Comprehensive Analysis of Two Malicious Arabic-language Twitter Campaigns

Reem Alharthi, Areej Alhothali, and Kawthar Moria.
Department of Computer Science, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia.
Emails: ralharthi0111@stu.kau.edu.sa, aalhothali@kau.edu.sa, kmoria@kau.edu.sa.

***Abstract*—Fake malicious accounts are one of the primary causes of the deterioration of social network content quality. Numerous such accounts are generated by attackers to achieve multiple nefarious goals, including phishing, spamming, spoofing, and promotion. These practices pose significant challenges regarding the availability of credible data that reflect real-world social media interactions. This has led to the development of various methods and approaches to combat spammers on social media networks. Previous studies, however, have almost exclusively focused on studying and identifying English-language spam profiles, whereas the problem of malicious Arabic-language accounts remains under-addressed in the literature. In this paper, therefore, we conduct a comprehensive investigation of malicious Arabic-language campaigns on Twitter. The study involves analyzing the accounts of these campaigns from several perspectives, including their number, content, social interaction graphs, lifespans, and day-to-day activities. In addition to exposing their spamming tactics, we find that these spam accounts are more successful in avoiding Twitter suspensions than has been previously reported in the literature.**

***Index Terms*—Social network security, Social spammers, Arab Twitter users, Malicious campaigns on Twitter, Data mining.**

I. Introduction

Social media networks have profoundly affected life today and have become massive platforms for communication and information exchange.A remarkable example of this influence is the role of social media in the Arab Spring, which was extensively reported in the literature [1], [2], [3].People have embraced such web platforms as independent media not subject to political parties or organizations. Today, in a highly volatile political setting, social media networks continue to play the same critical role in the Arab world. Researchers also utilize this social media data to extract information for various objectives such as opinion mining for the Arabic language [4], event detection [5], [6], and rumor detection [7].

The number of Arabic-language users on social media networks, including ill-intentioned individuals, increased tremendously after the Arab spring [8]. Some users misuse these websites to share inappropriate, deceptive, and offensive material, for intrusive advertising, public opinion manipulation, and to spread malicious malware, for example. They usually manage an enormous number of accounts known as campaigns and employ different spamming strategies to achieve their goals. Besides restricting freedom of expression, the quality of social media content as an informative tool used for various purposes is inevitably diminished by these actions. There exists an extensive literature on this topic [9], indicating that the problem of malicious users or campaigns is not recent; however, for Arab users, few studies have been presented to detect such activities at the account and individual tweet levels. Although some attempts have been made to analyze Arabic-language social media spammers [10], [11], this area is still insufficiently explored, as only traditional spammers have been investigated. Besides this and to our knowledge, there is no previous work investigating Arab spammers at the campaign level or investigating various aspects such as their spamming strategies. Therefore, an extensive analysis study has particular importance given the recent aggressive activity of malicious campaigns in the Arabic-language Twittersphere, where every day, malicious users flood trending topics with a tremendous amount of spam and low-quality content (more details in Section IV-B3).

This paper presents an empirical analysis of two malicious Arabic-language campaigns on Twitter. Compared to other groups in our dataset, they have the highest numbers of fake accounts and the longest periods of such harmful activities. The overall goal of this analysis is to extensively investigate the content and behavior of the two groups, including their numbers of accounts, tactics, lifespans, and methods used to control their large numbers of profiles. Accordingly, six hundred profiles of the two campaigns have been analyzed and split into two generations: one generation's activities were recorded for Apr-Dec 2018 and the second one's for Aug-Sep 2019. Through this analysis, many results are revealed regarding malicious Arabic-language groups, including how they coordinate accounts in clusters, how they use the clusters to target trending topics, their group interaction characteristics, the content characteristics described by the tweet functions, and the self-similarity ratios of the profiles.This study also addresses the expected lifetimes of such accounts as well as how they manage to run a large number of accounts (i.e., manually or using software).

Having greater insight into the activity of malicious campaigns will provide useful information about the quality and credibility of social network data. Such information is essential not only for the spam detection field but also for other areas that rely on social network data. Thereby, this study attempts to provide an in-depth insight into the current situation of Arabic-language trending topics and malicious campaigns, and more specifically, their strategies for abusing trending topics to maximize the distribution of their content as well as identifying the primary characteristics of such accounts and comparing them with the features of spammers as reported in the literature. In addition, through a two-month experiment, this study tries to analyze the lifespans and daily activities of the campaign accounts, as well as the Twitter system's ability to detect Arabic-language social media spam accounts.

The rest of this paper is structured as follows: *Section II* describes previous related works, and *Section III* discusses the dataset collection process. *Section IV* is divided into three major sections: the first introduces the number of campaigns' accounts (*Section IV-A*), the second provides an in-depth analysis of the main groups' characteristics, including the groups' interaction graphs, spamming strategies, and content attributes (*Section IV-B*), and (*Section IV-C*) discusses the practice of managing spam accounts. *Section V* summarizes the findings of this in-depth study. *Section VI* provides a summary of this paper, including the conclusion and suggestions for future work.

Finally, we note that a part of this paper appeared previously as a conference publication [12]. This part was included in the Data Analysis section in the conference paper, in which we briefly presented the clusters' organization and the automated behavior of the campaigns' accounts. Our main contributions for the journal version include an expanded detailed analysis that covers several aspects of these groups.

II. Related Works

A. Malicious Campaign Studies

Detecting individual malicious accounts on the scale of an entire social network is a costly and time-wasting approach. A study [13] highlighted the importance of addressing malicious accounts at the group level, which is often a more feasible and effective solution. The group-level detection approach identifies campaign accounts according to their common materials or objectives, which in turn raises the bar for the attackers to evade detection. Creating unique content or running every single account separately to hide the similarity among the group would greatly increase the costs and time to administer these accounts; therefore, researchers have suggested several systems and strategies expose various types of malicious accounts at the campaign level. For instance, a study [14] proposed examining the social graph between users and pages to reveal Fake-Likes campaigns on Facebook, and several studies have used the synchronized behavior and timing of social spammers' fraud activities, fake Twitter followers, and malicious retweeter groups to expose their accounts on Twitter [15], [16], [17], [18], [19]. Besides, a variety of analysis studies have been carried out to understand the various aspects of social spammers. Yang et al. [20] examined spammers' social graphs to identify the relationships and supporters of these accounts and showed that they are socially connected in communities and, in a number of cases, are supported by legitimate accounts. A study [21] investigated spammers' strategies to enhance their influence scores by following real users as well as each other on Twitter. Lastly, Gupta et al. [22] conducted a large-scale analysis study of spam campaigns on multiple social platforms that used telephone numbers to lure victims.

B. Detecting Spam in Arabic Social Networks Content

Most of the literature has focused mainly on English-language spammers and has made fewer attempts for non-English language users, even though spammers' strategies probably vary from one region to another given the fact that they evolve and find a new spamming method over time [23]. This section provides a brief overview of the body of related work, with a focus on detecting Arabic-language spammers in social networks.

One of the earliest empirical analyses of Arabic-language spammers presented by Al-Khalifa et al. [10], in which the authors examined the content and social graphs of these accounts, showed that they are still naive. A comprehensive investigation conducted by [11] studied the characteristics of long-surviving but eventually suspended Arabic Twitter accounts, and in this study, the authors compared these accounts with short-lived suspended accounts in terms of their content, activity, and linguistic attributes. Accordingly, they found that the short-lived group had a high self-similarity ratio compared to the long-lived group. Regarding the degree of activities such as gaining more followers or friends and posting tweets, the long-lived group was more active, and meanwhile, they avoided excessive behavior such as posting large numbers of tweets.

Research by [24] reported that spam tweets constitute about three-quarters of Saudi Arabia's trending tweets. The study also assessed the efficiency of well-known features that are designed based on English spam profiles in detecting Arabic spam accounts. First, they selected a range of features that combine profile and content characteristics to reflect the reputation and replication characteristics of an account, and then they compared the performance of the selected features and the model and features proposed by [25] that was also tested on the Arabic dataset. The results indicated that the selected features performed better than the model proposed by [25] in detecting Arab spam profiles. To classify spam at the tweet level, the authors in [26] designed a classification scheme that used content-based features such as the number of URLs, hashtags, phone numbers, and spam words present in a tweet.

Considering that automation technology can be used for malicious purposes such as spamming, an attempt by [27] was made to expose automated Arabic tweets. The proposed system tested different factors to identify a tweet as an automatic or manually generated tweet. In addition to the degree of formality of the tweet, several structure-based features such as the length of the tweet are employed in the classification decision. Nonetheless, automated tweets need not necessarily be malicious tweets, as there were no specific rules in the article to differentiate between malicious and benign automated tweets.

III. Datasets Collection

To collect a sufficient set of data, a Python crawler that uses Twitter API functions is developed to gather the required information. The crawler was first used to randomly collect tweets and users from Saudi Arabian trending topics, from which we excluded non-Arabic profiles. Using the crawler over $4,000$ different identifiers and $160,000$ tweets are assembled. We then manually annotated $1,000$ legitimate accounts out of the random set of users. To investigate malicious campaigns, a collection of accounts exhibiting spam-like behaviors such as sharing duplicate tweets and URLs were also gathered from trending topics in Saudi Arabia. Following that, we looked at the content and practices of these accounts, classifying individuals with similar or duplicate materials as a group, which is a sample strategy used by several previous studies [22], [28], [29]. The dataset eventually located two campaigns that stood out significantly from the rest of the collected groups. The first campaign consisted of 200 spamming accounts that shared a duplicate URL to an external website. The second group also had 200 accounts, and their work focused on the promotion of unfamiliar, unlicensed medicinal brands. Aside from having the longest periods of such harmful activities, two campaigns were discovered to have the highest number of fake accounts when compared to other groups. We, therefore, chose to focus our attention on these two campaigns (spammers and promoters). In addition, 200 accounts from second-generation (spammer and promoter) campaigns are added to the dataset in order to investigate them in depth and track their behavior over time.

IV. In-depth Analysis

A. The Number of Campaigns' Accounts

Our preliminary analysis of the number of accounts corresponding to the spammers' and promoters' campaigns shows that at any given time, they operate numerous accounts to spread their content, and that suspended or deleted campaigns' accounts will constantly be replaced by a new generation. As shown in Table I, the spammers' accounts (S_G_1), which had activity from Oct to Dec 2018, were all eventually suspended by Twitter, and the attackers replaced them with (S_G_2). The same thing goes for the promoters' group (P_G_1) and their second-generation (P_G_2). The accounts shown in Table I comprise the total number of accounts analyzed in the course of this study and do not reflect the actual number of campaign accounts.

For both campaigns, as Table I shows, the age of the accounts is above $1,500$ days (4 years) on average. By contrast, several studies for non-Arabic spammers found that these accounts tend to have a young age, of less than 200 days [25], [30], [31]. The long lifetime of the spammers' accounts implies that these accounts are compromised accounts, i.e., accounts that have been stolen from legitimate owners and which exhibit dramatic changes in behavior and content patterns [32], [33]. To confirm that,the account's tweets and profile are manually inspected to see if there was any evident behavioral change such as excessive posting rate, sharing spam URLs, and sharing duplicate content. The following points summarize the results of the experiment:

- According to the findings of this study, the majority of campaign accounts shared duplicated tweets in their most recent activities, whereas the first tweets were genuine tweets that did not include spam or duplicated links. In addition to the duplicated content, the tweet source, which is the utility used to post the tweet, was the most common sign of behavioral changes that we observed in our dataset.
- Through additional content analysis, two classes of account are identified; the first class includes accounts in which the old tweets (genuine tweets that are not duplicated, nor contain a spam link) are still there, and represents about 36.52% of the accounts in our dataset, and the second class involves accounts in which the old tweets were deleted (approximately 63.74% of the accounts). For the second class, we found several accounts that had explicitly used some service (such as Tweet Delete) to automatically delete all the old tweets[1].
- By analyzing the activity timelines of the campaign accounts, significant time gaps in the histories of these accounts can be observed which range between six months and four years; in the case of profiles in the first class particularly, the time between the last genuine tweet and the first spam tweet (Figure 1(a)), and in the case of accounts in the second class, the time between the account's creation date and the first tweet date, as shown in Figure 1(b). As Figure 1 shows, the creation date and the last genuine tweet most frequently occurred before 2018, and all the spam tweets were sent by the end of 2018 or in 2019.
- Figure 2 shows an example of a compromised account used in spamming activity. There is a significant difference in the behavioral patterns of the account's tweets before 2015 and the tweets in 2018, which involve duplicated tweets and using a different tweet source.
- The time gaps and the accounts' behavioral changes provide clear evidence that most of the accounts used by the two campaigns are compromised accounts.

B. The Main Characteristics of Malicious Groups

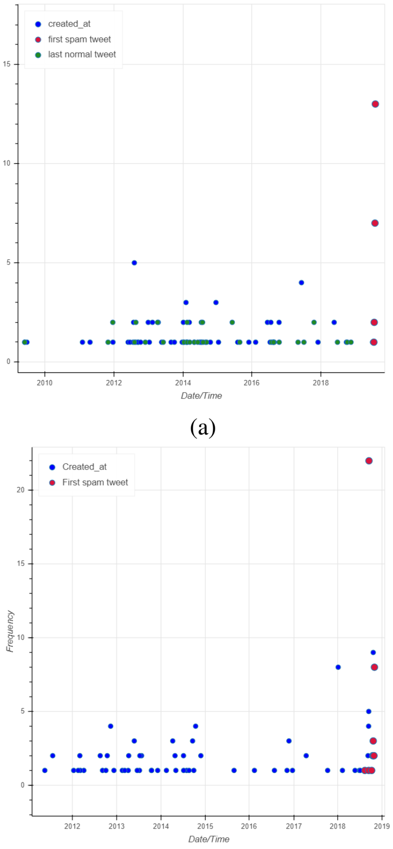
The results for the main groups' characteristics are presented in the subsections below, where various analytical methods are used to investigate the groups' interaction graphs, spamming strategies, and content attributes.

1) Groups' Interaction Graphs: In contrast to traditional spammers who aggressively and randomly share unsolicited content, both the promoters' and spammers' groups have shown high organizational levels. The two groups manage their large numbers of accounts in small-scale clusters of connected accounts, with about 4 to 18 profiles in each cluster. Individual clusters work autonomously toward a specific goal, e.g.,

[1] https://tweetdelete.net/

TABLE I: Group Descriptions and Metadata Statistics

	Activity Type	Activity Duration	Number of Accounts	Account Age	Followers	Friends
(S_G_1)	Spreading spam URLs	Oct-Dec 2018	200	1,604 days	116.32	236.85
(S_G_2)	Spreading spam URLs	Aug-Sep 2019	100	1,511 days	178.35	142.43
(P_G_1)	Promoting unlicensed medicines	Apr-Oct-Dec 2018	200	1,894 days	53.29	134.95
(P_G_2)	Promoting unlicensed medicines	Aug-Sep 2019	100	1,510 days	138.98	416.32



(a)

(b)

Fig. 1: (a) Class 1. Accounts for which old tweets were not deleted; (b) Class 2. Accounts for which old tweets were deleted

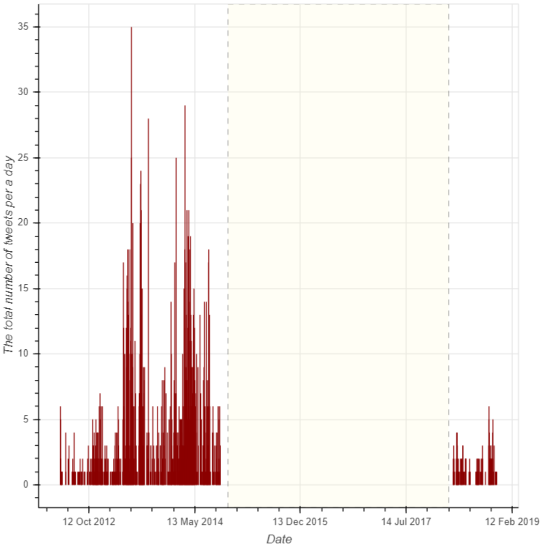


Fig. 2: An example of a spam account activity over time; last genuine tweet was on Oct 1, 2014, and first spam tweet occurred on Mar 26, 2018. Twitter for iPhone is the source of the last genuine tweet, and Twitter Web Client is the source of the spam tweets.

each group in a promoters' campaign promotes a particular medicine with the same brand name. Similarly, in a spammers' campaign, each cluster takes advantage of specific trending stories and encourages people to follow a hyperlink (a spam URL) to learn more details about the trending topic. Despite that, not all the groups are entirely independent in terms of their content, as we found that some of the groups share similar content.

A distinct spamming strategy is identified by studying the campaign clusters and their account interactions. This strategy is mainly developed to invade and flood hashtags and trending topics with unsolicited tweets. The accounts are organized and assigned to a specific social role inside a cluster, as follows. The clusters involve one or more central accounts whose role or task is posting original (spam or promoting) tweets. Then, a set of accounts regularly retweet, replay, and generally interact with whatever the central accounts share. Attackers with this strategy are guaranteed to reach a large group of users, especially in trending topics. This tactic will also promote spam to the "best tweets" tab by manipulating Twitter's tweet-rating tools, which assess tweets according to overall engagement, for instance, the number of retweets, likes, or replies. It is worth mentioning that the "best tweets" tab is the default tab for trending topics and the place where people often look for the most influential tweets.

In a general sense, the campaign accounts work or interact together within clusters. To visualize their organized behavior and interactions, a campaign interaction graph is constructed, which is defined as G=(V,E), where V is the graph's vertices/accounts and E is the edges that connect two vertices if there is an interaction between them [23]. In the interaction graph, three types of interaction between accounts is defined: retweets, replies, and mentions. We utilized the Networkx package[2] to build an undirected graph that shows the accounts' interactions from the topological point of view. The observations about the interaction graph are discussed in the following points:

[2] https://networkx.github.io/

- Because the clusters are often independent of each other and follow the same strategy, we have chosen to visualize the interaction graph at a cluster level rather than visualize all the campaign accounts' interactions. In addition, the interaction graph for three clusters is constructed to demonstrate the differences between the genuine and spammer classes: spammers, promoters, and genuine, with equal numbers of tweets and accounts. The clusters' interaction graphs were built according to their accounts' most recent 20 tweets ((Figure 3 (a), Figure 4 (a), and Figure 5 (a)) or their overall tweets ((Figure 3 (b), Figure 4 (b), and Figure 5 (b)) (see Figure 3 (a ,b), Figure 4 (a ,b), and Figure 5 (a ,b)).
- The groups organized behavior is very clearly shown in Figure 3 (a), and Figure 4 (a), in which accounts intensely interact with the central accounts in the clusters (the red node is the main central account). For the spammers' cluster, 26 nodes/accounts and 144 edges/interactions for nine accounts are found in their last 20 tweets, and similarly, there are 21 nodes/accounts and 104 edges/interactions in the promoters' graph. In contrast, in the genuine graph Figure 5 (a), there are 49 nodes/accounts and 49 edges/interactions, which indicates a more genuine or organic behavior.
- To assess graph connectivity, Table II presents the average clustering coefficients of the spammer and promoter groups' graphs. A clustering coefficient (CC) is a measure that indicates if the graph nodes are part of a highly connected graph [34]. Despite the large number of edges/interactions in the spammer and promoter groups', the average CCs of the accounts is zero, which indicates that their graph topology is a star. In other words, all the accounts' interactions are directed to the central accounts, and the central accounts do not interact with each other.
- As stated in Table II, the graph diameter of the spammer and promoter groups is equal to two, which indicates that there is a node that connects with every other node in the graph. Clearly, the central accounts are the nodes that connect all the other accounts, which is also reflected by the maximum degree of centrality and maximum closeness centrality properties of the groups' graphs.

Remarkably, we found a few isolated spam and promoter accounts that operated as a single account and posted tweets without further interaction with other accounts. Also, several promoters' accounts that had a strategy different from what has been discussed in this section are found. They abuse the "mention" function to reach a particular audience or user in the trending topics rather than many users. More specifically, they mention or reply to popular accounts or top tweets in a specific topic with their promotion tweets, which is a well-known strategy for spammers [35].

2) Groups' content attributes: Several previous studies have highlighted the importance of content-based features [31], [36], [37] in identifying social media spammers. Generally, the content or language model of the spammers is significantly different from genuine accounts' content as a

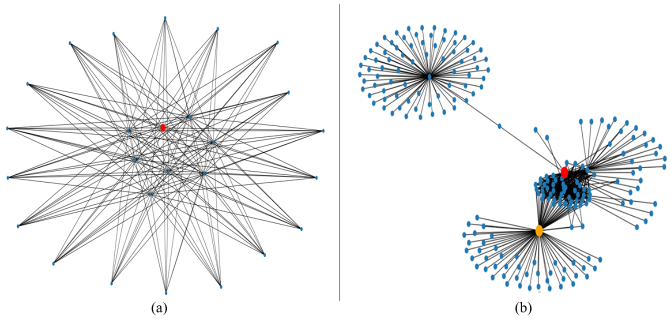


Fig. 3: (a) The interaction graph of 9 spam accounts according to their most recent 20 tweets. (b) The interaction graph for the same spam group according to all their tweets

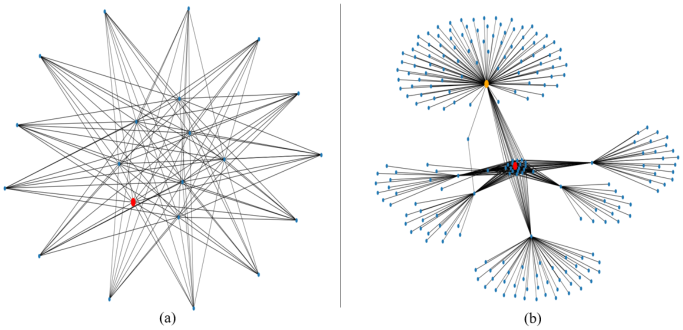


Fig. 4: (a) The interaction graph of 9 promoters' accounts according to their most recent 20 tweets. (b) The interaction graph for the same promoters' group according to all their tweets

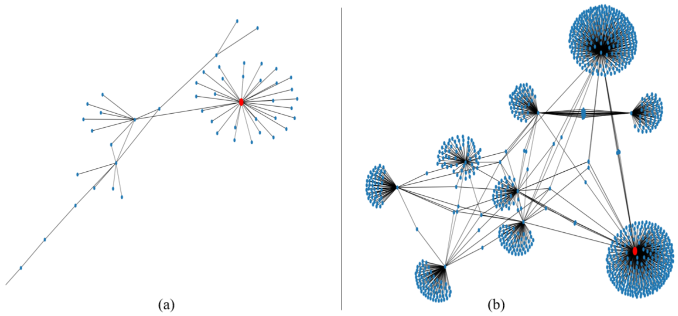


Fig. 5: (a) The interaction graph of 9 genuine accounts according to their most recent 20 tweets. (b) The interaction graph for the same genuine group according to all their tweets

TABLE II: Groups' Graph Properties.

	Spammers		Promoters		Genuine	
	(a)	(b)	(a)	(b)	(a)	(b)
Clustering Coefficient	0	0	0	0	0.0292	0.00889
Average Degree	11.1	4.9	9.9	3.1	2	2.1
Graph Diameter	2	6	2	4	7	6
Maximum Degree Centrality	0.72	0.42	0.65	0.5	0.61	0.31
Maximum Closeness Centrality	0.78	0.41	0.74	0.51	0.56	0.46
Maximum Betweenness Centrality	0.06	0.59	0.05	0.66	0.84	0.52

result of their distinct ill-intentioned use of social networks. In this way, they attempt to maximize their content distribution by intensively posting duplicate texts and URLs or aggressively exploiting the network's services, e.g., hashtags, mentions, URLs, and photos. Therefore, in this section investigates the content characteristics of Arabic spam campaigns in two aspects: using the tweet functions and the self-similarity and word frequency of the spammers' language.

Various statistics from the accounts' tweets for the content attributes are first collected, such as the numbers of URLs, photos, and hashtags. Figure 6 (a, b, c, and d) plotted a cumulative distribution function (CDF) for the content attributes to compare between the spammer and promoter groups and the genuine accounts in our dataset. The points listed below discuss the findings:

- Spammer groups exhibit aggressive behavior in using most of the tweet functions, in which they post many links, hashtags, and photos per tweet. Additionally, their content attributes vary considerably from the genuine accounts and promoters' accounts, as shown in Figure 6 (a, c, and d).
- Promoters' accounts show similar patterns across multiple attributes to the genuine accounts as opposed to spammers, as shown in Figure 6 (a, b, and c).
- As shown in Figure 6 (d), the unique URL ratio exhibits the highest divergence between the three classes in our dataset.

To estimate the semantic similarity of pairwise tweets for the campaign accounts, we take the average of the word embeddings of all words in the pair tweets and then compute the distance between the resultant sentences' vectors by using cosine similarity. The pre-trained word2vec model [38] is used to obtain the embedding vectors of the words. As shown in Figure 7 (a), the self-similarity [11] of 40% of both the spammers' and promoters' accounts is greater than 0.7, while less than 1% of the genuine accounts reach the same percent of similarity. Additionally, the new generations of the campaign accounts follow almost the same distributions as the old suspended accounts, as shown in Figure 7 (b). The high self-similarity ratio suggests that these accounts are designed to deliver or distribute one message, which conforms with our findings in the previous section. More precisely, these accounts concentrate on promoting, for example, a particular service or medicine in the case of promoters' accounts or taking advantage of a specific controversial story in the case of spammers' accounts.

- In addition to their high self-similarity, the spammers' and promoters' accounts often use the same sets of words to deliver their messages. Figure 8 shows the average of newly introduced words in the accounts' tweets throughout 100 tweets. As Figure 8 shows, genuine accounts use 6 to 8 new words in each new tweet, while both the spammers and promoters introduce two new words over their new tweets, which are more likely to be either keywords or hashtags in trending topics.
- A further interesting observation is that spammers' campaign tweets mostly relate to trending topics or hashtags. For example, if there is a trending story or viral news, these accounts usually claim that their spam URLs provide more information about the story. This stands in contrast to many previous studies that have described the spam tweets as tweets that are irrelevant to the topics [9], [23], [35]. In general terms, both campaigns post content that is easily detectable and varies from the material of real users.

3) The Campaigns' Accounts Lifespan and Daily Activities: As previously mentioned in Section IV-A, the two campaigns constantly replace the suspended accounts with new ones. As a result, they have the most extended durations of spamming activities and the most significant numbers of fake accounts (see Section IV-A). This section addresses the accounts' lifespans in order to answer the following questions: (1) how do attackers leverage these accounts to obtain the maximum possible output? (2) what are the average lifespans of these accounts? (3) how can these accounts avoid the Twitter detection system? (4) among the different kinds of spammer and promoter activities, which ones are the fastest to be detected by Twitter? And (5) among the spammer and promoter accounts, which ones can survive suspension for more extended periods of spamming activity?

To answer these questions, we conducted a two-month investigative study of the second-generation of the spammers' (S_G_2) and (P_G_2) promoters' groups (see Table I). During the two months, we followed or recorded the daily activities of these accounts, including their new tweets (i.e., original, retweet, and mention tweets), numbers of total tweets, numbers of followers and friends, and the current state of the accounts (i.e., active or suspended). Additionally, to examine their social interaction patterns, we tracked new tweets, total received retweets, and favorites.

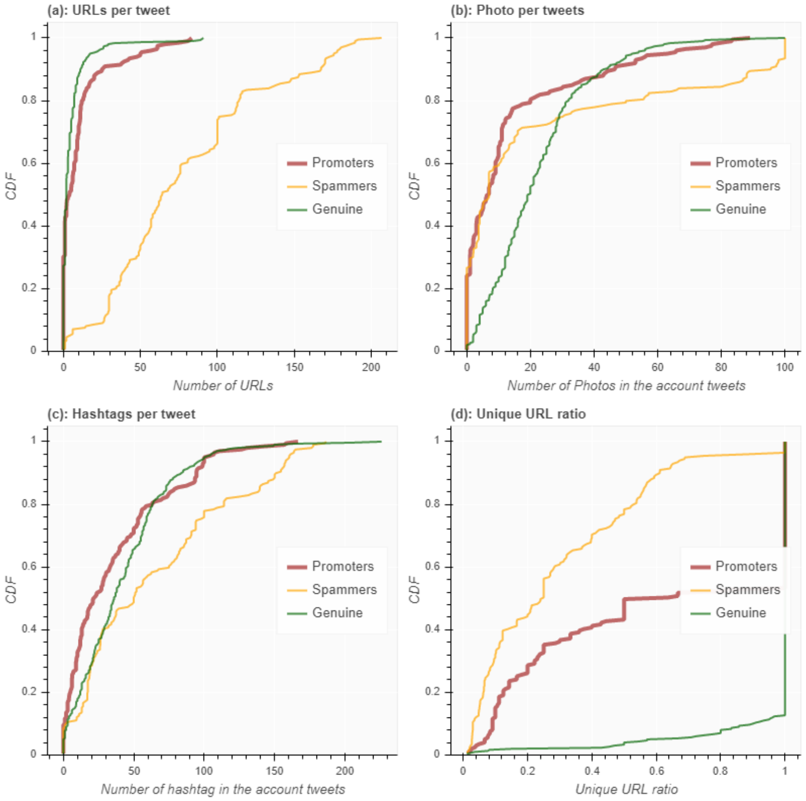


Fig. 6: (a). Number of URLs contained in the accounts' 100 tweets. (b). Number of photos in the tweets. (c). Number of hashtags contained in the accounts' last 100 tweets. (d). Number of unique URLs vs. number of shared URLs.

Over the two months, the 200 accounts of the (S_G_2) and (P_G_2) produced a total of $243,037$ low-quality tweets, as illustrated in Figure 9. The spammers' group accounted for a large percentage, at nearly 72% of total spam tweets. The fact that this vast number of tweets was generated by a subset of the actual number of campaigns' accounts is particularly alarming. We believe that the two campaigns had a much larger number of fake accounts, from which the 200 accounts are collected by searching a few keywords and trending topics. This massive number of spam tweets reflects the current crisis of Saudi Arabian trending topics, where such accounts flood the trends with disturbing and unsolicited content on a daily basis [39]. Twitter, on the other hand, had succeeded in suspending 55% of the campaigns' accounts over the two months but failed to identify about 45% of the total number, which were still active up to the last day of the experiment.

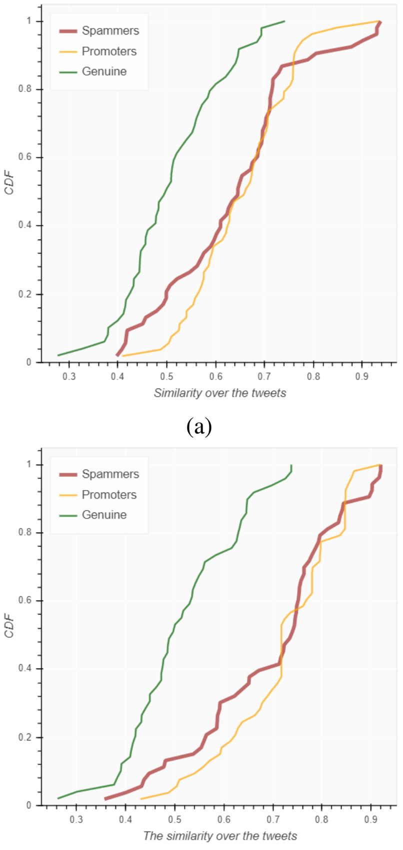
The average, median, and maximum number of tweets per day for the suspended and active accounts are computed to compare the posting ratios of the two groups. In addition, the posing ratios of the genuine accounts is estimated by computing these statistics for $1,000$ labeled accounts from our previously collected dataset (see Section III). As shown in Table III, the spammers continued to exhibit aggressive behavior, with the highest posting ratio among the three classes. Surprisingly, the active of (S_G_2) accounts were more aggressive than the suspended group, with an average of 90.21 tweets and an account share of over $6,143$ spam tweets per day. The promoters group (P_G_2) again showed a pattern that closely matched the genuine accounts, with an average of 22.57 tweets per day. Regarding the groups' social graphs, almost 90% of the accounts maintained the same numbers of followers and friends, and the rest of the accounts' numbers actually decreased during their spamming periods. That indicated that these accounts were aimed at targeting a broad audience in the trends instead of using the "follow" function to reach for specific victims.

We defined the lifespans of these accounts as the duration between the date of the first tweet that violated Twitter rules and the date of the last tweet. We believe that this provides a more precise definition of malicious account activities than the interval between the account creation date and the that of the latest tweet, since the interval time would incorporate all an account's activities, including its genuine early stage in the case of compromised accounts (see Section IV-A). Additionally, this definition gives a more accurate description of how long these accounts can avoid suspension after their first spam tweets.

Table IV provides the average, median, and maximum duration of activities in days for the second-generation of spammer and promoter accounts. In the case of suspended accounts, the duration is the time between the first spam tweet (most of these tweets occurred after Aug 1, 2019) and the date

TABLE III: Daily Average of Tweets for the Groups' Accounts and Genuine Accounts

		Average	Median	Maximum
(S_G_2)	Suspended accounts	59.22	29.5	707
	Active accounts	90.21	37.5	6143
(P_G_2)	Suspended accounts	31.11	25	229
	Active accounts	22.57	12.75	396
Genuine Accounts	-	23	8	100



(a)

(b)

Fig. 7: (a). Comparison between the three classes' accounts (spammers, promoters, and genuine) in terms of their self-similarity ratios. (b) Comparison between old suspended generation of campaigns and new active accounts.

of the last tweet before the account is suspended. For the active accounts, the duration is the time between the first spam tweets and the latest tweet shared by the account, and the accounts were still active until the last day of the experiment. As shown in Table IV, the average lifespan of the (S_G_2) accounts in the suspended group was less than that of the suspended (P_G_2) accounts, which is an expected outcome due to their aggressive behavior. Also, Twitter could detect the (S_G_2) accounts faster than the accounts of (P_G_2), as shown in Figure 10. Even though the active (S_G_2) accounts had a longer average lifespan than the (P_G_2) active accounts, and their total number of accounts was at some point greater than the (P_G_2) accounts during the experiment (see Figure 10), the promoters' campaign was more successful in leveraging their accounts. In consequence, the spammers' campaign exhibited very easily detectable behavior, as a result of which 60% of their accounts were suspended either on the first day or in less than five days (see Figure 13 (c)). Furthermore, this study discovered that the long-lived spammers' accounts were "sleepers," which meant that their activities would pause for a short period of time. The promoters' accounts, on the other hand, posted fair numbers of tweets daily; consequently, their activities (including their tweets that were still public in the trends) persisted for a more

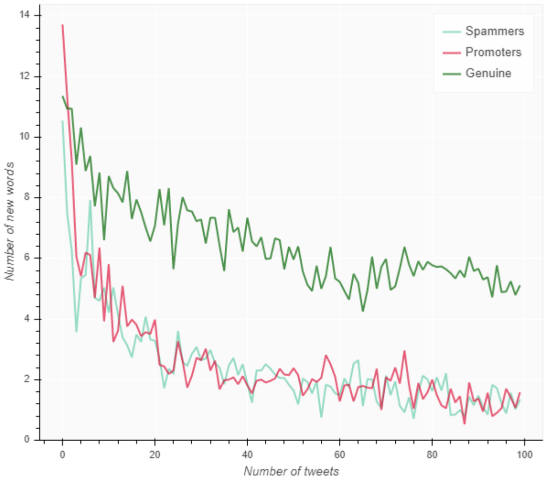


Fig. 8: Average number of new words over the accounts' last 100 tweets.

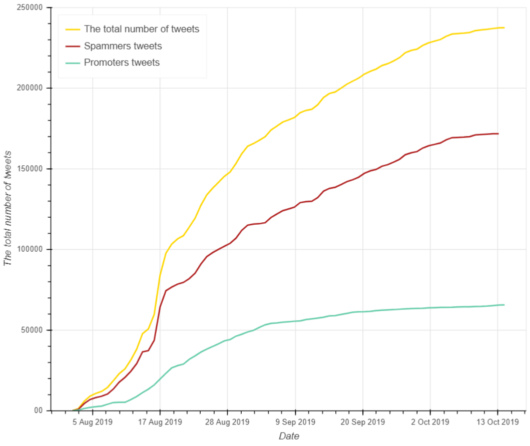


Fig. 9: Total number of (S_G_2) and (P_G_2) tweets over two months

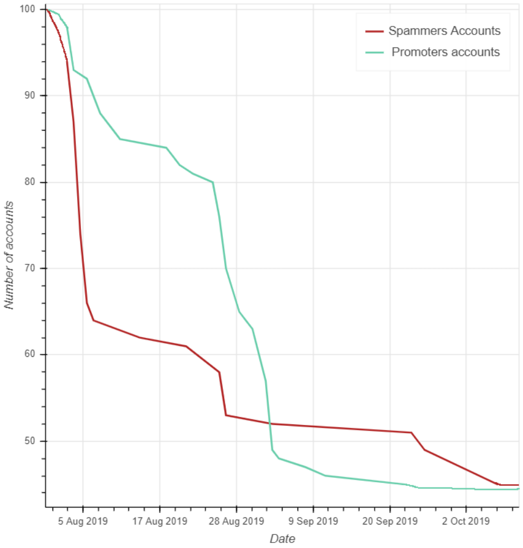


Fig. 10: Illustration of total number of (S_G_2) and (P_G_2) accounts over the two months

TABLE IV: Spammers' and Promoters' Accounts' Activity Duration in Days

		Average	Median	Maximum
(S_G_2)	Suspended accounts	16.75	3	190
	Active accounts	42.56	16.5	269
(P_G_2)	Suspended accounts	20.22	23	51
	Active accounts	21.18	6.5	73

extended period than those of the spammers. As shown in Figure 13 (b and d), only 18% of the promoters' accounts were suspended within 5 days of their activities.

Also, the sets of the suspended accounts in both campaigns are examined to clarify the relationship between the accounts' lifespans and posting patterns. The primary assumption of this experiment was that accounts with high posting ratios would be more likely to be identified than those with lower posting ratios. Figure 11 shows the average posting ratios per day plotted against the lifespans of the suspended accounts. Many promoters' profiles shared 30 tweets a day on average, and lived almost 30 days, as Figure 11 shows. For spammers' accounts, they were more likely, regardless of their posting ratio, to be identified within less than ten days. That situation, however, does not extend to all campaign accounts, which might be a result of other factors contributing to the suspension. For instance, accounts might be suspended after being reported or flagged for containing disturbing or spamming content by genuine users.

Similarly, the activities of accounts in the suspended set are examined and compared them to those of the active accounts of the two campaigns. The purpose of this analysis is to understand which account behaviors were the most detectable, e.g., retweets, mentions, and centering accounts (see Section IV-A). However, we could not adequately compare all the behaviors due to the small number of second-generation accounts and the different samples in each group (see Figure 12).

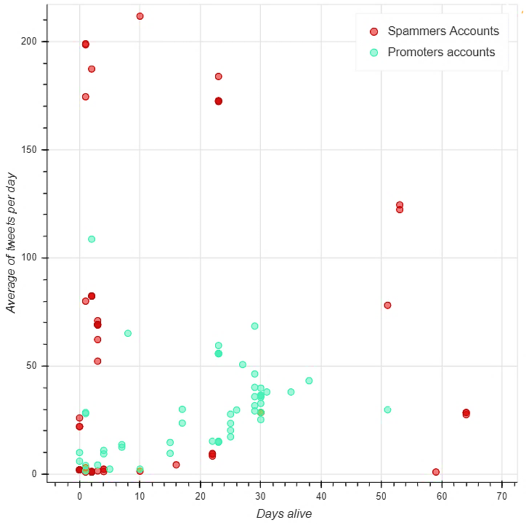


Fig. 11: Duration of suspended accounts' activity and the daily average of tweets

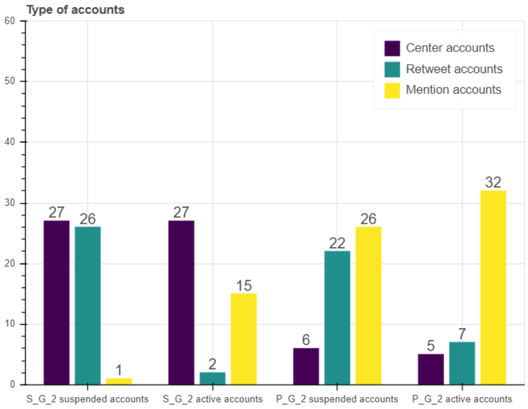


Fig. 12: Type of accounts that were suspended or still active for the two campaigns

Comparing our analysis results with previous studies on Arabic-language spammers [10], [11], we first found that the spam accounts' lifespans in our dataset were much longer than reported in other studies. In [11], for example, it was found that 50% of accounts were detected and suspended after their first spam tweet. In our case, only one account in the promoters' group was detected after three tweets, while the rest of the suspended accounts shared 6 to 1,780 tweets before suspension (see Figure 13 (b)). For the spammer groups, only 2% of total accounts were detected after five or fewer tweets, while the rest shared through their spamming period from 6 to 6,101 tweets (see Figure 13 (a)).

C. Practice of Managing Spam Accounts

Malicious campaigns are commonly known to use software to efficiently and quickly manage an enormous number of fake accounts [14], [40], [41]. Such accounts are known as botnets,

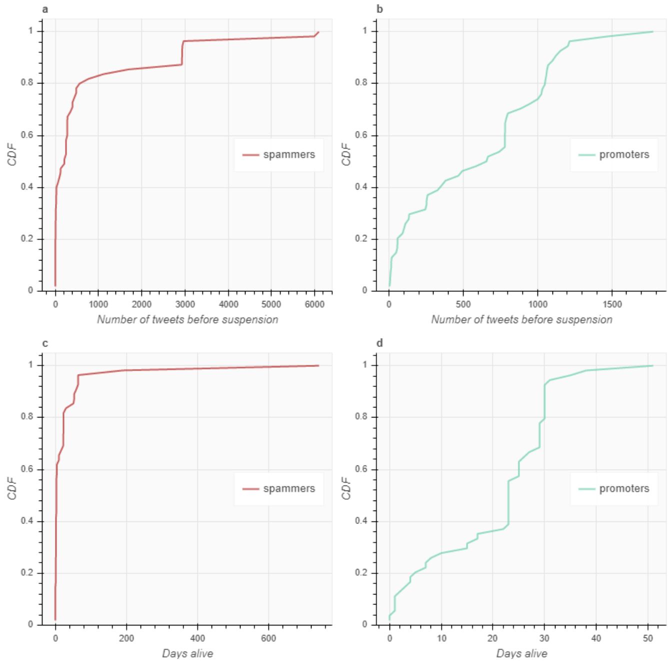


Fig. 13: Lifespans of suspended accounts in days and total numbers of tweets before the suspension

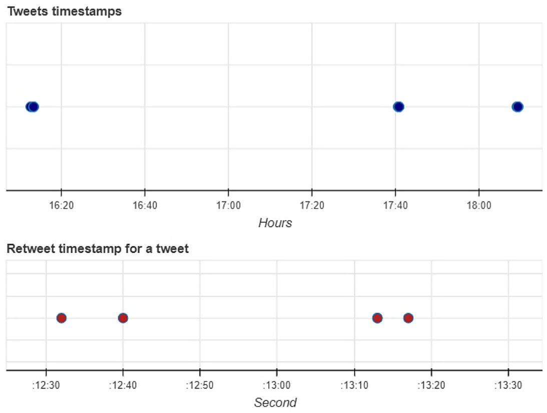


Fig. 14: Timestamps of original tweets and retweets

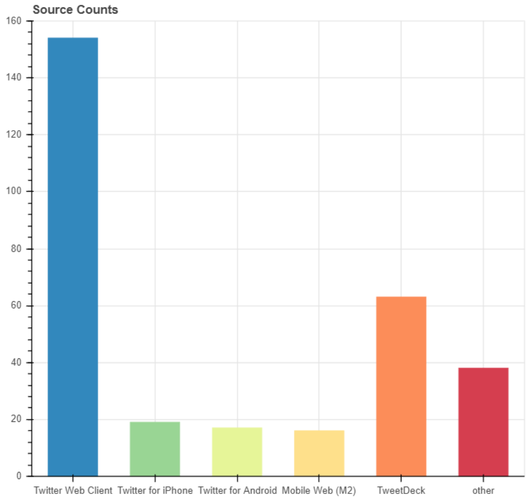


Fig. 15: Sources most used by spammer accounts

which are fake accounts on social media that are fully or partially controlled by software. A synchronized or identical timestamp is an essential feature that defines such users. In this section, therefore, the timing of activities of the campaigns' accounts are examined to identify botnet-like behavior.

In the case of our dataset, different spamming strategies involve various social interactions such as tweeting, retweeting, mentioning, and favoriting. Accounts that share a high volume of tweets in a short time are more likely to be part of a botnet [36], [42], [43], [44]. Among the two groups, the spammers' accounts tended to post a larger number of tweets daily (see Table III for average and maximum numbers of daily tweets). Secondly, synchronized retweets from a set of accounts are a strong indicator of software or botnet accounts. According to our previous experiment regarding several clusters' timestamps from the two campaigns [12], we found that the accounts have synchronized retweet timestamps. Figure 14 shows an example of a central account's activities over a couple of hours and the retweet timestamps. The account received all the retweets in seconds after posting the original tweets, as is shown in the figure. Thereby, we concluded that these accounts are simultaneously controlled to automatically perform specific tasks such as retweeting, replying, and favoriting. However, we found instances of accounts belonging to the promoters' group that exhibited human-like behavior in which they engaged in meaningful conversation with genuine accounts, for example, answering questions.

The source of an account's tweets might also indicate a sign of possible automated control, wherein some of these sources facilitate this process more than others [45]. We identified two major sources used by the two campaigns: Twitter Web Client and TweetDeck, as illustrated in Figures 1516. These two sources are generally known for services that involve scheduling future activities and managing multiple accounts. Also, we found that Twitter for iPhone and Twitter for Android were the sources most used by genuine accounts. Additionally, about 40% of the spammer accounts and 38% of the promoter accounts used two sources in their tweets, whereas less than 18% of genuine accounts used two sources. This suggests that some of these accounts are controlled by both humans and software, which results in human- and botnet-like behavior at the same time.

V. Summary of Arabic Spam Campaigns' Characteristics

The following points summarize the findings of our in-depth study:

- The spammers' and promoters' campaigns mainly involve accounts that are compromised or stolen from legitimate owners, and their lifespans are often greater than $1,500$ days, or four years.

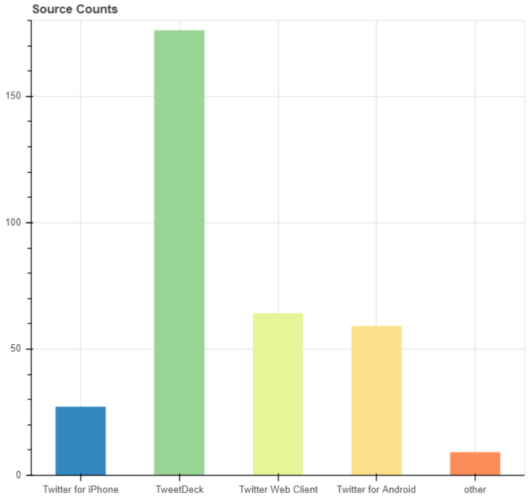


Fig. 16: Sources most used by promoter accounts.

- Both campaigns targeted trending topics through coordinated account groups, and they attempted to manipulate the reputations of their tweets in order to reach to the "top tweets" tab.
- Both campaigns had high self-similarity ratios, with 40% of accounts having a 0.7 similarity ratio, and in general, their content was significantly different from real users' content.
- The spammers' campaign exhibited very easily detectable behavior, as a result of which 60% of accounts were suspended in less than five days. The promoters' accounts, in contrast, posted a fair number of tweets daily, and consequently, their activities (including tweets that were still public in the trends) persisted for more extended periods than those of the spammers.
- We found that the campaigns' accounts were more successful in avoiding Twitter suspension than previously reported in the literature.
- These accounts are partially controlled by human and software, which results in human- and botnet-like behavior at the same time.

VI. Conclusion

This paper presents an in-depth analysis of two malicious Arabic-language campaigns on Twitter. They were selected due to illegal practices that were longer-running and more frequent compared to other groups. The primary purpose of this analysis is to examine the content and behavior of malicious Arabic-language groups, including their respective numbers of accounts, spamming tactics, lifespans, and methods used to control these accounts. To this end, we examined six hundred profiles of the two campaigns that were divided into two generations; the first generation's activity took place on Apr-Dec 2018, and the second-generation's activity on Aug-Sep 2019. Through this study, we have shown that compromised accounts that were usually over $1,500$ days or four years old were the accounts most used by the two campaigns. Both campaigns focused on trends through organized account groups, through which they tried to manipulate their tweets' reputations to reach the top tweets tab; this spamming tactic is clearly shown in their interaction graphs. Secondly, they had straightforward detectable content and their profiles had high ratios of text similarity, with 40% of the accounts having similarity ratios of over 0.7, and in addition, most of them used the same word sets to deliver their messages. Furthermore, we have demonstrated through our 2-month experiment on second-generation accounts that these accounts have avoided Twitter suspension more effectively than has been previously reported in the literature. Among the spammer and promoter campaigns, the promoter campaign was more successful in leveraging their accounts, and their profiles could avoid suspension for a longer period than the spammers' accounts. Finally, they are more likely to use script or software for managing and automating some of their actions, in particular, retweeting and responding.

The analysis provided in this paper has essentially revolved around two malicious Arabic-language campaigns on Twitter. Although most other malicious campaigns either disseminate spam URLs or promote a certain product or service, some of these groups are worth investigating in future research. For example, we found through this study that many groups of fake accounts that offer a service are explicitly manipulating trending topics. This is an interesting area for future work: first to study the trending topics and identify low-quality trending hashtags or topics, and second to investigate the techniques used by these accounts to manipulate topics.

References

[1] N. Eltantawy and J. B. Wiest, "The arab spring— social media in the egyptian revolution: reconsidering resource mobilization theory," *International journal of communication*, vol. 5, p. 18, 2011.

[2] P. N. Howard, A. Duffy, D. Freelon, M. M. Hussain, W. Mari, and M. Maziad, "Opening closed regimes: what was the role of social media during the arab spring?" *Available at SSRN 2595096*, 2011.

[3] A. Bruns, T. Highfield, and J. Burgess, "The arab spring and social media audiences: English and arabic twitter users and their networks," *American behavioral scientist*, vol. 57, no. 7, pp. 871–898, 2013.

[4] M. N. Al-Kabi, A. H. Gigieh, I. M. Alsmadi, H. A. Wahsheh, and M. M. Haidar, "Opinion mining and analysis for arabic language," *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 5, no. 5, pp. 181–195, 2014.

[5] N. Alsaedi, P. Burnap, and O. Rana, "Can we predict a riot? disruptive event detection using twitter," *ACM Transactions on Internet Technology (TOIT)*, vol. 17, no. 2, pp. 1–26, 2017.

[6] H. Almerekhi, M. Hasanain, and T. Elsayed, "Evetar: A new test collection for event detection in arabic tweets," in *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 2016, pp. 689–692.

[7] S. M. Alzanin and A. M. Azmi, "Rumor detection in arabic tweets using semi-supervised and unsupervised expectation–maximization," *Knowledge-Based Systems*, vol. 185, p. 104945, 2019.

[8] G. Wolfsfeld, E. Segev, and T. Sheafer, "Social media and the arab spring: Politics comes first," *The International Journal of Press/Politics*, vol. 18, no. 2, pp. 115–137, 2013.

[9] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: Dark of the social networks," *Journal of Network and Computer Applications*, vol. 79, pp. 41–67, 2017.

[10] H. S. Al-Khalifa, "On the analysis of twitter spam accounts in saudi arabia," *International Journal of Technology Diffusion (IJTD)*, vol. 6, no. 1, pp. 46–60, 2015.

[11] M. Alfifi and J. Caverlee, "Badly evolved? exploring long-surviving suspicious users on twitter," in *International Conference on Social Informatics*. Springer, 2017, pp. 218–233.

[12] R. Alharthi, A. Alhothali, and K. Moria, "Detecting and characterizing arab spammers campaigns in twitter," *Procedia Computer Science*, vol. 163, pp. 248–256, 2019.

[13] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 963–972.

[14] A. Beutel, W. Xu, V. Guruswami, C. Palow, and C. Faloutsos, "Copycatch: stopping group attacks by spotting lockstep behavior in social networks," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 119–130.

[15] M. Giatsoglou, D. Chatzakou, N. Shah, A. Beutel, C. Faloutsos, and A. Vakali, "Nd-sync: Detecting synchronized fraud activities," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015, pp. 201–214.

[16] A. Duh, M. Slak Rupnik, and D. Korošak, "Collective behavior of social bots is encoded in their temporal twitter activity," *Big data*, vol. 6, no. 2, pp. 113–123, 2018.

[17] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake twitter followers," *Decision Support Systems*, vol. 80, pp. 56–71, 2015.

[18] N. Vo, K. Lee, C. Cao, T. Tran, and H. Choi, "Revealing and detecting malicious retweeter groups," in *2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2017, pp. 363–368.

[19] H. AlMahmoud and S. AlKhalifa, "Tsim: a system for discovering similar users on twitter," *Journal of Big Data*, vol. 5, no. 1, p. 39, 2018.

[20] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter," in *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 71–80.

[21] 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 *Proceedings of the 21st international conference on World Wide Web*, 2012, pp. 61–70.

[22] S. Gupta, D. Kuchhal, P. Gupta, M. Ahamad, M. Gupta, and P. Kumaraguru, "Under the shadow of sunshine: Characterizing spam campaigns abusing phone numbers across online social networks," in *Proceedings of the 10th ACM Conference on Web Science*, 2018, pp. 67–76.

[23] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," *Computers & Security*, vol. 76, pp. 265–284, 2018.

[24] N. El-Mawass and S. Alaboodi, "Detecting arabic spammers and content polluters on twitter," in *2016 Sixth International Conference on Digital Information Processing and Communications (ICDIPC)*. IEEE, 2016, pp. 53–58.

[25] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers on twitter," in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*, vol. 6, no. 2010, 2010, p. 12.

[26] N. Al Twairesh, M. Al Tuwaijri, A. Al Moammar, and S. Al Humoud, "Arabic spam detection in twitter," in *The 2nd Workshop on Arabic Corpora and Processing Tools 2016 Theme: Social Media*, 2016, p. 38.

[27] H. Almerekhi and T. Elsayed, "Detecting automatically-generated arabic tweets," in *AIRS*. Springer, 2015, pp. 123–134.

[28] Z. Chen and D. Subramanian, "An unsupervised approach to detect spam campaigns that use botnets on twitter," *arXiv preprint arXiv:1804.05232*, 2018.

[29] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao, "Detecting and characterizing social spam campaigns," in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*, 2010, pp. 35–47.

[30] Z. Alom, B. Carminati, and E. Ferrari, "Detecting spam accounts on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 1191–1198.

[31] A. Almaatouq, E. Shmueli, M. Nouh, A. Alabdulkareem, V. K. Singh, M. Alsaleh, A. Alarifi, A. Alfaris *et al.*, "If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts," *International Journal of Information Security*, vol. 15, no. 5, pp. 475–491, 2016.

[32] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Compa: Detecting compromised accounts on social networks." in *NDSS*, 2013.

[33] X. Ruan, Z. Wu, H. Wang, and S. Jajodia, "Profiling online social behaviors for compromised account detection," *IEEE transactions on information forensics and security*, vol. 11, no. 1, pp. 176–187, 2015.

[34] P. Bindu, R. Mishra, and P. S. Thilagam, "Discovering spammer communities in twitter," *Journal of Intelligent Information Systems*, vol. 51, no. 3, pp. 503–527, 2018.

[35] A. A. Amleshwaram, A. N. Reddy, S. Yadav, G. Gu, and C. Yang, "Cats: Characterizing automation of twitter spammers." in *COMSNETS*, 2013, pp. 1–10.

[36] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization. corr abs/1703.03107 (2017)," *arXiv preprint arXiv:1703.03107*, vol. 3, 2017.

[37] E. M. Clark, J. R. Williams, C. A. Jones, R. A. Galbraith, C. M. Danforth, and P. S. Dodds, "Sifting robotic from organic text: a natural language approach for detecting automation on twitter," *Journal of computational science*, vol. 16, pp. 1–7, 2016.

[38] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "Aravec: A set of arabic word embedding models for use in arabic nlp," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017.

[39] R. Alharthi, A. Alhothali, and K. Moria, "A real-time deep-learning approach for filtering arabic low-quality content and accounts on twitter," *Information Systems*, vol. 99, p. 101740, 2021.

[40] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[41] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Social fingerprinting: detection of spambot groups through dna-inspired behavioral modeling," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 561–576, 2017.

[42] F. Morstatter, L. Wu, T. H. Nazer, K. M. Carley, and H. Liu, "A new approach to bot detection: striking the balance between precision and recall," in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016, pp. 533–540.

[43] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots," in *Proceedings of the 25th international conference companion on world wide web*, 2016, pp. 273–274.

[44] A. H. Wang, "Detecting spam bots in online social networking sites: a machine learning approach," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2010, pp. 335–342.

[45] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving twitter spammers," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 8, pp. 1280–1293, 2013.