [REDACTED]: Siga todos que FAV e RT esse tweet ●

These retweets are encouraging users to follow or like a particular account or tweet. Similar patterns are observed in follower markets; services that use accounts to build fake influence and reputation. Previous research extensively studied follower markets [50], [53], [65] and the detection of accounts involved in such activity [69]. However, it is important to note that follower-market tweets are not the only kind of retweets these accounts make; we exclude other examples for brevity.

Impersonating Third-Party Sources. We find that coordinated campaigns rely on specific third-party sources for posting content. Many of these sources are fake versions of existing applications, since Twitter forces each application name to be unique. For example, the fake version of "Twitter␣for␣Android" will be "Twitter␣for␣␣Android" with an extra space (␣) between "for" and "Android." In Table II, we provide several examples of original sources alongside their corresponding fake sources, along with their frequency of occurrence. Upon manual inspection and from the examples in Table II, we conclude that fake sources often contain spelling errors, leading spaces, and other similar changes from the original app source name that are difficult to distinguish visually. We search the latest Twitter 1% Data from 2022, which contains a random 1% sample of tweets made on each day of 2022, to find out if regular users also post with these sources. Of all the applications listed in Table II, we only find two messages from "␣Twitter for iOS," indicating these sources are not used by regular users and are more likely to be used by state-backed accounts. Table II shows that an impersonated version of "Twitter for Android" was used to send 106,636 tweets. State-sponsored accounts might use fake third-party sources for a variety of reasons. For example, these sources might aid in the management of accounts by offering automation, added functionalities, and extra features that prove beneficial for the campaigns. Another reason to use such sources is their legitimate-looking names, which make the account look legitimate at a cursory glance by users on the platform. Figure 2 shows how the source used to appear on Twitter and an extra space or basic string manipulation can go easily unnoticed. We also did not observe regular users using these applications in our dataset. Additionally, we search Google Play Store and Apple App Store and do not find these impersonated third party applications, indicating that these are not publicly available apps and are used exclusively by malicious actors.

Original Application	Third-Party Version(s)	Frequency
"Twitter for Android"	"Twitter for␣␣Android",	106,636
	"Twidere for Android #2",	34
	"Twidere for Android #5",	54
	"Twidere for Android #7",	73
	"Twitter from Android"	3,283
	"android apps for twitter"	1,555
"Twitter for iPad"	"Twitter for␣␣iPad",	20
	"twtkr for iPad",	275
	"␣␣␣␣twtkr for iPad"	73
"Twitter for iPhone"	"Twitter for iphons",	12,523
	"twitter for Iphone ios"	782
"Instagram"	"␣␣␣Instagram"	3
"Twitter for iOS"	" Twitter for iOS"	364
"HTC Peep"	"␣␣HTC Peep"	427
"Hootsuite",	"hootsuite",	6,127
"Hootsuite Inc."	"␣␣␣␣␣␣␣HootSuite"	43

TABLE II: Third-party versions of original applications used by state-sponsored campaigns.

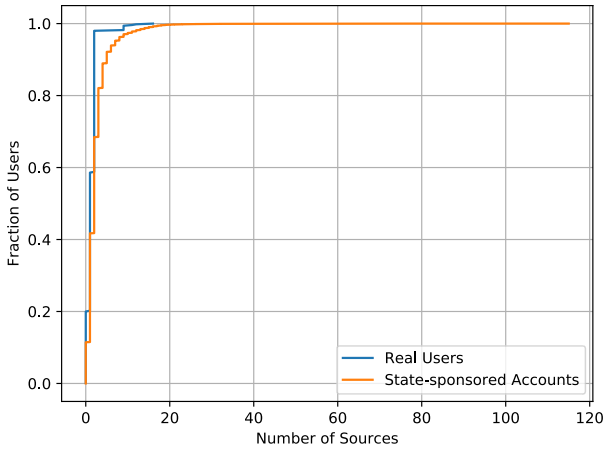
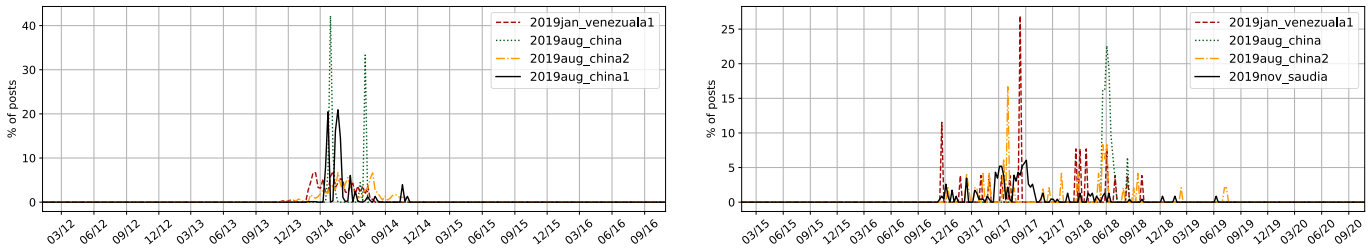


Fig. 3: CDF of number of sources used by real users and accounts from all campaigns.

Number of Sources. We randomly select 500 accounts from the Twitter 1% data discussed in Section III and collect their tweets from 2016 to 2020 using the Twitter Search API.[1] Figure 3 shows that, on average, a state-backed account uses more sources to post its messages than a regular account. The maximum sources used by a regular user are 16 as compared to 115 for a state-backed account. The discrepancy in the number of sources can be attributed to several factors. For a regular user, it is likely that they might switch between different devices or platforms depending on their circumstances or contexts. For example, they might use a mobile application while on the go and switch to a web client when using a com-

[1]The Twitter API is no longer publicly available and does not show the tweet source.



(a) "Twitter for␣␣Android" (b) "Moments Internal Auth"

Fig. 4: The time graphs show the same applications being used around the same time period by different campaigns.

puter. However, in the case of a state actor, utilizing diverse sources can help manage their campaigns. For example, using scheduled applications for sending messages that make the accounts appear real (e.g., "advice messages"), and using other applications to spread the actual harmful content. It might also be advantageous to spread out application use when it comes to third-party applications, since abnormally high activity on fake sources might lead the sources to get flagged and suspended. For example, the campaign *2019jan_venezuela1* used the fake application "Twitter for␣␣Android" primarily for retweets, where 4,061 out of the total 4,783 messages posted from the application were either retweets or replies to other tweets.

Simulating Legitimacy. As seen before, we find a trend of retweeting messages both inter- and intra-campaign. On a similar note, we find that a number of messages "unrelated" to any specific agenda(s) are posted. For example, below are "advice" messages that are verbatim shared by various unrelated campaigns (i.e., 2019jan_iran, 2019aug_china1, 2019jan_russia, 2018oct_ira, 2019nov_saudia and 2019aug_china2) suggesting access to a common knowledge base:

> When a friend does something wrong, don't forget all the things they did right.
>
> ———
>
> When two people are meant for each other, no time is too long, no distance is too far, and no one can ever tear them apart.
>
> ———
>
> My past is my past, it made me who I am, I have no regrets, wouldn't change a thing. I just don't live there anymore.
>
> ———
>
> Just because I don't talk to you, or text you first, doesn't mean I don't miss you. I'm just waiting for you to miss me.
>
> ———
>
> I get jealous because I'm afraid someone is going to make you happier than I do.

These quotes, presented as inspirational or uplifting content, might serve as camouflage for the dissemination of misinformation while building trust and credibility among their audience. A similar behavior of posting cute dog pictures to appear legitimate and blend into the community has been shown in past research [38]. The presence of positive and motivational content may divert attention from the deceptive nature of the underlying narratives and can also aid in avoiding detection [48]. By appealing to emotions, these accounts might seek to amplify the impact of their disinformation, as individuals are more likely to share content that elicits emotional responses [62].

Timing and Coordination. We analyze various third-party applications and the times when they were used by various campaigns. Figure 4 highlights two applications being used by different campaigns. The fake application "Twitter for␣␣Android" was actively used between 2013 and 2015 by four different campaigns. Similarly, "Moments Internal Auth" shows various coordinated spikes between 2016 and 2019 for four different campaigns. "Moments" is an application that allows users to curate and present collections of tweets around specific topics or events. It provides a way to aggregate tweets, including text, images, and videos, into a single, immersive narrative. The plot in Figure 4 shows an example of possible strategic coordination among these actors, potentially driven by shared objectives or alliances. It is important to note that the accounts from each campaign were actively posting from other applications before and after these posts, further implying a strategic reason for using these applications during the specific timeframes (i.e., 2013-2014 in Figure 4a and 2016-2019 in Figure 4b). The synchronized usage hints at the existence of information-sharing networks or coordinated efforts to exchange tactics and strategies among state-sponsored actors. It is also plausible that these custom applications were being marketed and sold to state actors, offering functionalities that align perfectly with their specific campaigns.

Pushing Agendas and Stylometry. We observe instances where different campaigns simultaneously push identical narratives. The similarity can be attributed to geopolitical strategies, economic objectives, or ideological alignments. When two or more campaigns have aligned goals, they may coordinate their disinformation efforts to amplify their desired message and enhance its impact. It is also likely that in the realm of disinformation, successful strategies and tactics are emulated. Below are some tweets in Arabic that are shared amongst three different Iranian campaigns (2019jun_iran, 2019jun_iran1 and

2019jun_iran2), along with their translation (**NB:** Farsi, not Arabic, is the language spoken in Iran):

Message 1: من لم يكفّر الشيعة فهو كافر!
Translation: Whoever does not declare the Shiites to be infidels is an infidel!
Message 2: ای شیعیان! شما به ما منسوب هستید، پس مایه ی زینت ما باشید نه مایه ی آبروریزی ما
Translation: O Shiites! You belong to us, so be our adornment and not my disgrace
Message 3: جز راه _حسین ،باقی راه ها،براهه است
Translation: Except the way of Hossein, the rest of the ways are misguided
Message 4: نکاح حیواناتزابطه باحیوانات
Translation: Marriage of animals: relationship with animals...
Message 5: نکاح الحیواناتمعاشرةالحیوانات لتجنب الزنا
Translation: #Nekaah_animals: Living with animals to avoid fornication!

Sunni and Shia are two different sects of Islam that hold different beliefs and have ongoing tension; notably with respect to Iran/Saudi Arabia relations. The first tweet calls out Shia Muslims as infidels, whereas tweets two and three are in favor of Shias. In the context of the third tweet, Hossein is the grandson of the Muslim Prophet and is particularly respected in the Shia community. The tweet calls all castes other than Shias misguided. This is textbook controversy, which is the cornerstone of troll account behavior, i.e., accounts take both sides of the argument to cause strife [38]. On the other hand, the last two tweets are examples of exaggerated claims and blatant disinformation while appealing to religious sentiment. The author is promoting marriage (or "Nekaah" in Arabic) with animals to prevent the sin of fornication.

Takeaways. In a nutshell, we identify certain characteristics of state-sponsored campaigns, shedding light on their potentially coordinated nature and deceptive strategies. We find seven key characteristics of state-backed operations: 1) the extensive use of scheduling applications, 2) impersonating third-party applications, 3) leveraging a variety of sources, 4) retweeting similar messages, 5) simulating legitimacy, 6) coordinating their timings, and 7) pushing certain agendas.

Our findings suggest that there are shared behavioral patterns exhibited by state-sponsored accounts from different campaigns despite originating in different countries. Recent work showed possible coordination amongst state actors in individual state-sponsored campaigns (e.g., retweeting behavior [64]). We find a similar trend across different campaigns on a varying scale. We analyze state-backed operations from various vantage points (e.g., information sharing behaviors and stylometry), in order to build a generalized feature fingerprint on how troll accounts disseminate information, communicate, and share content. We find that state-backed operations exclusively use impersonated third-party applications that are not available on official app stores. Multiple state-backed actors use these applications around the same timeframes. The state-backed accounts also make use of scheduling applications and on average, use more sources than regular users to post content. Although scheduling messages is traditionally associated with marketing campaigns, we observe that state-sponsored influence campaigns use this strategy too. We also find stylometric similarities in the posted content: retweeting of similar messages and amplifying similar content at large scale. The use of scheduling applications could serve as an easier method to scale such activities. On a similar note, recent work points towards a posting behavior called *copypasta* [59] networks where the same "disinformation" is duplicated in troll operations. The stylometric similarities we observe in our dataset and the retweeting and duplication of content across state actors points towards similar behavior at scale.

Putting it all together, in the following sections, we translate our empirical findings into account-level features to demonstrate that state-backed accounts show clear differences from benign users for these metrics.

B. Account-Level Characteristics

We now map the identified campaign characteristics into account-level signals. We prepare a list of 12 signals that capture various characteristics we have identified. The signals incorporate posting styles in the form of stylometric features (i.e., unique words, sentence length, and others), group activity in terms of retweets and mentions, and the use of "regular" well-known sources (i.e.,"Twitter Web Client," "Twitter for Android," "Twitter for iPhone," "Twitter for iPad," and "Twitter Web App") by the users. To determine the possibility of leveraging these features to build a generalized troll account detection system, we first perform statistical tests on the populations to observe key differences. In Table III, we show the mean value of each of the given features derived from the campaign-level characteristics in a set of 1000 real world Twitter users obtained from Twitter 1% data described in Section II and all troll accounts belonging to state-backed operations. We perform the Kolmogorov–Smirnov [67] test to determine whether the differences in scores are statistically significant. For each feature, we perform the test and report the scores in Table III, along with P-values. We set the value of α to 0.01, therefore, we only reject the null hypothesis if the P-value is <0.01. Overall, we can reject the null hypothesis for all metrics, which means that for all samples, the difference in values are statistically significant. Next, we aim to use these features along with others to build a machine learning classifier for detecting accounts from state-backed operations.

IV. Building the Detection Model

Based on our observations from Section III, we identify a set of features to train a supervised model to distinguish state-sponsored campaign accounts from legitimate Twitter accounts. Figure 5 lays out the entire pipeline of our system from feature extraction to finding potential state-backed accounts in the wild and alerting relevant authorities. Unlike previous works [30], [38], our model is designed to detect accounts

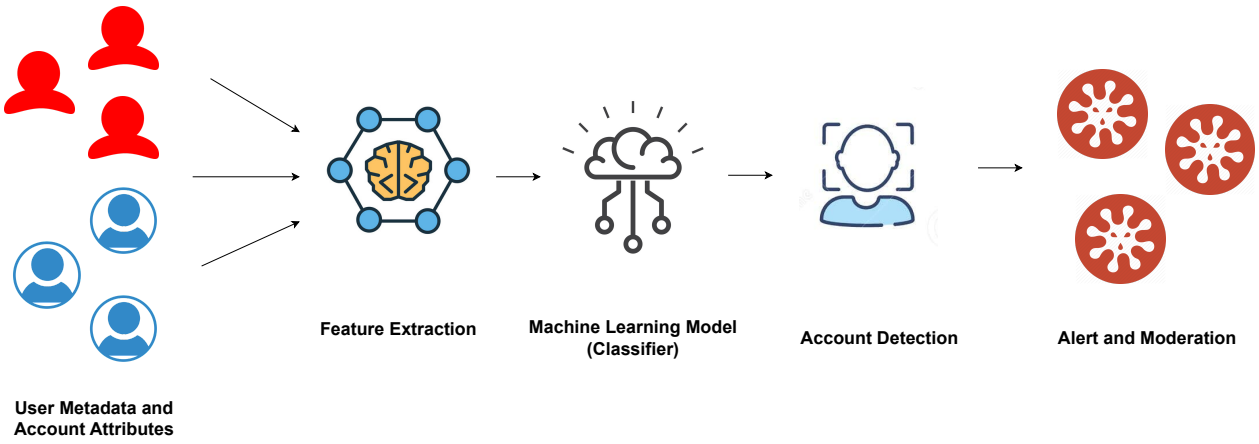


Fig. 5: Overview of the system: two input streams are fed to the system: (1) a dataset of known state-sponsored malicious accounts and (2) a set of benign accounts. The system extracts the features of both sets of accounts. Next, a detection model is built using the identified features and used to detect unseen accounts in the wild. Finally, the system alerts users to potential malicious accounts in the wild, and those accounts can be moderated accordingly.

Characteristic	Trolls	Real	KS	P-Val
Fraction of Messages by Regular Sources	0.67	0.72	0.10	**<0.01**
Cumulative Mentions per Tweet	9.06	0.63	0.53	**<0.01**
Fraction of Messages that are Retweets	0.16	0.31	0.17	**<0.01**
Number of Sources Used on Average	2.50	2.26	0.11	**<0.01**
Average Tweet Word Count	15.18	12.66	0.14	**<0.01**
Average Tweet Unique Words	13.45	11.21	0.13	**<0.01**
Average Tweet Stopword Count	1.42	1.90	0.10	**<0.01**
Average Tweet Punctuation Count	2.89	2.50	0.11	**<0.01**
Average Word Length	5.44	4.64	0.15	**<0.01**
Average Sentence Length	12.33	9.66	0.17	**<0.01**
Average Sentence Complexity	52.52	40.14	0.21	**<0.01**
Function to Non-function Words Ratio	0.065	0.084	0.11	**<0.01**

TABLE III: Statistical comparison between real users and state-sponsored accounts.

from previously unseen campaigns, i.e., it can be trained on one campaign and used to detect accounts from other unknown campaigns.

A. Feature Extraction

To train a machine learning system, it is first important to identify and select relevant attributes or characteristics from the raw input data. This process helps to transform the data into a suitable representation that the machine learning model can understand and learn from effectively. In Table III, we demonstrate the difference between state-backed actors and benign users for several features. Since there are clear distinctions, we are positive that we can use certain features to discern between malicious and non-malicious users and build a machine-learning based classifier. Therefore, in this section, we use those features and motivate several more that can be leveraged to build a detection system. The complete list of features is shown in Table IV. Later on in Table VI, we show how removing certain types of features from the model (e.g., stylometric features) impacts the overall classifier performance and the importance for using all the features for the most optimal performance.

No.	Feature	Novelty
1	tweet_count	[15], [21], [34], [57]
2	account_age	[12], [14]
3	no. of followers	[15], [21], [29], [34], [44], [57]
4	no. of following	[21], [34]
5	language	[14]
6	description_length	[29], [44], [57]
7	description_language	ours
8	cumulative_mentions_per_tweet	[57]
9	average_tweet_length	[15]
10	retweet_fraction	[57]
11-34	percentage of tweets in a given hour	[1]
35	average_tweet_word_count	[57]
36	average_tweet_unique_words	ours
37	average_tweet_stopword_count	ours
38	average_tweet_punctuation_count	ours
39	average_word_length	[21]
40	average_sentence_length	[21]
41	average_sentence_complexity	ours
42	function_to_non-function_words_ratio	ours
43	no_of_sources	[15]
44	fraction_of_messages_by_fake_sources	ours
45	fraction_of_messages_by_regular_sources	ours

TABLE IV: System features.

Overall, we divide the features into four main categories:

User Attributes. User metadata attribute features are valuable pieces of information that provide insights into the characteristics and behavior of users.

- *Tweet Count:* The number of tweets posted by an account

provides insight into its activity level. State-sponsored accounts have an average tweet count of 2,483 as opposed to 5,433 of real users, with a KS-Score of 0.42 and P-Value<0.01, making the difference statistically significant.

- *Account Age:* The age of an account can be indicative of suspicious behavior. According to past research, there is evidence of state-sponsored accounts being created around same the time frame or in batches [38].
- *Followers and Following:* State-sponsored accounts might follow each other to increase social proof and have a different follower-following ratio than a regular user. A state-sponsored account has 2,090 followers and 961 following while a regular user has 1,162 followers and 552, with a KS-Score of 0.17 and 0.16 respectively, with each P-Value<0.01, making the differences statistically significant.
- *Language and Description Language:* On average, 37% of state-sponsored accounts are non-English speaking, therefore we use the language of the account and its description as a feature. To model the language as a feature, we assign an integer value to each language. For example, "English" represented as "en" in the raw data would correspond to "1." In total, we consider 31 languages for our analysis, which are derived from both real-world users and state-backed accounts.
- *Description Length:* The length of an account's description can provide insight into its authenticity. State-sponsored accounts have an average description length of 38.7 as opposed to 43.8 of genuine users, with a KS-Score of 0.078 and P-Value<0.01, making the difference statistically significant.
- *Cumulative Mentions per tweet:* Twitter allows users to mention or tag other users in a tweet. To mention someone in a tweet, the user includes their username preceded by the "@" symbol (e.g., "@username [insert tweet text]") within the tweet's text. By calculating the cumulative average number of mentions per tweet, we identify patterns of interactions. Higher-than-average mentions as shown in Table III indicates a coordinated effort to promote a specific agenda or target specific individuals or groups. For this feature, we count the instances of other users being mentioned by the target account.
- *Average Tweet Length:* State-sponsored accounts may exhibit distinct patterns in tweet length. They have an average tweet length of 86.4 characters as compared to 68.7 characters of real world users, with a KS-Score of 0.194 and P-Value<0.01, making the difference statistically significant.
- *Retweet fraction:* This feature indicates the fraction of tweets that are retweets. By analyzing the number of retweets made by the users, we identify the sharing rate.

Temporal Characteristics. These features are represented as a size 24 vector, where each entry depicts the percentage of messages in that hour from the 24-hour window. Past research has shown evidence that troll accounts post during specific times of day, sometimes even as an office job [3].

Stylometry. Stylometry features allow us to capture linguistic nuances, such as vocabulary choices, sentence structure, punctuation, and grammatical patterns, that are characteristic of these accounts. These features are derived from the differences we observe in Table III.

- *Average Tweet Word Count:* By analyzing the average word count, the classifier can identify accounts that consistently produce unusually short or long tweets compared to real users. Deviations from the expected average word count can be indicative of templated content commonly found in such accounts.
- *Average Tweet Unique Words:* By calculating the average number of unique words in tweets, the classifier can identify the diversity in the vocabulary used by the accounts.
- *Average Tweet Stopwords Count:* Stopwords are common words like "and," "the," or "is" that carry little semantic meaning. Incorporating the stopwords count adds linguistic nuance to the classifier.
- *Average Tweet Punctuation Count:* Punctuation usage can reveal certain writing styles or patterns associated with state-sponsored accounts. Excessive use of punctuation marks, such as exclamation marks or ellipses, may indicate attempts to convey emotions or manipulate reader perception.
- *Average Word Length:* State-sponsored accounts may utilize specific writing techniques, such as elongating words, to mask their writing style or bypass content filters.
- *Average Sentence Length:* Sentence length can provide insights into the writing style and coherence of state-sponsored accounts. By calculating the average sentence length, the classifier can identify accounts that show signs of unnatural content generation.
- *Average Sentence Complexity:* By measuring average sentence complexity, such as the presence of complex sentence structures or syntactic constructions, the classifier can identify accounts that consistently produce content with a linguistic complexity indicative of malicious activity. We use the Flesch Reading Ease [18] score to compute complexity. It takes into account both the sentence length and the average number of syllables per word, providing a comprehensive measure of text complexity. The score ranges from 0 to 100, with higher scores indicating easier-to-read text.
- *Function to Non-Function Words Ratio:* Function words (e.g., articles, prepositions, and pronouns) and non-function words (e.g., nouns, verbs, and adjectives) have different roles in language. State-sponsored accounts may exhibit patterns of overusing or underusing function words to manipulate or convey specific messages.

Source Features. We consider source features to be important because they provide valuable information about the origin and

credibility of the content shared by state-backed accounts. The following source features are included in the classifier:

- *Number of Sources:* State-sponsored accounts often use multiple sources to share content. By analyzing the number of sources used in tweets, the classifier can identify accounts that exhibit an unusually high number of sources compared to real users.
- *Fraction of Messages by Fake Sources:* State-sponsored accounts may purposefully share content from sources known for spreading misinformation or propaganda (e.g., fake third-party versions of original applications). By tracking the number of sources with errors, the classifier can identify accounts that consistently share content from unreliable or deceptive sources. We consider a source fake if it contains extra or leading spaces and all lower-case letters.
- *Fraction of Messages by Regular Sources:* State-sponsored accounts use various scheduling applications and other third-party sources to post their messages. Since the list of scheduling applications and new third-party applications is ever-expanding, we account for the fraction of messages made by accounts from "regular" well-known sources (i.e.,"Twitter Web Client," "Twitter for Android," "Twitter for iPhone," "Twitter for iPad," and "Twitter Web App").

B. Training the System

We create a balanced dataset for training, with 500 accounts randomly selected from the Twitter 1% dataset from Section II as the negative class. We use the Twitter Search API to retrieve their tweets. To select accounts for the positive class, we use two campaigns with a high number of accounts, i.e., 2018oct_ira with 3,608 accounts and 2018oct_iran with 770 accounts. We randomly select 500 accounts from both campaigns and use those as positive classes to train the system. By using a balanced dataset, we ensure that the model is exposed to enough examples from both classes, reducing the risk of bias. In real-world scenarios, the class of interest (the positive class) is often the minority class. By balancing the dataset, the model also becomes more sensitive to the minority class, allowing it to learn patterns and features specific to that class [26], [54]. We report the average result of both iterations in Table V. For each iteration, we perform 10-fold cross-validation to train the system. We experiment with four classifiers: K-Nearest Neighbors (KNN) [66], Decision Tree [39], Support Vector Machines (SVM) [55], and Random Forest [28]. We evaluate the performance of each classifier based on accuracy, precision, recall, and F1-score. We find that Random Forest classifier works the best, achieving an accuracy of 98.5% and an F1-Score of 97.8%. Therefore, we select Random Forest for performing evaluation of our system and identifying accounts from unseen campaigns in the next section.

Component Testing. We also test each component of our system independently and present the results in Table VI. We find that, when considering individually, the metadata classifier performs the best, however it important to note that the stylometry and temporal classifier individually perform well. By manually reviewing the misclassified accounts, we find that the source classifier is able to pick up accounts which are misclassified by the other three classifiers, adding overall nuance and signal towards prediction capabilities. However, we achieve the best F1-score and accuracy when all components are put together.

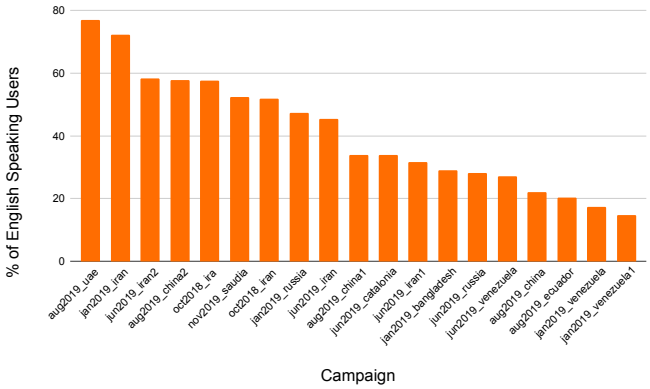


Fig. 6: The graph shows the percentage of English-speaking users per campaign.

Classifier	Accuracy	Recall	Precision	F1-Score
KNN	92.4%	92.6%	92.8%	92.7%
Decision Tree	95.4%	94.2%	94.6%	94.4%
Linear SVM	97.7%	97.6%	97.6%	97.6%
Random Forest	**98.5%**	**97.6%**	**97.9%**	**97.8%**

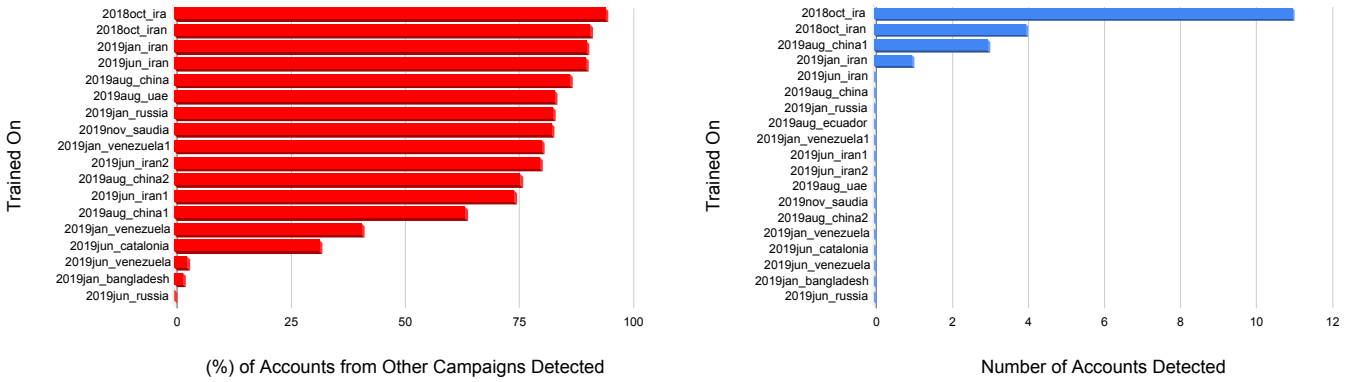
TABLE V: Classification performance of our system using 10-fold cross-validation.

Classifier	Accuracy	Recall	Precision	F1-Score
Metadata	95.4%	96.2%	96.3%	96.3%
Temporal	91.4%	92.9%	84.8%	88.9%
Stylometry	90.4%	90.0%	84.7%	87.4%
Source	76.7%	71.1%	57.2%	64.2%
All	**98.5%**	**97.6%**	**97.9%**	**97.8%**

TABLE VI: Classification performance of each classifier component individually.

Performance Using Unbalanced Dataset. It is important to understand how a real-world imbalance could affect the performance of our system. For this purpose, we use the 2019jun_venezuela with 33 accounts as positive class for the system and 500 real-world users as negative class. Due to limited data access, this is our best attempt to mimic a scenario with more examples of real-users. We find that in this setting the accuracy of our model slightly increases, with an accuracy of 99.5%, a precision score of 99.6%, a recall of 99.6% and an F1 score of 99.6%.

Structural Imbalances in Dataset. To identify differences in



(a) Training on Individual Campaigns and Evaluating on Others

(b) Training on Individual Campaigns and Evaluating on Real Users

Fig. 7: System performance and results.

state-sponsored actors and benign users, it is also important to account for structural imbalances in the dataset, for instance the language used by Twitter accounts. Figure 6 shows the percentage of English-speaking users per campaign, as determined by the "language" attribute in user metadata. We see diversity among campaigns ranging from a small percentage of English-speaking users to nearly all users being English-speaking. We find that the set of real users that we collect for our work consists of 26.7% English-speaking users. While language imbalance between state-sponsored accounts and benign accounts can introduce unintended bias in the model, we attempt to reduce the biases by evaluating our system by cross-validating on campaigns with different proportions of English-speaking users.

Feature Importance. We also examine the importance assigned to each feature by our machine learning model. We compute the Gini importance [40], which assigns a value between 0 and 1 for each feature. The more significant the feature for prediction, the higher its value. We find that "Retweet Fraction" and "Cumulative Mentions per Tweet" are the best predictors, confirming our findings from Section III that state-sponsored accounts have distinct attributes in sharing, promoting and pushing their content.

Our system leverages several modalities to detect troll campaigns. We propose a combination of user-level, temporal, stylometric, and source features for detection of state-backed accounts, where every component and its individual performance is highlighted in Table VI. A portion of these features have been used in the past for different account detection systems as shown in Table IV. However, the combination of features we propose is obtained by studying campaigns from various state-actors emerging from different countries and comparing their behavior with benign users. Our feature combination is unique and tailored towards influence operations. We also propose novel features (e.g., impersonated application use and stylometry). Past works have touched on some aspects of troll behavior, like loose coordination patterns [38], retweeting and promoting of each other's content [64] and duplicating or copy pasting text [59]. However, our system provides a holistic approach towards detecting accounts, where the performance is best by using a combination of factors that make a successful state-backed campaign.

V. Evaluation

In the previous section, we show that the model performs well on previously seen campaigns. Now, we use our system to detect accounts from unseen campaigns, i.e., we train the system individually on each campaign and detect the accounts from all the other campaigns. We use the negative class accounts from Section IV-B, and accounts from each campaign are considered positive class. Figure 7a highlights the performance of our system. When trained on the October 2018 IRA campaign, the system is able to detect roughly 94% of the accounts from other campaigns, i.e., 24,247 out of 25,667 total accounts. There are certain campaigns that train the system poorly, such as 2019jun_russia. However, it is important to note that these campaigns have very few accounts for the classifier to train on, like 2019jun_russia consists of only three accounts. With enough examples of the positive class, the classifier is able to detect more accounts from unseen campaigns.

Estimating False Positives. We also test our system on a separate set of real-world users randomly selected from the Twitter 1% dataset. From Figure 7b, the highest number of accounts flagged is eleven out of 629 when the system is trained on 2018oct_ira campaign. Therefore, we estimate the upper-bound false detection rate at roughly 1.75%, which is the worst-case scenario. For the majority of campaigns our system shows a 0% false detection rate.

A. Detection in the Wild

We search the Twitter 1% data from 2017 to find messages made from sources that are popularly used by the coordinated campaigns identified in Section III. We find 2,696 accounts that posted one or more messages from the application "Twitter for␣␣Android". We then run our system to identify accounts

Flagged Accounts	Other Accounts
admtvosenlucha	si
followers	quiero
اللهم	vida
الله	hoy
stats	día
igualdad	gracias
todas	mejor
partes	siempre
8m	así
unfollowers	amor
twitter	love
come	tan
join	bien
media	solo
policy	dios

TABLE VII: Top 15 words computed using TF-IDF for accounts that were flagged and not flagged by our system in the wild.

that might belong to state-backed operations. Our system marks 116 out of all the accounts.

Comparison of Flagged vs Other Accounts. Now, we compare the content features of accounts flagged by our system with the ones that are not flagged to look for linguistic signals. We use TF-IDF (Term Frequency - Inverse Document Frequency) for this task. A word's importance to a document in a collection or corpus is meant to be reflected by the TF-IDF statistic, which is a numerical measurement. Since some words are used more frequently than others overall, the TF-IDF value rises according to the number of times a word appears in the text and is offset by the number of documents in the corpus that contain the term. We use it to ensure the selection of words that are both important within a specific document and distinctive in the overall dataset. We compute the top 15 words using TF-IDF for tweets from both sets of accounts: 1) flagged and 2) non-flagged or other accounts. The results are shown in Table VII. Most words from the accounts not flagged by the system are generic in nature. The presence of religious terms like "اللهم" and "الله" (which mean "God" in Arabic) may suggest the exploitation of religious sentiments or potential attempts to influence religious communities through propaganda or divisive content. On the other hand, we observe terms like "si," "quiero," "vida," "hoy," and "día" in the non-flagged dataset, which are common Spanish words related to daily life. These terms and others like "gracias," "mejor," and "siempre" show typical social interactions rather than promoting specific agendas.

VI. Case Studies

In this section, we show various case studies from the accounts identified by our system in Section V-A which highlight their involvement in state-backed operations. We do so to present additional evidence that these accounts are potentially malevolent and operate in ways that recognized disinformation campaign accounts do.

Retweeting Similar Agendas. As discussed in Section III-A, a common strategy accounts in state-backed operations use is to retweet a large chunk of messages. A major reason to retweet certain messages can be to boost a given narrative and increase the visibility and exposure of selective content to a larger audience. Retweeting each other's messages also helps create an echo chamber effect by giving the illusion of widespread support or consensus. To the outside eye, repeated retweets from seemingly independent accounts give the impression that multiple sources support and validate the information being shared, thereby enhancing the perceived credibility of the messages. It is in the best interest of malicious actors to create the perception that their propaganda has gained traction and popularity. By retweeting each other's content, these accounts can also inflate engagement metrics, such as retweet counts and likes, while drowning out or overshadowing genuine voices and opinions. When targeting an important event, such as a political event, it is specifically advantageous for a malicious actor to flood the platform with repetitive content, which can aid in spinning a desired narrative and creating the illusion of overwhelming support for their agendas. This can discourage or intimidate genuine users from expressing dissenting views, ultimately stifling open and balanced discussions. The following examples contain three retweets that were made by accounts identified by our system as well as those from known disinformation campaigns (i.e., 2019jan_iran and 2019jan_venezuela1).

Message 1: RT [REDACTED]: Al Assad: El ataque químico fue "ciento por ciento fabricado" y los reportajes son "falsos"
Translation: RT [REDACTED]: Al Assad: The chemical attack was "one hundred percent fabricated" and the reports are "false"
Message 2: RT [REDACTED]: La Patria de Bolívar rechaza la enfermiza obsesión de Luis Almagro contra Venezuela, exigimos respeto absoluto. Somos libres...
Translation: RT [REDACTED]: The Homeland of Bolívar rejects the sick obsession of Luis Almagro against Venezuela, we demand absolute respect. We're free...
Message 3: RT [REDACTED]: Este 19 de abril decimos con Alí: yo no me quedo en casa pues al combate me voy!. Vamos a la Calle, vamos a la Batalla, vamo...
Translation: RT [REDACTED]: This April 19 we say with Ali: I'm not staying at home because I'm going to fight! Let's go to the Street, let's go to the Battle, let's go...

The first message is regarding a "chemical attack" about which the mainstream media reports are flawed according to the tweet. Casting suspicion on authority is one of the major themes revolving around disinformation campaigns, as has been seen in past research (e.g., about mistrusting the

government on COVID-19 vaccines) [68]. The second message is a pro-Venezuela tweet, which is also a common theme amongst campaigns: to pick a side of the story and spread the message as much as possible. The last retweet is a call-to-action and can be considered an instigation to start protests. It is one of the key components of disinformation campaigns and is aimed at causing disruptive behavior. Past works have named this as *Rapid disinformation attacks*, whereby, disinformation is unleashed quickly to sabotage a crucial event (e.g., right before an important election) [58].

Advice Messages. Another important aspect of accounts involved in disinformation campaigns is their ability to slip under the radar and appear "legitimate." To do so, these accounts often employ deceptive tactics to gain the trust of their target audience. One such tactic is posting "advice" messages, which are designed to provide seemingly helpful or insightful information while subtly promoting a specific agenda. By posting advice messages, state-sponsored accounts might aim to establish themselves as trustworthy and knowledgeable sources. These messages also target the user's emotions and aspirations. In the past, various techniques have been used to fulfill this objective, ranging from posting cute dog pictures to sharing jokes [38]. As discussed in Section III-A, the commonly observed strategy is to post "advice" messages that are unrelated to any specific agenda(s) being pushed by the campaign. Following are two examples of such messages that were posted by accounts identified by our system and those in a known campaign (i.e., 2019jan_venezuela1):

> **Message 1:** Una persona que realmente te conoce es alguien que ve el dolor en tus ojos, mientras los demás creen que sonríes.
> **Translation:** A person who really knows you is someone who sees the pain in your eyes, while others think you smile.
> **Message 2:** La diferencia entre "puedo" y "no puedo" es solo de una palabra y una cuestión de actitud; si hay ganas, todo se puede.
> **Translation:** The difference between "I can" and "I can't" is only one word and a matter of attitude; if there is desire, everything is possible.

It is interesting to observe that these messages are copied and pasted verbatim, highlighting the possibility of a corpus of such messages that are sent through these accounts just to make them look like normal users. Both the messages are in Spanish, and therefore, we provide an English translation for them. The first message is more uplifting, and the next one is motivational. By posting advice messages, they can blend seamlessly with the larger user base, making it challenging to distinguish them from authentic accounts. This tactic might also lead to accounts remaining active for extended periods and continue their propaganda efforts undetected.

VII. Related Work

In this section, we examine previous studies that have investigated fake social media accounts and disinformation campaigns.

Detection of Fake Accounts. A wide variety of research looks at detecting fake accounts on social media. Some efforts have been made to determine features found commonly in fake accounts, e.g., a disproportionate friend-to-follower ratio or the similarity of posted content [6], [51]. More sophisticated systems, such as the one proposed by Yang et al., discovered more resilient features in fake accounts that are difficult for adversaries to evade (e.g., average neighbor's followers and following rate) [71]. Chu et al. propose a classifier to distinguish between a human, a cyborg, and a bot using a system built on various components (e.g., entropy and spam detection) [12]. Ghosh et al. delved into the phenomenon of *link farming*, which spam accounts use to amass a large number of followers [23], while Wang et al. analyzed user click patterns to create user profiles and employed supervised and unsupervised learning techniques to detect fake accounts [63]. Viswanath et al. utilized Principal Components Analysis (PCA) to identify patterns among extracted features from spam accounts [60] while Egele et al. concentrated on detecting compromised legitimate accounts, finding that regular users exhibit consistent habits over time, and any sudden deviations from these patterns indicate a compromise [16]. Davis et al. propose *BotorNot*, a system that uses over 1000 features (including Temporal, User, and Friend features) to detect whether an account is real or a bot [14]. Another distinction found between regular and fake accounts is *social connections*. In this direction, Danezis et al. applied a Bayesian Inference approach to detect compromised accounts by identifying communities with similar characteristics [13], while Cai et al. partitioned social networks into communities and sought out ones that displayed inconsistent connections with the rest of the network [8].

A separate line of research focuses on the synchronized behavior of fake accounts, which are commonly controlled by a single entity. Cao et al. propose *SynchroTrap*, a detection system that groups malicious accounts based on their synchronized actions and timing [9], while Stringhini et al. introduce *EvilCohort*, a system that identifies sets of social network accounts utilized by botnets by examining communities of accounts accessed by shared IP addresses [52].

Closely related to our work is a system proposed by Alhazbi et al. [1]. The authors use a set of behavioral features to detect state-sponsored troll accounts on Twitter. They train and evaluate on a set of Saudi trolls disclosed by Twitter in 2019, with an overall classification accuracy of up to 94.4%. However, in our work we identify features that are common across several campaigns spanning multiple countries. Our work also uses features from different modalities (e.g., temporal and stylometric) as we demonstrate the interplay of different features to capture different modus operandi used by troll accounts. We also demonstrate the efficacy of

our system by training on many different campaigns and identifying unseen state-sponsored accounts, thus generalizing the utility of our system beyond a single campaign. Another system proposed by Fornacciari et al. [21] uses six groups of features based respectively on the analysis of writing style, sentiment, behaviors, social interactions, linked media, and publication time to detect state-sponsored troll accounts on Twitter. The system, TrollPacifier, uses a neural network model and achieves a classification accuracy of 95.5%.

Disinformation Campaigns. Numerous studies have examined the role of social bots in the spread of political disinformation [7], [19], [20], [57]. This body of work demonstrates that bots are capable of large-scale public opinion manipulation, which may have an impact on important political events, such as election results. Mihaylov et al. show that there are two types of accounts involved in spreading disinformation: independent actors and those who are financially incentivized to propagate specific messages [33]. Steward et al. study Russian-sponsored troll accounts participating in the Black Lives Matter (BLM) discussion on Twitter and show that they infiltrated both left- and right-leaning communities with the objective of promoting particular narratives [49], while Ratkiewicz et al. use machine learning techniques to identify the spread of false political information on Twitter [37]. Zannettou et al. conduct a variety of studies examining state-sponsored troll accounts operating on Twitter and Reddit between 2014 and 2018 and evaluate their effectiveness in disseminating content across various platforms and web communities [72], [73], [74]. They show that the accounts involved in such campaigns are often created in *waves* and later present a pipeline that focuses on studying the images shared by these accounts on Twitter. Similarly, Hegelich et al. analyze the utilization of 1.7K Twitter bots during the Russia-Ukraine conflict and uncover a range of behaviors exhibited by these bots, such as attempts to conceal their identity, promotion of specific topics through hashtags, and the retweeting of messages with particularly engaging content [25]. More recently, Wang et el. [64] find some inter-state coordination patterns in state-backed operations on Twitter highlighting that influence campaigns attract greater attention than baseline information operations, however their contribution is a measurement instead of a detection system.

VIII. Discussion and Conclusion

In this paper, we study a publicly available dataset of state-sponsored coordinated campaigns on Twitter. These campaigns are geared towards spreading disinformation, trolling, and other disruptive and malicious behaviors. We show that there are patterns in the way these campaigns operate (e.g., use of scheduling services for automation and impersonating third-party applications). From our findings, we derive a machine-learning based model that can identify accounts from unseen campaigns. Unlike past research, we take a step towards building systems that can perform cross-campaign detection. We show that it is possible to build systems that are tailored towards the intricate nature of state-sponsored campaigns. Additionally, we find various instances of potential inter- and intra-state coordination (e.g., coordinated posting patterns, shared applications and copy-pasted messages), hinting at a larger market or shared modus operandi used by different state actors. I.e., it is highly unlikely that these campaigns are completely homegrown. In this direction, some past research has pointed towards black markets for reusable political disinformation bots [19]. We believe that it is important for future research to look at disinformation and influence campaigns through a larger lens.

Resilience to Evasion. Our multi-factor analysis approach is a key component of our detection system's resilience to evasion. The system considers multiple indicators of disinformation spread, including user metadata attributes, linguistic patterns, and temporal behaviors making the system less susceptible to evasion targeting specific features. Disinformation campaigns attempting to evade detection by using one specific technique may still exhibit other suspicious patterns that can be captured by the system. This multi-factor analysis helps reduce the effectiveness of evasion strategies and enhances the system's overall robustness.

For example, an attacker might attempt to avoid detection by minimizing one of the popular tactics (e.g., the use of scheduling applications to automate messages). Although it might avoid detection, it would make their campaigns much less effective since the messages would have to be manually sent and require much more effort, reducing the efficiency (and likely the efficacy) of the campaign to a large degree.

Another evasion strategy could be to reduce the frequency of retweeting or sharing each other's messages. However, this will also reduce the efficacy of their operations since the reach and visibility of the disinformation content will be diminished.

Implications for the Design of Disinformation Detection Systems. There are several implications for the design of new safety features for social media, especially platforms like Twitter, Threads, and Mastodon that can be derived from our findings. First, our analysis provides a crucial step toward identifying the characteristics of inauthentic accounts that spread disinformation to enable automated detection (machine learning based) of troll accounts in the future. In particular, we establish the distinct ways in which these accounts operate, strategize and often present themselves to appear human-like. Understanding how accounts belonging to state-sponsored disinformation campaigns attempt to mimic legitimate users can help inform behavioral models. These models can track account behavior over time and compare it to established patterns of disinformation spreaders. By continuously monitoring and analyzing user behavior, detection systems can adapt to evolving tactics, identifying previously unseen strategies.

Next, we use a combination of features that can help future systems distinguish troll accounts from legitimate ones. Rather than relying on individual signals alone, detection systems can combine multiple indicators to assess the likelihood of an account engaging in disinformation campaigns. This holistic approach increases the robustness and accuracy of the detection process.

It is also worth noting that while machine learning models play a crucial role in automated detection, human expertise is invaluable in fine-tuning and validating the system's output. Combining the insights gained from our research with human analysts' expertise can act as a force multiplier for the detection system.

Another implication for social media platforms is to be more cognizant of the idea of third-party agencies aiding different state actors in mass-producing disinformation. The presence of copy-pasted content, coordinated posting, third-party clients shared by multiple actors and other nefaroius methods being used should inform the development of future systems countering disinformation.

Limitations and Future Work. Like any other malicious actor, disinformation campaigns are likely to adapt to counter-detection efforts. Since these campaigns are ever-evolving and are now becoming *smarter* and more *outsourced*, thereby expanding their attack vector by using a variety of sources (e.g., blogs, websites, fake articles and dedicated accounts) [10], it is possible that the detection system's efficacy might drop against novel strategies in a long enough time period.

We advance the field of combating online disinformation by identifying emerging trends and tactics used by disinformation campaigns and using those to build a detection system that can effectively perform campaign-agnostic detection. However, we also argue that in order to maintain a robust defense, continuous monitoring and updates are essential to effectively address the changing climate of disinformation with the presence of LLM tools and the possibility of widespread disinformation [76], as is true for many information security related problems (ranging from fuzzers to spam detection to browser fingerprinting).

Another constraint in our study is the selection process of real accounts used for training our classifier in Section IV-B. We cannot guarantee with absolute certainty that the chosen set of accounts does not contain any accounts from a disinformation campaign.

In Section V-A, we test the system on a set of 2,696 accounts and identify 116 accounts that potentially belong to state-backed operations. However, we do not have definite proof of these accounts being malicious, but we illustrate their resemblance with the activity of accounts belonging to known campaigns through case studies in Section VI. A related limitation is the discontinuation of the Twitter API, leaving us unable to test our system on a larger, empirical dataset. Unfortunately, this is a problem that will be faced by all social media research moving forward, but at the same time, our dataset is "clean" in that it predates the widespread availability of generative text models like ChatGPT.

Future research could delve deeper into the mechanisms of coordination, investigate the objectives and motivations behind such convergence, and explore the impact of shared application usage on the effectiveness of disinformation campaigns. It is also imperative for researchers to look at disinformation campaigns from a broader perspective and understand the commonalities between them to build better detection systems. Additionally, understanding the role of platform policies, algorithmic biases, and countermeasures in mitigating the influence of state-sponsored disinformation remains an important area for further investigation.

Acknowledgments. This work was supported by the National Science Foundation under Grants CNS-1942610, CNS-2114407, CNS-2114411, CNS-2247867, and CNS-2247868. Any opinions, findings, and conclusions or recommendations expressed in this Report are those of the PI and do not necessarily reflect the views of the NSF.

REFERENCES

[1] S. Alhazbi. Behavior-based machine learning approaches to identify state-sponsored trolls on twitter. *IEEE Access*, 8:195132–195141, 2020.

[2] B. Allyn. Study Exposes Russia Disinformation Campaign That Operated In The Shadows For 6 Years. https://www.npr.org/2020/06/16/878169027/study-exposes-russia-disinformation-campaign-that-operated-in-the-shadows-for-6-, 2020.

[3] E. Antonov. Professions that people hate: Internet troll about the "factory", patriotism and Milonov. https://paperpaper-ru.translate.goog/troll/?_x_tr_sl=auto&_x_tr_tl=en&_x_tr_hl=en, 2017.

[4] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: Quantifying influence on twitter. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, WSDM '11, page 65–74, New York, NY, USA, 2011. Association for Computing Machinery.

[5] M. T. Bastos and D. Mercea. The brexit botnet and user-generated hyperpartisan news. *Social Science Computer Review*, 37(1):38–54, 2019.

[6] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, 2010.

[7] A. Bessi and E. Ferrara. Social bots distort the 2016 US Presidential election online discussion. *First Monday*, 21(11), 2016.

[8] Z. Cai and C. Jermaine. The latent community model for detecting sybil attacks in social networks. In *ISOC Network and Distributed Systems Security Symposium (NDSS)*, 2012.

[9] Q. Cao, X. Yang, J. Yu, and C. Palow. Uncovering large groups of active malicious accounts in online social networks. 2014.

[10] B. Chandra and L. N. Chao. Dismantling the Disinformation Business of Chinese Influence Operations. https://www.rand.org/blog/2023/10/dismantling-the-disinformation-business-of-chinese.html, 2023.

[11] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, page 925–936, New York, NY, USA, 2014. Association for Computing Machinery.

[12] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on Dependable and Secure Computing*, 9(6):811–824, 2012.

[13] G. Danezis and P. Mittal. Sybilinfer: Detecting sybil nodes using social networks. In *ISOC Network and Distributed Systems Security Symposium (NDSS)*, 2009.

[14] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. BotOrNot: A System to Evaluate Social Bots. In *The Web Conference (WWW)*, 2016.

[15] J. Echeverrï£¡a, E. De Cristofaro, N. Kourtellis, I. Leontiadis, G. Stringhini, and S. Zhou. Lobo: Evaluation of generalization deficiencies in twitter bot classifiers. ACSAC '18, page 137–146, New York, NY, USA, 2018. Association for Computing Machinery.

[16] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna. Towards detecting compromised accounts on social networks. *Transactions on Dependable and Secure Computing (TDSC)*, 2015.

[17] F. Ezzeddine, O. Ayoub, S. Giordano, G. Nogara, I. Sbeity, E. Ferrara, and L. Luceri. Exposing influence campaigns in the age of LLMs: a behavioral-based AI approach to detecting state-sponsored trolls. *EPJ Data Science*, 12(1), oct 2023.

[18] G. Fenwick. Flesch Reading Ease: Everything You Need to Know. https://writingstudio.com/blog/flesch-reading-ease/, 2020.

[19] E. Ferrara. Disinformation and social bot operations in the run up to the 2017 French presidential election. *ArXiv 1707.00086*, 2017.

[20] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini. The rise of social bots. *Communications of the ACM*, 2016.

[21] P. Fornacciari, M. Mordonini, A. Poggi, L. Sani, and M. Tomaiuolo. A holistic system for troll detection on twitter. *Comput. Hum. Behav.*, 89(C):258–268, dec 2018.

[22] P. Galán-García, J. Puerta, C. Laorden, I. Santos, and P. Bringas. *Supervised Machine Learning for the Detection of Troll Profiles in Twitter Social Network: Application to a Real Case of Cyberbullying*, volume 239, pages 419–428. 01 2014.

[23] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi. Understanding and combating link farming in the twitter social network. In *The Web Conference (WWW)*, 2012.

[24] J. A. Goldstein and S. Grossman. How disinformation evolved in 2020. https://www.brookings.edu/techstream/how-disinformation-evolved-in-2020/, 2021.

[25] S. Hegelich and D. Janetzko. Are Social Bots on Twitter Political Actors? Empirical Evidence from a Ukrainian Social Botnet. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2016.

[26] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intell. Data Anal.*, 6(5):429–449, oct 2002.

[27] S. Lewandowsky, U. K. Ecker, and J. Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition*, 6(4):353–369, 2017.

[28] A. Liaw and M. Wiener. The r journal: Classification and regression by randomforest. *R News*, 2:18–22, 2002. https://journal.r-project.org/articles/RN-2002-022/.

[29] Y. Liu and Y.-F. B. Wu. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[30] L. Luceri, S. Giordano, and E. Ferrara. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *ICWSM*, 2020.

[31] S. Mahtani and R. Cabato. Why crafty Internet trolls in the Philippines may be coming to a website near you. https://www.washingtonpost.com/world/asia_pacific/why-crafty-internet-trolls-in-the-philippines-may-be-coming-to-a-website-near-you/2019/07/25/c5d42ee2-5c53-11e9-98d4-844088d135f2_story.html, 2019.

[32] A. Martin. After more than 200 takedowns, Meta confirms covert online campaigns have gone global. https://therecord.media/after-more-than-200-takedowns-meta-confirms-covert-online-campaigns-have-gone-global, 2022.

[33] T. Mihaylov and P. Nakov. Hunting for Troll Comments in News Community Forums. In *ACL*, 2016.

[34] L. H. X. Ng and K. M. Carley. Botbuster: Multi-platform bot detection using a mixture of experts, 2022.

[35] S. I. Observatory. A US PR Firm Steps Into Contested Elections. https://cyber.fsi.stanford.edu/io/news/us-pr-firm-steps-contested-elections, 2020.

[36] F. Pierri, A. Artoni, and S. Ceri. Investigating italian disinformation spreading on twitter in the context of 2019 european elections. *PLOS ONE*, 15(1):1–23, 01 2020.

[37] J. Ratkiewicz, M. Conover, M. R. Meiss, B. Gonçalves, A. Flammini, and F. Menczer. Detecting and Tracking Political Abuse in Social Media. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2011.

[38] M. H. Saeed, S. Ali, J. Blackburn, E. D. Cristofaro, S. Zannettou, and G. Stringhini. Trollmagnifier: Detecting state-sponsored troll accounts on reddit. In *IEEE Symposium on Security and Privacy*, 2022.

[39] S. Safavian and D. Landgrebe. A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674, 1991.

[40] Scikit. Feature importances with a forest of trees. https://scikit-learn.org/stable/auto_examples/ensemble/plot_forest_importances.html, 2023.

[41] C. Shao, G. L. Ciampaglia, O. Varol, K. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots, 2018.

[42] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer. The spread of low-credibility content by social bots. *Nature Communications*, 9(1), nov 2018.

[43] C. Shao, P.-M. Hui, L. Wang, X. Jiang, A. Flammini, F. Menczer, and G. L. Ciampaglia. Anatomy of an online misinformation network. *PLOS ONE*, 13(4):1–23, 04 2018.

[44] A. Shevtsov, C. Tzagkarakis, D. Antonakaki, and S. Ioannidis. Identification of twitter bots based on an explainable machine learning framework: The US 2020 elections case study. *CoRR*, abs/2112.04913, 2021.

[45] L. Singh, S. Bansal, L. Bode, C. Budak, G. Chi, K. Kawintiranon, C. Padden, R. Vanarsdall, E. K. Vraga, and Y. Wang. A first look at COVID-19 information and misinformation sharing on twitter. *CoRR*, abs/2003.13907, 2020.

[46] F. B. Soares and R. Recuero. Hashtag wars: Political disinformation and discursive struggles on twitter conversations during the 2018 brazilian presidential campaign. *Social Media + Society*, 7(2):20563051211009073, 2021.

[47] K. Starbird. Examining the alternative media ecosystem through the production of alternative narratives of mass shooting events on twitter. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2017.

[48] K. Starbird, A. Arif, and T. Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 2019.

[49] L. Steward, A. Arif, and K. Starbird. Examining Trolls and Polarization with a Retweet Network. In *MIS2*, 2018.

[50] G. Stringhini, M. Egele, C. Kruegel, and G. Vigna. Poultry markets: On the underground economy of twitter followers. *SIGCOMM Comput. Commun. Rev.*, 42(4):527–532, sep 2012.

[51] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Annual Computer Security Applications Conference (ACSAC)*, 2010.

[52] G. Stringhini, P. Mourlanne, G. Jacob, M. Egele, C. Kruegel, and G. Vigna. {EVILCOHORT}: Detecting communities of malicious accounts on online services. In *USENIX Security Symposium*, 2015.

[53] G. Stringhini, G. Wang, M. Egele, C. Kruegel, G. Vigna, H. Zheng, and B. Y. Zhao. Follow the green: Growth and dynamics in twitter follower markets. In *Proceedings of the 2013 Conference on Internet Measurement Conference*, IMC '13, page 163–176, New York, NY, USA, 2013. Association for Computing Machinery.

[54] Y. Sun, A. Wong, and M. S. Kamel. Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, 23, 11 2011.

[55] J. A. K. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural Process. Lett.*, 9(3):293–300, jun 1999.

[56] Twitter. Twitter Moderation Research Consortium. https://transparency.twitter.com/en/reports/moderation-research.html, 2023.

[57] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini. Online Human-Bot Interactions: Detection, Estimation, and Characterization. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2017.

[58] J. Villasenor. How to deal with AI-enabled disinformation. https://www.brookings.edu/articles/how-to-deal-with-ai-enabled-disinformation/, 2020.

[59] P. Vishnuprasad, G. Nogara, F. Cardoso, S. Cresci, S. Giordano, and L. Luceri. Tracking fringe and coordinated activity on twitter leading up to the us capitol attack. *Proceedings of the International AAAI Conference on Web and Social Media*, 18:1557–1570, 05 2024.

[60] B. Viswanath, M. A. Bashir, M. Crovella, S. Guha, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In *USENIX Security Symposium*, 2014.

[61] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 2018.

[62] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[63] G. Wang, T. Konolige, C. Wilson, X. Wang, H. Zheng, and B. Y. Zhao. You are how you click: Clickstream analysis for sybil detection. In *USENIX Security Symposium*, 2013.

[64] X. Wang, J. Li, E. Srivatsavaya, and S. Rajtmajer. Evidence of inter-state coordination amongst state-backed information operations. *Scientific reports*, 13(1), Dec. 2023. Publisher Copyright: © 2023, The Author(s).

[65] J. Weerasinghe, B. Flanigan, A. Stein, D. McCoy, and R. Greenstadt. The pod people: Understanding manipulation of social media popularity via reciprocity abuse. In *The Web Conference (WWW)*, 2020.

[66] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, 10:207–244, jun 2009.

[67] Wikipedia. Kolmogorov-Smirnov test. https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test, 2022.

[68] M. D. Witte. Disinformation about the COVID-19 vaccine is a problem. Stanford researchers are trying to solve it. https://news.stanford.edu/press-releases/2022/02/24/curbing-spread-cs-disinformation/, 2022.

[69] Y. Wu, D. Guozhong, W. Wei, S. Guowei, G. Liangyi, Y. Miao, L. Jiguang, and H. Yaxue. Detecting bots in follower markets. *Communications in Computer and Information Science*, 472:525–530, 01 2014.

[70] Y. Xia, J. Lukito, Y. Zhang, C. Wells, S. J. Kim, and C. Tong. Disinformation, performed: self-presentation of a russian ira account on twitter. *Information, Communication & Society*, 22(11):1646–1664, 2019.

[71] C. Yang, R. C. Harkreader, and G. Gu. Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers. In *International Workshop on Recent Advances in Intrusion Detection*, 2011.

[72] S. Zannettou, B. Bradlyn, E. De Cristofaro, G. Stringhini, and J. Blackburn. Characterizing the use of images by state-sponsored troll accounts on twitter. In *AAAI International Conference on Web and Social Media (ICWSM)*, 2019.

[73] S. Zannettou, T. Caulfield, E. De Cristofaro, M. Sirivianos, G. Stringhini, and J. Blackburn. Disinformation warfare: Understanding state-sponsored trolls on twitter and their influence on the web. In *WWW Companion*, 2019.

[74] S. Zannettou, T. Caulfield, W. Setzer, M. Sirivianos, G. Stringhini, and J. Blackburn. Who let the trolls out?: Towards understanding state-sponsored trolls. In *ACM Conference on Web Science*, 2019.

[75] S. Zannettou, M. Sirivianos, J. Blackburn, and N. Kourtellis. The web of false information. *Journal of Data and Information Quality*, 11(3):1–37, may 2019.

[76] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi. Defending against neural fake news. *Advances in neural information processing systems*, 32, 2019.