Towards Understanding Malicious Actions on Twitter

Completed Research Full Paper

Agnieszka Onuchowska
University of South Florida
aonuchowska@mail.usf.edu

Donald J. Berndt
University of South Florida
dberndt@usf.edu

Abstract

In this study we investigate the characteristics of malicious account behaviors on Twitter based on the analysis of the published data archive. We investigate emergent behavior of malicious accounts that Twitter tagged as connected to state-backed information operations, identified as malicious and removed from the Twitter network. We focus on the analysis of four types of malicious accounts' features: (1) Account reputation, (2) Account tweeting frequency, (3) Age of account and (4) Account activity score. With the use of descriptive statistics and unsupervised learning, we attempt to extend past research that defined behavioral patterns of malicious actors on Twitter. Our research contributes to the understanding of behavior of malicious actors and enriches current research in that area. In this paper we analyze the dataset published by Twitter in January 2019, which contains details on suspended malicious accounts' activities initiated in Bangladesh.

Keywords

Twitter, Malicious Accounts, Tweet Categorization, Bangladesh

Introduction

Over the past years social media has become an important medium for public life (Bradshaw and Howard, 2017), being a global platform that allows people to express their views (Kelly et al., 2012). At the same time, the emergence of activities instigated by malicious accounts has challenged the foundations of today's information society (Ferrara, 2015). Manipulation initiatives have an overwhelming negative effect on societies (Ferrara, 2015) whereas disinformation campaigns embrace confusion and distrust and are often used to deepen societal divisions based on nationalistic, racial or religious strains (Wardle & Derakhshan, 2017). For example, it has been reported that in the past malicious actors' actions on social media focused on diverting attention from discussed issues (Bradshaw and Howard, 2017), influencing beliefs and values of social media users (Gorwa, 2017), providing fake political support (Ratkiewicz et al., 2011) or manipulating online discussions on presidential elections (Bessi and Ferrara, 2016).

Bradshaw and Howard (2017) indicated that the first organized attempts to manipulate social media were reported in 2010. As of 2017, organizations that attempted to manipulate social media outlets were found in 28 countries. Activities that showcased nudging of public opinion environments were reported both among countries with democratic governments and authoritarian regimes. On top of that, social media campaigns were found to target both local populations and global audience, aiming at foreign countries (Bradshaw and Howard, 2017).

Social media companies have been actively targeting malicious accounts, with a goal to reduce their overall number as well as to minimize the impact of such accounts' adverse actions. For example, before the US midterm elections in November 2018, Facebook and Instagram blocked accounts showing "coordinated inauthentic behavior" that seemed designed to interfere in the elections' outcome (Gleicher, 2018). In November 2017, Twitter (2017) updated the Automation Rules policy, in which the company publicly communicated that automated tweets which seek to manipulate trending topics would be filtered out and any originating accounts would be suspended. Consequently, in October 2018, Twitter created a data archive "to strengthen Twitter against attempted manipulation, including malicious automated accounts

and spam" (Twitter, 2018). A published data archive that was posted on Twitter presents removed malicious accounts and their tweets and other media (such as images or videos) generated by state-backed information operations in Russia, Venezuela, Bangladesh and Iran.

Our primary motivation for this study is to understand the characteristics of malicious account behaviors on Twitter based on the analysis of the published data archive. We investigate emergent behavior of malicious accounts that Twitter tagged as connected to state-backed information operations, identified as malicious and removed from the Twitter network. We focus on the analysis of four types of malicious accounts' features: (1) Account reputation, (2) Account tweeting frequency, (3) Age of account and (4) Account activity score. With the use of descriptive statistics and unsupervised learning, we attempt to extend past research that defined behavioral patterns of malicious actors on Twitter. Our research contributes to the understanding of behavior of malicious actors and enriches current research in that area. In this paper we analyze the dataset published by Twitter in January 2019, which contains details on suspended malicious accounts' activities initiated in Bangladesh. To the best of our knowledge, this article is the first effort to extend research on malicious account behavior based on labeled proprietary data on removed malicious accounts identified and released by Twitter itself.

The paper is structured as follows. First, we present past research characterizing malicious account behavior on Twitter. Second, we review past research, which deployed clustering techniques to investigate characteristics of the Twitter network. Next, we present the methods of our study and discuss the findings. Finally, we elaborate on implications of the study and present our plans for future research.

Literature review

Past research on malicious accounts' characteristics

An extensive portion of past research about Twitter is dedicated to the identification of malicious accounts and their behaviors. Badawy et al. (2018) provided distinction between trolls and bots. Trolls can be defined as malicious accounts whose goal is to manipulate the audience, whereas bots are automated accounts. Both bots' and trolls' goals are to misinform the audience and populate politically biased information (Badawy et al., 2018). Previous research estimated that bot accounts represent 9 -15 % (Varol et al., 2017), 5-8.5% (Twitter, 2014), 15% (Bessi & Ferrara, 2016) of all accounts registered on Twitter. Malicious deployment of social bots emulates fake political support or is commonly used to disseminate rumors or spam (Varol et al., 2017, Chu et al., 2012). Social bots, also referred to as sybil accounts, are known to algorithmically generate news content, interact with other user accounts or to coordinate online activities (Davis et al., 2016, Varol et al., 2017). When social bots are grouped together, they create botnets, which are used in coordinated activities (e.g., bot attacks, advertisement campaigns) which are run by botmasters (Varol et al. 2017). The presence of botnets can be indicated by the increased level of spamming activity as accounts owned by humans would rarely be used to retweet spam tweets (Cook et al., 2014). Abokhodair et al. (2015) described misdirection and smoke screening as two types of tactics bots can potentially employ to point the Twitter audience away from legitimate and truthful information. Botnet campaigns are often targeted to populate content which is unrelated to the discussed topic and by this they try to refocus a reader away from legitimate news. Spam bots randomly follow users and expect that some users will follow bots back (Chu et al., 2012). By this, spam posts with email spam leading to ad-infested intermediary sites (Lokot & Diakopoulos, 2016) or unsolicited commercial information populated by bot accounts get disseminated across the network. Chu et al. (2012) estimated that among all bot accounts, about 60% of such accounts would have fewer followers than followees. Varol et al. (2017) identified two types of bots: simple and sophisticated. Sophisticated bots, which try to build normal-looking social ties with other accounts, are known to retweet humans but at the same time, they are not likely to mention human-generated content. Also, sophisticated bots would not be able to lead meaningful discussion exchanges with humans. In turn, simple bots tend to follow other accounts in a random manner, retweet other simple bots and mention sophisticated bots and interact with other bots which show human-like behavior (Varol et al., 2017). Bessi & Ferrara (2016) found that bots, as opposed to legitimate human account holders, tend to tweet in much larger bursts than humans and they tend to retweet existing content more often than creating their own tweets. What is more, they usually have more followees than followers and their lifespan is usually short (shortly after being created, they usually get suspended or removed) and their user names are randomly generated. In turn, Chu et al. (2012) found that bots can become dormant or hibernated. When bots get

hibernated, they would not be disseminating any tweets. However, during the time when bots are active, the number of tweets generated by bots is higher than the number generated by humans.

Clustering of Twitter accounts

Past research often relied on clustering methods to investigate Twitter actors' characteristics and actions. For example, Vaast et al. (2017) used clustering to check the consistency of key feature patterns such as mentions, likes, retweets and hashtags, which were used to define connective action episodes (CAEs). Gunarathne et al. (2018) analyzed customer complaints shared through Twitter and deployed the K-means algorithm to group tweets which share similar complaint types into clusters. Soman and Murugappan (2014) used the fuzzy K-means (FKM) algorithm to cluster similar user profiles, which shared similar trending topics in their tweets. By this, they proposed a method that distinguishes tweet and non-tweet spam users within similar trending topic areas. Wei et al. (2015) used cluster analysis to distinguish between suspended and non-suspended accounts. Next, they performed Gaussian Mixture Modelling to find suspended accounts' subtypes, such as spammers or bots. Becker et al. (2011) used the online incremental clustering algorithm to group tweets on similar topics. Then they used defined cluster to conduct event classification on Twitter. Kaleel and Abhari (2015) used K-means and LHS techniques to cluster tweets, label the clusters and identify Twitter events based on the defined clusters. Takhteyev et al. (2012) clustered tweets based on the geographical location of user accounts' profiles to find that Twitter users tend to establish ties with other users from the same region or metropolitan area. Vosecky et al. (2014) proposed a Multi-faceted Topic Model (MfTM) to evaluate the quality of clustering methods applied on Twitter datasets. Alsaedi et al. (2017) clustered individual tweets according to an investigated event on Twitter using an online clustering algorithm. With the clustering output, the researchers were able to summarize topics discussed in each of the identified clusters. Bakerman et al. (2018) used K-means clustering to analyze tweets' textual information combined with the geotagged coordinates information (latitude and longitude) that characterized each tweet.

Method and Testbed

Methods of Data Analysis

To understand the characteristics of a given dataset, we first conducted descriptive statistical analysis of malicious account behaviors presented in the dataset. We identified the characteristics of malicious accounts and referred to past literature, which showcased that one can draw inferences about Twitter accounts' characteristics based on user metadata without the need to run sentiment analysis (Kelly et al., 2012, Lee et al., 2011). In order to find common patterns characterizing malicious accounts, we conducted K-means cluster analysis, which historically serves as a classic way of finding emerging patterns in raw data (Lohrmann & Luuka, 2018, van Dam & van de Velden, 2015). We focused on four types of characteristics: (1) Account reputation, (2) Account tweeting frequency, (3) Age of account and (4) Account activity score.

Baseline Dataset of Malicious Accounts

In October 2018, Twitter started publishing archives of tweets and media which had been identified as malicious and removed by Twitter from the network. According to Twitter, identified accounts had been proliferating potentially state-backed information operations instigated by Russia, Iran, Venezuela and Bangladesh (Twitter, 2018). The data archive presented on the Twitter website contains the information about malicious tweets followed by other media content, such as videos or pictures.

In this study we decided to analyze the dataset related to Bangladesh's state-backed information operations, which we then analyzed in line with the earlier literature review on malicious account types. The dataset was published by Twitter in January 2019 and contains the following information about the attributes of 11 malicious accounts (Twitter, 2018):

Numerical attributes: tweet identification number, user identification number, the number of accounts following the user, the number of accounts followed by the user, date of user account creation, the time when the tweet was published, the tweetid of the original tweet that this tweet is in reply to, the number of tweets quoting this tweet, the number of tweets replying to this tweet, the number of likes that this tweet received, the number of retweets that this tweet received

Categorical attributes: the name of the user (hashed), the user's self-reported location, the user's profile description, the user's profile URL, the language of the account, as chosen by the user, the language of the tweet, the text of the tweet, the name of the client app used to publish the tweet, True/False, is this tweet a retweet, geo-located latitude and longitude, a list of hashtags used in this tweet, a list of URLs used in this tweet.

The primary data which we decided to use for our analysis involved the following attributes: the number of accounts following the user, the number of accounts followed by the user, date of user account creation, the time when the tweet was published, the number of likes that this tweet received, the number of retweets that this tweet received.

In order to investigate tweeting frequency, we analyzed 11 malicious account behaviors presented in the Bangladesh dataset. Table 1 describes malicious accounts presented in the dataset:

Account #	Account creation	Reported location	Profile description	Follower count	Following count	Number of tweets	Number of retweets
Account 1	12/4/2016	Dhaka, Bangladesh	Only news portal provides news in Bangla, English and Arabic	17	23	956	0
Account 2	11/14/2017	Dhaka, Bangladesh	https://t.co/wtfVbjri8C is a Bangladesh based multimedia platform for news, opinion and entertainment. Itâ€™s a 24/7/365 display to keep readers updated.	156	8	13317	0
Account 3	12/7/2017	Bangladesh	N/A	4	36	8	0
Account 4	8/28/2016	Dhaka, Bangladesh	News service in Bangla and English	17	25	2877	0
Account 5	4/22/2012	N/A	My FB: https://t.co/8oUmDbaS36 Activist, Bangladesh Student League	1626	822	1095	820
Account 6	8/11/2018	Narayanganj, Bangladesh	Juger Chinta	4	7	122	0
Account 7	5/20/2009	Dhaka	#SoftwareEngineer, #Photographer, #Activist, #Traveler	2101	462	2097	37
Account 8	4/12/2018	Dhaka, Bangladesh	N/A	0	6	8	0
Account 9	1/22/2014	Dhaka, Bangladesh	You can either love me or hate me but either way I am still on your mind. ðŸ˜Ž	1789	90	4027	158
Account 10	12/14/2017	Sylhet, Bangladesh	https://t.co/H8kHPOlQCX is an online news portal from Bangladesh.	6	90	14	0
Account 11	5/8/2017	Bangladesh	FB: https://t.co/cmBuPz7Whp	760	11	671	7

Table 1: Description of 11 Malicious Accounts

For each of the 11 accounts, we extracted information about their respective tweeting history such as tweeting frequency, the number of tweets generated each day, number of days when accounts stayed dormant and number of days between account creation and account removal. From this information, we

drew characteristics of malicious accounts' tweeting behavior and we focused on the measurement of the following features: (1) Account reputation, (2) Account tweeting frequency, (3) Age of account and (4) Account activity score. The description of identified account features is presented in Table 2.

Feature	**Description**
Account reputation	Follower count/ (follower count + following count) (Chu et al., 2012)
Tweet frequency	The average number of tweets generated by an account on a daily basis (Dickerson et al., 2014)
Days tweeted	Number of days when an account tweeted at least one time
Age of account	Total number of days between account creation and removal (Freitas et al., 2015)
Activity score (%)	Days tweeting/ days active (%)
Total number of tweets	Total number of tweets posted during account's lifetime (Freitas et al., 2015)
Total like count	Total number of likes all tweets received (Cristofaro et al., 2014)
Total retweet count	Total number of retweets all tweets received (Cha et al., 2010)
Average like count	Total like count/ total number of tweets
Average retweet count	Total retweet count/ total number of tweets

Table 2: Description of account features

The values characterizing each of the features can be found in Table 3. The description of each feature is as follows: An account reputation score which is close to 1 would suggest a celebrity account (such accounts have many followers and few friends). A score of less than 0.5 indicates that an account is likely to be a bot (bots have fewer followers than friends) whereas accounts with a reputation score that oscillates around 0.5 suggests an account owned by a simple human (Chu et al., 2012). Tweet frequency is defined as an average number of tweets generated by an account on a daily basis. Dickerson et al. (2012) indicates that a high tweet frequency score could indicate bot behavior. In turn, a comparison of the number of days tweeted and the age of account computed as activity score (%) should help us define the percentage of time in which an account was actively tweeting. Low activity score could indicate dormant or hibernated malicious accounts (Chu et al., 2012).

	Account reputation	Tweet frequency	Days tweeted	Age of account	Activity score (%)	Total number of tweets	Total like count	Total retweet count	Avg like count	Avg retweet count
Account 1	0.43	5.40	177	306	57.84%	956	14	1	0.01	0.00
Account 2	0.95	42.68	312	402	77.61%	13317	667	19	0.05	0.00
Account 3	0.00	2.67	3	3	100.00 %	8	0	0	0.00	0.00
Account 4	0.40	10.70	269	845	31.83%	2877	15	2	0.01	0.00
Account 5	0.66	6.02	318	2426	13.11%	1915	792	90	0.41	0.05
Account 6	0.36	3.94	31	113	27.43%	122	8	0	0.07	0.00
Account 7	0.82	3.33	641	3393	18.89%	2134	602	321	0.28	0.15
Account 8	0.11	4.00	2	5	40.00%	8	0	0	0.00	0.00
Account 9	0.95	3.16	1324	1793	73.84%	4185	2750	754	0.66	0.18
Account 10	0.06	7.00	2	13	15.38%	14	1	0	0.07	0.00
Account 11	0.99	2.25	302	592	51.01%	678	2738	301	4.04	0.44

Table 3: Features of Malicious Accounts

The descriptive statistics for the analyzed dataset are presented in Table 4.

	Account reputation	Tweet frequency	Days tweeted	Age of account	Activity score (%)	Total number of tweets	Total like count	Total retweet count	Avg like count	Avg Retweet count
Mean	0.52	8.29	307.36	899.18	46.09%	2383.09	689.73	135.27	0.51	0.07
Standard deviation	0.37	11.66	389.74	1142.44	28.62%	3876.51	1061.19	238.61	1.19	0.14
Median	0.43	4.00	269.00	402.00	40.00%	956.00	15.00	2.00	0.07	0.00

Table 4: Descriptive Statistics of Identified Features

K-means Clustering

In order to define the number of analyzed clusters, we decided to compare internal consistency measures of the provided data such as pseudo-R2 and total sum of squared errors (SSE). SSE refers to the sum of squared differences between each observation and its' group mean and is used as a measure of variation within the cluster. To find the suitable number of clusters, we used the elbow method to find a low enough value of 'k' with a low enough SSE. On top of the SSE measure, we decided to analyze R-squared, being an internal consistency measure which defines how much of the variation in data is being captured by the clustering methodology. Similar to SSE, we used the elbow method to find the suitable number of 'k' which yielded a high enough value of R-square. Based on SSE and R2 measures' outputs, we observed that using four clusters was most suitable for our study.

#of clusters	SS 1	SS 2	SS 3	SS 4	SS 5	SS 6	Total SSE(*)	R square
2	684757.3	1298598.1					1983355.40	84.8 %
3	200348.8	0.0	684757.3				885106.10	93.2%
4	67899.50	99695.89	200348.8	0.0			367944.19	97.2%
5	8493.6	0.0	5303.0	32040.4	200348.8	0.0	246185.80	98.1%
6	0.0	0.0	8493.6	5303.0	0.0	32040.4	45837.00	99.6%

Table 5: SSE and R2 of Clusters

# of clusters/ Cluster #	**1**	**2**	**3**	**4**	**5**	**6**
2	8	3				
3	2	1	8			
4	1	2	5	3		
5	4	1	2	2	2	
6	1	1	4	2	1	2

Table 6: Cluster Membership

Results

Once the suitable number of clusters was set to four, we performed a cluster analysis based on the following account characteristics: account reputation, tweet frequency, age of account and activity score. The four clusters that we achieved from the data are demonstrated below:

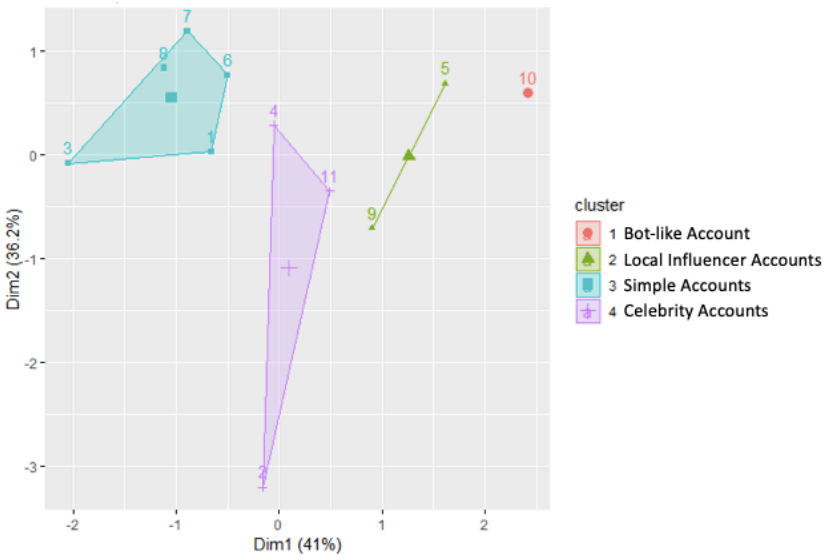

Figure 1: Demonstration of clusters on a two-dimensional plane

Based on the outcomes of our computations presented in Table 5 and Table 6, we obtained four clusters, which are defined as follows: Cluster 1 contained one account and Cluster 2 comprised two accounts, whereas Cluster 3 had five and Cluster 4 had three accounts. Looking at the attributes used for the cluster analysis, we observed that the average account reputation of the whole group is close to 0.5, showcasing that the total number of followers was equal to total number of followees when combining all eleven accounts. On average the tweeting frequency score oscillated around 8.28 which demonstrated that on average an account was tweeting more than 8 times a day. The age of accounts was also quite varied. Cluster 1 represented a group of accounts which on average were 13 days old. On the contrary, Cluster 2 presented a group of accounts with an average age of nearly 3000 days. The activity score was also quite varied, ranging from accounts tweeting very rarely to accounts tweeting on a regular basis. We describe the clusters achieved in the section below:

	Account Reputation	Tweet Frequency	Age of Account	Activity Score
Full data	0.5209	8.2864	899.18	0.4609
Cluster 1	0.06	7	13	0.1538
Cluster 2	0.74	4.675	2909.5	0.16
Cluster 3	0.26	5.342	254.4	0.5142
Cluster 4	0.9633	16.03	929	0.6749

Table 7: Clusters of Identified Features

Cluster 1 (Bot-like Account) had only one member and this account had a very low reputation (0.06) and age (13 days). Another interesting aspect is even though it was active for a very small fraction of its' age (activity score of 0.1538), the accounts' tweeting frequency of 7 was quite close to the average for the whole data (8.29). Therefore, we could possibly categorize this account as a representative of those bot accounts which either tweet a lot or do not tweet at all.

Cluster 2 (Local Influencer Accounts) had two member accounts and these accounts enjoyed a much higher than average reputation of 0.74. The age of these accounts was considerably very high (2909.5 days). Their activity score (0.1538) also demonstrated that they tweeted quite intermittently. Their tweeting frequency (4.675) was also much lower than the average (8.29). These accounts represented those who do not tweet very frequently but still enjoy quite a high number of followers and behave like local influencers.

Cluster 3 (Simple Accounts) was the biggest cluster with 5 members. The account reputation (0.26) was much lower than the average (0.5209). Their tweeting frequency (5.3) and age (254.4) was also lower than the average for the group. Their activity score of 0.51 denoted that they were quite active. These represented simple accounts which are following a lot of other accounts and are not very influential themselves but help in propagation.

Cluster 4 (Celebrity Accounts) was the last cluster and it had three members. These accounts enjoyed a very strong reputation (0.9633) signifying that they had a much larger number of followers than the number they were following themselves. They also tweeted very frequently (16.03). The accounts also were quite old (on average 929 days) and were very actively tweeting as well (activity score of 0.67). Consequently, these we represented as celebrity accounts.

Conclusions

The contribution of this paper is as follows. We were able to extract four different measures of malicious account behavior from a dataset released by Twitter. The features extracted captured an account's longevity, tweeting behavior and reputation. Due to the high heterogeneity in malicious accounts' behavior and description, we employed a cluster analysis to group these accounts according to these characteristics. We found out that even though there were only 11 accounts in the dataset, they showcased a wide variety of user behavior. We identified very young accounts which were highly volatile in their tweeting behavior and we were also able to capture significantly older accounts which were more consistent in their tweeting activity.

We identified four clusters of malicious accounts, which showed different tactics of propagating misinformation on a Twitter platform. Cluster 1 represented by one account, showed a bot-like behavior due to the fact that the account reputation was low (0.06). The account's lifespan was short – the account was taken out from the Twitter network only after 14 days, which can also suggest that Cluster 1 representative was a bot account. In turn, Cluster 2 representatives' behavior reflected the way local influencer accounts usually act. Cluster 2 accounts did not tweet frequently but still enjoyed a high number of followers. Next, Cluster 3 accounts resembled a set of behaviors specific to simple accounts. The accounts had a large number of followees but were not followed by many accounts. Finally, Cluster 4 accounts showed a celebrity-like behavior. They had a strong reputation (0.9633) indicating higher number of followers than followees. Cluster 4 accounts had high activity score which implicates that in order to spread misinformation they tweeted on a regular basis to propagate misinformation.

Identified clusters have enabled us to get a better understanding of malicious account behavior in populating misinformation . One of the major shortcomings of this paper is related to the fact that there are only 11 accounts in the data which we analyzed. Nevertheless, in the near future we are looking to analyze datasets related to Russian, Iranian and Venezuelan state-backed operations, which will help us get a more holistic understanding of malicious account behavior in larger networks.

Future Research

As a next step, we plan to extend the study with the use of the datasets on malicious accounts published by Twitter. Based on the findings, we intend to build an agent-based model simulation. With the use of the simulation, we aim to undertake inductive extraction of malicious accounts data published by Twitter. We understand that constructing an agent-based model of the Twitter environment will help us to draw inferences on the direction of future research and guide us towards future uses of ABM simulations of Twitter and data collection of the same.

References

Abokhodair, N., Yoo, D., & McDonald, D. W. 2015. "Dissecting a Social Botnet: Growth, Content and Influence in Twitter," In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 839-851). ACM.

Alsaedi, N., Burnap, P., & Rana, O. 2017. "Can We Predict a Riot? Disruptive Event Detection Using Twitter," *ACM Transactions on Internet Technology (TOIT)*, *17*(2), 18.

Bakerman, J., Pazdernik, K., Wilson, A., Fairchild, G., & Bahran, R. 2018. "Twitter Geolocation: A Hybrid Approach," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *12*(3), 34.

Badawy, A., Ferrara, E., & Lerman, K. 2018. "Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign," In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 258-265). IEEE.

Becker, H., Naaman, M., & Gravano, L. 2011. "Beyond Trending Topics: Real-world Event Identification on Twitter," In *Fifth International AAAI Conference on Weblogs and Social Media*.

Bessi, A., & Ferrara, E. 2016. "Social Bots Distort the 2016 US Presidential Election Online Discussion".

Bradshaw, S., & Howard, P. (2017). Troops, trolls and troublemakers: A global inventory of organized social media manipulation.

Cook, D. M., Waugh, B., Abdipanah, M., Hashemi, O., & Rahman, S. A. 2014. "Twitter Deception and Influence: Issues of Identity, Slacktivism, and Puppetry," *Journal of Information Warfare*, 13(1), 58-71.

Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, K. P. 2010. "Measuring User Influence in Twitter: The Million Follower Fallacy," In *Fourth International AAAI Conference on Weblogs and Social Media*.

Chu, Z., Gianvecchio, S., Wang, H., & Jajodia, S. 2012. "Detecting Automation of Twitter Accounts: Are you a Human, Bot, or Cyborg?" *IEEE Transactions on Dependable and Secure Computing*, 9(6), 811-824.

De Cristofaro, E., Friedman, A., Jourjon, G., Kaafar, M. A., & Shafiq, M. Z. 2014. "Paying for likes?: Understanding facebook like fraud using honeypots," In *Proceedings of the 2014 Conference on Internet Measurement Conference* (pp. 129-136). ACM.

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. 2016. "Botornot: A System to Evaluate Social Bots," In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 273-274). International World Wide Web Conferences Steering Committee.

Dickerson, J. P., Kagan, V., & Subrahmanian, V. S. 2014. "Using Sentiment to Detect Bots on Twitter: Are Humans More Opinionated Than Bots?" In *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 620-627). IEEE Press.

Ferrara, E. 2015. "Manipulation and Abuse on Social Media" by Emilio Ferrara with Ching-Man Au Yeung as coordinator. *ACM SIGWEB Newsletter*, (Spring), 4.

Freitas, C., Benevenuto, F., Ghosh, S., & Veloso, A. 2015. "Reverse Engineering Socialbot Infiltration Strategies in Twitter," In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 25-32). IEEE.

Gleicher, N. 2018. Election Update | Facebook Newsroom. Retrieved from https://newsroom.fb.com/news/2018/11/election-update/

Gorwa, R. 2017. "Computational Propaganda in Poland: False Amplifiers and the Digital Public Sphere," *Computational Propaganda Research Project Working Paper*, (2017.4).

Gunarathne, P., Rui, H., & Seidmann, A. 2018. "When Social Media Delivers Customer Service: Differential Customer Treatment in the Airline Industry," *MIS Quarterly*, *42*(2), 489-520.

Kaleel, S. B., & Abhari, A. 2015. "Cluster-discovery of Twitter Messages for Event Detection and Trending," *Journal of Computational Science*, 6, 47-57.

Kelly, J., Barash, V., Alexanyan, K., Etling, B., Faris, R., Gasser, U., & Palfrey, J. G. 2012. "Mapping Russian Twitter".

Lee, K., Eoff, B. D., & Caverlee, J. 2011. "Seven Months with the Devils: A Long-term Study of Content Polluters on Twitter," In *Fifth International AAAI Conference on Weblogs and Social Media*.

Lohrmann, C., & Luukka, P. 2018. A Novel Similarity Classifier with Multiple Ideal Vectors Based on k-means Clustering," *Decision Support Systems*, 111, 27-37.

Lokot, T., & Diakopoulos, N. 2016. "News Bots: Automating News and Information Dissemination on Twitter," *Digital Journalism*, 4(6), 682-699.

Ratkiewicz, J., Conover, M. D., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. M. 2011. "Detecting and Tracking Political Abuse in Social Media," In *Fifth international AAAI conference on weblogs and social media*.

Soman, S. J., & Murugappan, S. 2014. "Detecting Malicious Tweets in Trending Topics Using Clustering and Classification," In *2014 International Conference on Recent Trends in Information Technology* (pp. 1-6). IEEE.

Takhteyev, Y., Gruzd, A., & Wellman, B. 2012. "Geography of Twitter Networks," *Social Networks*, 34(1), 73-81.

Twitter. 2014. United States Securities and Exchange Commission Report. Retrieved from https://www.sec.gov/Archives/edgar/data/1418091/000156459014003474/twtr-10q_20140630.htm

Twitter. 2017. Automation rules. Retrieved from https://help.twitter.com/en/rules-and-policies/twitter-automation

Twitter. 2018. Dataset Readme. Retrieved from https://storage.googleapis.com/twitter-election-integrity/hashed/Twitter_Elections_Integrity_Datasets_hashed_README.txt

Twitter. 2018. Elections integrity. Retrieved from https://about.twitter.com/en_us/values/elections-integrity.html#data

Vaast, E., Safadi, H., Lapointe, L., & Negoita, B. 2017. "Social Media Affordances for Connective Action: An Examination of Microblogging Use During the Gulf of Mexico Oil Spill," *MIS Quarterly*, *41*(4).

Van Dam, J. W., & Van De Velden, M. 2015. "Online Profiling and Clustering of Facebook Users," *Decision Support Systems*, 70, 60-72.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. 2017. "Online Human-bot Interactions: Detection, Estimation, and Characterization," arXiv preprint arXiv:1703.03107.

Vosecky, J., Jiang, D., Leung, K. W. T., Xing, K., & Ng, W. 2014. "Integrating Social and Auxiliary Semantics for Multifaceted Topic Modeling in Twitter," *ACM Transactions on Internet Technology (TOIT),* 14(4), 27.

Wardle, C., & Derakhshan, H. 2017. "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making," *Council of Europe report, DGI* (2017), 9.

Wei, W., Joseph, K., Liu, H., & Carley, K. M. 2015. "The Fragility of Twitter Social Networks Against Suspended Users," In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 9-16). IEEE.