Detecting Social Bots on Twitter: A Literature Review

Eiman Alothali*[1], Nazar Zaki[2], Elfadil A. Mohamed[3], and Hany Alashwal[4]
[1, 2, 4] College of Information Technology
United Arab Emirates University, AlAin, UAE
Email: {201790016, nzaki, halashwal}@uaeu.ac.ae
[3] College of Information Technology, Ajman University, UAE
Email: elfadil.abdalla@ajman.ac.ae

***Abstract*—Due to the exponential growth in the popularity of online social networks (OSNs), such as Twitter and Facebook, the number of machine accounts that are designed to mimic human users has increased. Social bots accounts (Sybils) have become more sophisticated and deceptive in their efforts to replicate the behaviors of normal accounts. As such, there is a distinct need for the research community to develop technologies that can detect social bots. This paper presents a review of the recent techniques that have emerged that are designed to differentiate between social bot account and human accounts. We limit the analysis to the detection of social bots on the Twitter social media platform. We review the various detection schemes that are currently in use and examine common aspects such as the classifier, datasets, and selected features employed. We also compare the evaluation techniques that are employed to validate the classifiers. Finally, we highlight the challenges that remain in the domain of social bot detection and consider future directions for research efforts that are designed to address this problem.**

Keywords—Social Bots; Twitter; Detection; Sybil

I. Introduction

Online social networks (OSN) represent a global platform through which people share and promote products, links, opinions, and news. In the third quarter of 2007, Twitter had 330 million active users [1]. By 2015, the estimated number of users had grown to 1.3 billion [2]. The data-sharing feature of social networks allows users to distribute content and links; however, this feature is also commonly used by spammers and fraudsters. Social bot accounts make OSNs vulnerable to adversaries. Social bots are programs that automatically generate content, distribute it via a particular social network, and interact with its users [3]. According to a recent study by Varol et al., between 9% and 15% of Twitter accounts are bot accounts [4], which is the equivalent of 48 million accounts [2]. A further study found that social bots are responsible for generating 35% of the content that is posted on Twitter [5].

Many studies have aimed to address the problems associated the use of automated accounts on social networks [6][7], [8], [9], which can spread spam, warms, and phishing links or manipulate legitimate accounts by hijacking and deceiving users [10]-[12]. Malicious accounts typically operate under a botmaster, who controls a group of social bots to distribute spam or manipulate behaviors on a given social network [13]. For example, in Syria, a social bot was employed to flood Twitter with hashtags related to the Syrian civil war with irrelevant topics that redirected the attention of users from controversial government actions [5]. Social bots have also played a significant role in the uprisings that occur in the aftermath of major events such as elections or conflicts [14]. Gupta et al. [15] studied the fake content that was proliferated via Twitter during the Boston Marathon blasts and the role such content played in spreading rumors and misinformation. They found that bot accounts were created and generated after the blasts, many of which impersonated real accounts [15]. The malicious activities of bots during events such as these can be used to spread spam. In addition, they can also cause financial harm, as was observed in the case of Cynk, which suffered a 220-fold drop in market price as a result of the activities of automated stock trading social bots [3].

The activities of social bots also impact the social graph of OSNs because of the large number of non-genuine social relationships. If social bots successfully infiltrate users' accounts, they can harvest social bot private data and subsequently use it for phishing and spamming activities [9] [16]. In addition, they can aggregate information from the web to impersonate others, replicate human behaviors, and influence people by ranking and retweeting. In addition to essentially misleading users, social bots can damage the ecosystem of the social network by establishing fake fellowship relations [17] and/or poisoning the network content.

In an attempt to limit the threats posed by social bots, researchers have proposed different methods by which social bots can be detected and blocked. The majority of the studies in this domain to date have focused on studying behavior patterns [18]-[22]. For example, a recent study that was performed by Fu et al. [23] proposed a dynamic metric to measure the change in users' activities as a means of identifying the strategies employed by spammers. Another detection scheme aimed to identify malicious account groups by understanding the algorithms associated with the generated account names and subsequently relating these to creation time [24]. This study analyzed 4.7 million accounts that were collected from Twitter and achieved reasonable accuracy. In the same area, Stringhini et al. [8] studied data from three large social networks after creating large and diverse honey-profiles. They successfully detected 15,857 spam profiles that had been deleted by Twitter.

It is important to note that not all social bot accounts can be classified as malicious accounts. Some even explicitly state their nature in the profile of the account. Social bots that operate without malicious intent may serve positive purposes, such as managing news feeds or acting as customer care responders. The problem we are concerned with in this paper is undisclosed social bots that have malicious intentions. As outlined above, these social bots can pose fundamental financial, social, political, and security risks. They have become increasingly sophisticated in their designs and capabilities to avoid social bot detection techniques [5], [14]. A study by Freitas et al. [25] found that only 38 out of every 120 social bots were detected and removed by Twitter.

In light of the above, there is a requirement to gain in-depth insights into the capabilities and limitations of the social bot detection techniques that are currently in use on the Twitter social network platform. By comparing and evaluating the existing approaches, we can develop an understanding of the different solutions that are available based on the selected features and trained classifiers, and can subsequently apply this understanding to identify which technologies achieve the best accuracy and detection results.

The rest of this paper is organized as follows. The review methodology is presented in Section II. In Section III, we evaluate the datasets employed in existing studies. Section IV then progresses to identify the methods by which social bots can be detected. Section V presents a discussion and evaluation of the techniques mentioned in the previous section. Section VI highlights the challenges that remain in the domain of social bot detection and considers future directions for research efforts that aim to address this problem.

II. Review Methodology

In this paper, we focus on studying the detection techniques that are commonly employed to detect social bots or fake accounts on the Twitter social network platform. The analysis does not include alternative social networks, such as Facebook or Tumbler, or other malicious activities and problems such as spam or hijacking.

III. Datasets & Preprocessing

The approaches that underpin social bot detection techniques can vary significantly. However, they can broadly be categorized into three common methods: graph-based, crowdsourcing, and machine learning [3], [38]. The process of detecting social bots commences by retrieving data from the Twitter stream. Once the data is collected, the next step involves preparing this data for the chosen classifier by extracting and selecting the features that can be studied through statistical methods, as [26], [27], or those that can be manually labelled using previous work, as [6], [28], [29].

In this section, we review the datasets that were employed in the studies identified in the literature review and determine whether they use public or private datasets. Through the use of a graph-based method, we identify the

978-1-5386-6673-9/18/$31.00 ©2018 IEEE

social graph data employed and the total number of nodes for sampling both classes: Sybil and legitimate. In the machine-learning methods analysis, we identify the tools employed in the data collection process and assess the total number of accounts included in the testing phase. In addition, we state the number of features that were involved in the preprocessing phase.

Within their graph-based approach, [30] and [31] used a public dataset that was compiled using data from previous studies. For example, the authors in [31] used a sample of 100 Sybil nodes and a sample of 100 benign nodes for the synthesized social network in addition to a real dataset from Twitter to compare their proposed bot-detection method with other random walk-based efforts. They employed a Twitter dataset sample of 50,000 nodes for the Sybil region and 50,000 for benign nodes for training and testing after processing a dataset that contained 41,652,230 nodes and 1,202513,046 edges. Moreover, to complete their dataset for experimenting, both [32], [30] purchased a number of fake Twitter accounts to implement within the Sybil social network region.

In studies that have focused on machine learning methods, the reviewed studies used datasets that consisted of a combination of privately obtained accounts and the public datasets that were employed in previous studies (See Table I). In certain cases, some researchers, such as [27] and [35], used the available public datasets as a ground truth baseline for testing their techniques. In general, most of the research employed the Twitter API to collect data and compile the datasets with the exception of [33], who used their own API to collect data [34]. Feature selection methods are commonly applied to increase the speed of the classier, reduce the training time, improve generalization, and avoid the overfitting problem. For example, as part of their preprocessing phase, [27] used a correlation-based system in combination with a principal components analysis method. The selected features were then analyzed using a cumulative distribution function for each selected feature.

IV. Social bot Detection Methods

Generally, social bot detection on social networks is performed by one or more of the three common methods mentioned earlier: Graph-based, crowdsourcing, and machine learning.

The graph-based method involves using the social graph of a social network to understand the network information and the relationships between edges or links across accounts to detect bot activity. The crowdsourcing method involves using expert annotators to identify, evaluate, and determine social bot behaviors. Finally, the machine learning method involves developing algorithms and statistical methods that can develop an understanding of the revealing features or behavior of social network accounts in order to distinguish between human- and computer-led activity.

In this section, we provide an overview of the three methods that have been used by researchers to detect social bot accounts on Twitter. In each subsection, we discuss the related studies and the datasets, detection mechanisms, classifiers involved, and the process by which the results were validated.

A. *Graph-Based Detection*

Social network graphs are commonly employed to understand and distinguish between users' relationships on social networking platforms. Three social graph-based methods are typically employed to detect social bots and malicious accounts [38]. The first method is based on trust propagation, which evaluates whether the trust relationship that exists between two graph objects is strong or weak. The second method is graph clustering, by which related nodes of a social graph are grouped based on similar characteristics such as users' distance. The third method involves studying graph metrics and properties, where probability distribution, scale-free graph structure, and centrality metric measures are addressed in a social graph.

In this subsection, we present three graph-based Sybil detection systems that were evaluated using datasets from Twitter. These three systems were chosen because they involved the analysis of a Twitter dataset. They are presented in chronological order, starting with the most recent work.

SybilWalk [31] is a proposed Sybil detection method that employs a random walk-based method on an undirected social graph. The idea of the random walk method is to label legitimate users with benignness scores and Sybil users with badness scores. Therefore, these scores will help to classify users into two classes: legitimate and Sybil. In addition, this method can rank all users as a means of identifying top-ranked accounts that are likely to be Sybils. The authors of SybilWalk assumed that the graph satisfies the homophily property, for which two linked nodes tend to share the same label. They labeled the legitimate node badness score with 0 and the Sybil node with 1 and employed a directed Twitter graph dataset that was obtained from a previous study [43]. To evaluate their experiment, the authors used the Area Under the Receiver Operating Characteristic Curve (AUC) as a standard metric to measure the quality of their ranking method, which they awarded with a score of 0.96. They also presented the classification results of Sybil and legitimate nodes in the form of the false positive rate (FPR) and false negative rate (FNR) of 1.3% and 17.3% respectively.

Mehrotra et al. proposed a method to detect fake followers using social graph-based features that relate to the centrality of all nodes in the graph [30]. They claimed that their proposed method can be applied to all social networking platforms. They employed five datasets, two of which were of the legitimate follower type and the remaining three of fake followers. They used six features of centralities of graph-based centrality measurements for the purposes of the classification. After computing the centrality measures of all the given nodes in the graph, they applied three classifiers: Artificial Neural Networks, Decision Tree, and Random Forest. The random forest classifiers scored the highest accuracy of 95%, with the precision of 88.99%, and recall of 100%.

TrueTop [32] is another influence measurement system that employed a graph-based approach to test Sybil resilience. The authors employed a synthetic simulation of users on Twitter to avoid violating the platform's terms of service. They employed four datasets to implement the system, evaluate the accuracy of the model, and test Sybil resilience against a set of predetermined metrics. They presented a model of the strength of Sybil attacks based on the α parameter, which represented the ratio of total weight of the edge in the non-Sybil region against that of the Sybil region. They assumed the worst-case scenario of Sybil attacks in which there was no interaction between the two regions.

B. *Crowdsourcing*

As previously described, the crowdsourcing approach to social bot detection involves leveraging human detection to identify patterns across given account profiles or the content shared by human and social bot accounts. The role of the human is to distinguish between bot accounts and

TABLE I
Summary of Datasets collection & selected features for Machine Learning studied articles

Ref#	Public	Private	Total # of Tweets	Total # of Sample	Content Features	Profile Features	Behavior Features	Extracted Features	Best Accuracy
EMD[29]		✓		1,000	✓	✓	✓	16	90%
Kantepe M. & Ganiz M. [28]		✓		1,800	✓	✓	✓	62	86%
BeDM[35]	✓	✓	5,122,000	5,658	✓		✓	*	87.32%
BotOrNot [26]	✓	✓	5.6 M	31 K	✓	✓	✓	1000	95%
Gilani Z. et. al [33]		✓	722,109	3,536	✓	✓	✓	15	86.44%
Alarifi A. et al. [27]	✓	✓		3,020	✓		✓	8	88%
DeBot [36]		✓		9,134	✓	✓	✓	**	94%
Chu Z. et al [37]		✓	8,350,095	6,000	✓	✓	✓	11	96%

* They focused in two types of features content and behavior without giving details.
**This study technique is different from all, they focus on dynamic time warping and temporal activity of a user account

TABLE II
SUMMARY OF USED LEARNING MACHINE TECHNIQUES

Machine Learning Classifier	#Ref
Bayes-theorem	[37], [27], [28], [29]
Support Vector Machine	[27], [28]
Neural Network	[27], [30]
Decision Tree	[37],[27], [30]
Random Forest	[33],[30],[26] , [4], [27]
Logistic Model Tree	[27]
Logistic Regression	[4], [28]
Gradient Boosted Trees	[28]
Pairwase	[36]

human accounts. For example, DARPA held a Twitter bot challenge competition in March 2015[6]. The competition involved identifying influential bots that supported pro-vaccination discussions on Twitter to serve as ground truth. Teams were asked to submit their guesses to a web server, which calculated and presented the results in real-time. All teams used human judgment to identify bots after using their own implementation of bot detection techniques based on the features they had chosen to inform their guesses. Three teams out of the six scored the highest in terms of their ability to detect the bot accounts.

Another use of human annotation in bot detection involved constructing ground truth datasets [33], [27]. Four annotators were tasked with classifying and labeling Twitter profiles into two categories, bot or human, and providing a justification for their choice [33]. They were provided with a list of attributes or features to consider, such as creation date of account, number of tweets, number of favorited tweets, etc. To ensure the reliability of the annotation experiments, the researchers employed Cohen's kappa (k) coefficient and average pairwise inter-annotator agreement across all dataset bands.

A study by Alarifi et al. involved 10 volunteers rating and labeling 2000 random accounts to build a ground truth dataset [27]. They evaluated the accuracy and reliability of their labeling process by injecting their ground truth dataset with a subset of 1020 Twitter bot accounts from another ground truth reference. They achieved an accuracy of 96% with 4% error rate during the labeling process.

C. Machine Learning

In this section, we present a review of the different machine learning methods that have been employed in recent social bot detection efforts (see Table II). The objective of machine learning techniques is to solve the problem through the use of large amounts of data that have many variables. Using machine-learning techniques can facilitate the detection of behavioral patterns based on the features of users' accounts to ascertain the likelihood of those accounts being bots or human [3].

A framework for bot detection on Twitter was proposed by Kantepe and Ganiz, who applied machine learning algorithms after an extensive process of data preprocessing and feature extraction [28]. They employed the Twitter API and Apache Spark to collect the data. They collected 1,800 Twitter accounts and extracted 62 features that were categorized into three types: User Features, Tweet Features, and Periodic Features. They used percentages of 60-40% and 70-30% and 90-10% for the training set and test set. They used four classifiers: logistic regression, multinomial naive-byes, support vector machine, and gradient-boosted trees. The highest accuracy result was 86% for gradient-boosted trees, and the F1 score was 83%.

A study by Ersahin et al. presented a classification method to detect fake accounts on Twitter by testing the dataset using entropy minimization discretization (EMD) [29]. They tested their data before and after the discretization was performed and employed the naive base classifier and F-measure to calculate the prediction accuracy of the system. The results were 85.5% before the proposed technique was applied and this was improved to 90.41% after the preprocessing process using the discretization technique on selected features.

TABLE III
SUMMARY OF 20 COMMON FEATURES

Protected account	Screen name
Profile image	number of links
tweets count	mention
retweet	# of friends
verified	follower count
favorite count	age of account
lists	user replies
rate of media	description
Entropy of tweets	# of words per tweet
URL rate	# hashtage

BeDM [35] is a proposed behavior-enhanced deep learning model for bot detection in social networks. It implements convolutional neural network (CNN) layers with LSTM long-short-term memory and hidden layers to capture the latent temporal patterns of users' tweet history and behavior. They used a public dataset in their experiment and an additional collection of 1000 tweets using the Twitter API. They employed tenfold cross-validation with measuring precision, recall, and F1 scores to evaluate their model and compare it against similar baselines. The BeDM performance achieved the highest in the F1 score at 87.32% in comparison to alternative baselines[34], [44] [45]. The recall score was 86.26%, and the precision was 88.41%.

Davis et al. proposed a system called BotOrNot [26] that employed the random forest classifier to evaluate social bots. Their classification system extracted more than 1000 features from 6 main classes through which they analyzed network features, user features, friends features, temporal features, content features, and sentiment features. They used tenfold cross-validation to measure their system performance and scored 95% AUC. In 2017, they extended their work to improve the accuracy of the evaluation through the use of a new training data [4]. The new system achieved 0.85% AUC less than their previous results. This was attributed to the challenging sample of bot accounts that were added to evaluate their classifier. They repeated the experiment with two datasets and achieved 0.94% AUC in detecting simple and sophisticated bot accounts. Moreover, they tested their system with 14 million accounts to estimate the fraction of bot population. The results indicated that this ranged between 9% and 15% [4], as mentioned earlier in this article. More interestingly, they used clustering analysis to group accounts according to behavior, and they identified three types of bot accounts: spammers, self-promoters, and accounts that use applications to post content.

Gilani et al. classified Twitter accounts into automated agents and human users [33]. They used their own platform, Stweeler [34], to collect their data. They collected 2.5 and 3 million tweets a day and partitioned their dataset into four subsets: 10M, 1M, 100 K, and 1K, each of which represented the popularity of the account based on the number of followers. They used human annotation for the labeling process and Cohen's kappa coefficient to maintain the reliability of the annotator judgments. The total number of accounts included in the testing phase was 3,536 across the four bands. After performing a statistical calculation, the authors extracted 15 features and employed the random forest classifier. They performed three sets of experiments through which they ran fivefold cross-validation by training and testing. The accuracy rate was 86.44%, precision was 85.4%, recall was 82.2%, and F-measure was 83%. They found 6 features scored the highest among the 15 features.

Alarifi et al. analyzed the detection features of bot accounts. They collected data that consisted of 1.8 million accounts and then randomly selected 2000 accounts for the sample after manually labeling them into human, bot, and hybrid accounts [27]. They employed two feature methods to extract effective features and used principal component analysis and correlation-based methods. They selected eight features to evaluate their models by applying four machine learning algorithms: decision tree, Bayesian network, support vector machine, and multilayer artificial neural network. For the purpose of performance measurements, they used six indicators: detection rate, error rate, TP/FP, precision, recall, and F-measure. Both random forest and Bayes net classifiers performed better with 88% and 86.74% in sequence.

DeBot is a bot detection system that used a pairwise approach by applying a lag-sensitive hashing technique to cluster user accounts into correlated sets in real time [36][39]. They employed dynamic time warping to

TABLE V
SUMMARY OF USED PERFORMANCE MEASUREMENTS

Measurement technique	#Ref
Random Walk	[31]
ROC (FPR/FNR)	[31], [4], [27]
AUC	[31], [4]
Precision	[30] , [35], [27] , [33][36]
Recall	[30], [35], [27],[33], [28]
Accuracy	[30], [27], [4], [28],[33], [29]
F-measure	[35],[27], [28], [33], [29]
Counting Credits at vertex	[32]
Error Rate	[27]
CDF	[27] ,[37]
Confusion Matrix	[29] ,[37]

capture the correlation between posting activities using a cross-correlation-based random projection technique. As such, the synchronized behavior in a sequence of 40 activities acted as an indicator of automated accounts. They calculated and compared their model against five alternative methods such as Twitter, BotORNot, etc. They achieved 94% precision in their generated daily reports.

Chu et al. studied the features related to tweeting behavior, tweet content, and account properties [37] to detect the automation of bots. They categorized accounts into human, bot, and cyborg according to the investigated features. Their classification system incorporated an entropy-based component to detect regularity of timing to measure automation, a spam detection component in the form of a Bayesian classification that detected text patterns as a means of detecting spam, account properties, and a random forest classifier as a decision maker. They collected data covering 512,407 accounts using the Twitter API. They constructed their ground truth sample by using 6000 accounts divided equally per human, bot, and cyborg. They extracted 8 features and implemented a random forest classifier with tenfold cross-validation. They employed a confusion matrix to measure the system performance and achieved an average score of 96%.

V. DISCUSSION

Thus far, this paper has examined some of the approaches that have been employed to detect the activities of social bots on Twitter. As mentioned earlier, social bot accounts are more deceptive than ever before, and it is becoming increasingly difficult to develop systems that can detect these applications. To make progress in this area, there is a need to consider the main challenges and factors that impact social bot detection activities, as understanding these challenges will facilitate the development of new technologies that can address the issues that are at play. In this section, we highlight the factors that commonly represent challenges in social bot detection. These factors are datasets, common features, methods employed, and performance measures.

A. *Datasets:*

To study and understand the behavior of social bots in comparison to human behavior in social networks, it is essential to maintain datasets that consist of both human and bot accounts. In Table I, the difference in data sizes across the reviewed papers is obvious. Researchers encounter two issues with datasets. The first of these concerns the availability of recent public datasets on which to perform experimentation. Some studies use their own platform to collect data as a solution to avoid this issue and avoid situations in which they have a limited amount of Twitter API request per hour during the process of collecting the data, as was the case in [34]. However, it usually takes time to collect data that is of a decent size and contains enriched content. For example, the maximum tweets per user that the API can provide is 3200. In this regard, a good number of studies have established a system of sharing processed datasets so that other researchers can then use as a baseline or for comparative purposes. However, the datasets that are available will be limited to the features and size in which they were issued to avoid violating the privacy of the users.

The second issue relates to developing a trained dataset that is diverse in terms of the content of the bot accounts. This is especially significant in studies that employ a machine learning approach to bot detection. Labeling a sample that is well defined in terms of size and content can be very difficult to achieve. Therefore, many researchers employ human annotation of a reasonable training sample to perform this task, even though it takes time and is prone to human error. One solution to this problem that some studies have identified is using accounts that Twitter has suspended for use as social bots. However, this solution is not significantly accurate because human users are sometimes suspended for violating Twitter's terms of use. In addition, this approach will rely on the researcher's ability to obtain the data for suspended accounts, and this is not readily available.

B. *Common Features*

Social bot detection is based on classifications of selected features to sort accounts into either legitimate or bot accounts. However, the studies reviewed in this paper highlight how common features are used to detect social bot accounts. These include factors related to timing, automation, text use, sentiment, and clickstream behavior. Therefore, we cannot assume a social bot depends on one feature without addressing the other features [37]. In Table III, we summarize the common features that are extracted from a full set of features in the reviewed papers to measure the likelihood of an account being a human or bot. In general, the extracted features can address the network features to identify the community features. We can also identify the social connections of users and ranking through performing content and behavioral analysis. For example, if an account is verified or protected, it is a logical indicator that it is a human account, not a bot account. The profile features that are extracted from the metadata, such as profile image, screen name, and description, may also indicate the nature of the account. For example, a default profile image is a sign of a new user or a bot account [27]. The temporal pattern, such as the average of tweeting and retweeting ratios, for example, can be a sign of bot activity if it occurs with small inter-arrivals [35][40]. Therefore, using an entropy component to detect behavior as part of the classification system is essential.

In addition, the rate of posting similar content with URL can be an indicator of a spammer [10][41]. In other words, the URL feature can be used to detect the link farming behavior that is typically employed by spammers and bot accounts [42]. Also, using the mention feature in association with the URL and number of link feature and entropy of tweets can indicate a bot account with malicious intention [7]. Moreover, if the number of followers is high yet the account is relatively new, it's likely that the followers are fake and the account is a bot.

C. *Methods Employed to Detect Bots*

The literature review of the recent studies that have been performed in this domain highlights how different approaches to detecting social bots have been implemented. The main methods focus on the primary components of social networks, such as network structure, content, and behavior features. In general, the content and behavioral characteristics of bot accounts are employed in off-the-shelf machine learning algorithms [4]. This section evaluates some of the methods that were most commonly used within the studied papers to detect social bot accounts. Table IV presents a summary of the advantages and disadvantages of each method (See Table IV).

Using a graph-based method, [31] studied the trust propagation of the network in which the ranking of users is easier based on the trust scores. However, this approach is sensitive to the selection of trust seeds, and it works based on assumptions, which are not always accurate. Using the same method, [32] and [30] measured the influence of social users based on the social network graph. This approach is useful for visualization and measuring the influence rate of a social network based on the distance measurement of the centrality of the influencer nodes. However, the computation cost can be high if the targeted network is large and real. Therefore, using a synthetic network to apply this approach is useful, even though the results in the real network are sometimes unpredictable.

The crowdsourcing method can be employed to effectively build ground truth data and annotation tasks. This approach employs human intelligence to identify different patterns. The problem with this approach is that it consumes time and is prone to human error. However, as described in Section III, there are some solutions by which the annotation task can be validated during the preprocessing phase to maintain the best results.

The survey of the existing literature revealed that researchers are more likely to employ machine learning methods than the other two approaches. The majority of reviewed papers used tree-based approaches and Bayes-theorem. The random forest classifier was the most commonly employed

TABLE IV
ADVANTAGES AND DISADVANTAGES OF DETECTION METHODS

Method		Ref#	Advantages	Disadvantages
Graph Based	Trust Propagation	[31]	- good and easy to rank users - good to represent two trust classes	- depends on assumptions - sensitivity to selection
	Centrality and properties	[32], [30]	- good to measure influence nodes - graph visualization	- large network will be computationally cost - depends on assumptions - sensitive to neighborhood
Crowdsourcing	Annotation task to build ground truth	[27], [33], [6]	- using human intelligence - good accuracy	- paid task usually - time consuming
Machine Learning	Tree -Based	[33], [37], , [4], [27] [30]	- Performance accuracy - widely used	- overfitting problem
	Bayes-theorem	[37], [27], [28], [29]	- easy to implement - good performance	- less performance in case of large features - learning time
	Support Vector Machine	[27], [28]	- Good accuracy based in kernel choice - solve regression problems	- Training & testing time is affected by data size - sensitive to choice of kernel and parameter
	Neural Network	[27], [30] [35]	- performance accuracy	- Training time is slow - need large sample to perform well
	Pairwise	[36]	- easy to implement in behavior change	- evasion of similarity metrics

classifier within the reviewed methods. The advantages of this classifier are that it is less complex in terms of tuning and achieves a more accurate performance. However, the complexity of the tree will generate overfitting, as is the case with most decision tree algorithms. Bayes-theorem and random forest were widely used in the studies described in the literature. Bayes-theorem as a statistical theorem is fast in terms of training and prediction time. However, the performance of this classifier is better when the data set contains a relatively low number of features. Many researchers have employed a support vector machine to reduce the error rate in the classification process. However, SVM depends on the selective kernel and parameter. In addition, a major disadvantage of this classifier is that it depends on the use of a large training set to increase performance. Similar to SVM, neural network effectiveness depends on the sample size; when the sample size is large, the vector performs well. One detection method that was applied in the literature was that of the pairwise similarity technique [36]. This method can effectively detect bots based on the similarity of the profile activity. However, the extent to which the method can be scaled depends on storing and analyzing the user's history.

D. *Performance Measures:*

Within the investigated studies, different performance measurements have been used to evaluate social bot detection classification techniques. The approach that is commonly used to measure performance in these articles is the accuracy rate, which relates to the percentage of accounts that are correctly classified with respect to the whole sample. However, using the accuracy rate alone is not sufficient to evaluate the chosen classifier. The Chu, Zi et al. study evaluated the accuracy of each feature, and the result was meaningless when compared to using a confusion matrix to evaluate the whole features for each class [37].

Table V presents the list of the performance measures that were employed in each of the reviewed papers. The majority used classifiers and tenfold cross-validation and fivefold cross-validation to validate their results. Five to six studies used F-measure, precision, and recall to measure performance. These performance measurements are appropriate for the bot detection problem since it is ultimately a binary classification problem.

VI. CONCLUSION

In this paper, we reviewed the bot detection methods that have been employed in recent studies on the Twitter social network. We quantified the existing papers according to the detection scheme and classifiers employed. We then summarized the main observations on the reviewed literature within four main subsections: dataset, analyzed features, classier, and performance measures.

The findings revealed that social bot detection is challenging and this challenge is exacerbated as the social network volume increases. Bots employ sophisticated mechanisms to avoid detection and researchers have yet to develop viable methods by which such mechanisms can be identified. As such, there is a requirement for ongoing studies into bot detection approaches. Twitter is encouraged to develop systems that can recognize automated tweets and tag them with a unified label so that they can be readily identified by users. The research community is encouraged to collaborate to build a periodically updated public dataset that includes recently detected bots.

REFERENCES

[1] Twitter, October 2017. [Online]. Available: https://investor.twitterinc.com/results.cfm

[2] C. Smith. (2017, November) 400 amazing twitter statistics and facts. [Online]. Available: https://expandedramblings.com/index.php/

[3] 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.

[4] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," arXiv preprint arXiv:1703.03107, 2017.

[5] N. Abokhodair, D. Yoo, and D. W. McDonald, "Dissecting a social botnet: Growth, content and influence in twitter," in Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing. ACM, 2015, pp. 839-851.

[6] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, "The darpa twitter bot challenge," Computer, vol. 49, no. 6, pp. 38-46, 2016.

[7] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@ spam: the underground on 140 characters or less," in Proceedings of the 17th ACM conference on Computer and communications security. ACM, 2010, pp. 27-37.

[8] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proceedings of the 26th annual computer security applications conference. ACM, 2010, pp. 1-9.

[9] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach." DBSec, vol. 10, pp. 335-342, 2010.

[10] X. Zhang, S. Zhu, and W. Liang, "Detecting spam and promoting campaigns in the Twitter social network," in Data Mining (ICDM), 2012 IEEE 12th International Conference on. IEEE, 2012, pp. 1194-1199.

[11] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, and J. H. Park, "Social network security: Issues, challenges, threats, and solutions," Information Sciences, vol. 421, pp. 43-69, 2017.

[12] M. Shafahi, L. Kempers, and H. Afsarmanesh, "Phishing through social bots on Twitter," in Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016, pp. 3703-3712.

[13] J. Zhang, R. Zhang, Y. Zhang, and G. Yan, "The rise of social botnets: Attacks and countermeasures," IEEE Transactions on Dependable and Secure Computing, 2016.

[14] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, "The socialbot network: when bots socialize for fame and money," in Proceedings of the 27th annual computer security applications conference. ACM, 2011, pp. 93-102.

[15] A. Gupta, H. Lamba, and P. Kumaraguru, “$1.00 per rt# bostonmarathon# prayforboston: Analyzing fake content on twitter,” eCrime Researchers Summit (eCRS). IEEE, 2013, pp. 1-12.

[16] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu, “Design and analysis of a social botnet,” Computer Networks, vol. 57, no. 2, pp. 556-578, 2013.

[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] F. Amato, A. Castiglione, A. De Santo, V. Moscato, A. Picariello, F. Persia, and G. Sperlí, “Recognizing human behaviours in online social networks,” Computers & Security, 2017.

[19] S. Sivanesh, K. Kavin, and A. A. Hassan, “Frustrate twitter from automation: How far a user can be trusted?” in Human-Computer Interactions (ICHCI), 2013 International Conference on. IEEE, 2013, pp. 1-5.

[20] G. Laboreiro, L. Sarmento, and E. Oliveira, “Identifying automatic posting systems in microblogs,” Progress in Artificial Intelligence, pp. 634-648, 2011.

[21] C. M. Zhang and V. Paxson, Detecting and Analyzing Automated Activity on Twitter. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 102-111. [Online]. Available: https://doi.org/10.1007/978-3-642-19260-9_11

[22] N. Chavoshi, H. Hamooni, and A. Mueen, “Temporal patterns in bot activities,” in Proceedings of the 26th International Conference on World Wide Web Companion. International World Wide Web Conferences Steering Committee, 2017, pp. 1601-1606.

[23] Q. Fu, B. Feng, D. Guo, and Q. Li, “Combating the evolving spammers in online social networks,” Computers & Security, vol. 72, pp. 60-73, 2018.

[24] S. Lee and J. Kim, “Early filtering of ephemeral malicious accounts on twitter,” Computer Communications, vol. 54, pp. 48-57, 2014.

[25] C. Freitas, F. Benevenuto, S. Ghosh, and A. Veloso, “Reverse engineering socialbot infiltration strategies in twitter,” in Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015. ACM, 2015, pp. 25-32.

[26] 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. International World Wide Web Conferences Steering Committee, 2016, pp. 273-274.

[27] A. Alarifi, M. Alsaleh, and A. Al-Salman, “Twitter turing test: Identifying social machines,” Information Sciences, vol. 372, pp. 332-346, 2016.

[28] M. Kantepe and M. C. Ganiz, “Preprocessing framework for twitter bot detection,” in Computer Science and Engineering (UBMK), 2017 International Conference on. IEEE, 2017, pp. 630-634.

[29] B. Er¸sahin, Ö. Akta¸s, D. Kılınç, and C. Akyol, “Twitter fake account detection,” in Computer Science and Engineering (UBMK), 2017 International Conference on. IEEE, 2017, pp. 388-392.

[30] A. Mehrotra, M. Sarreddy, and S. Singh, “Detection of fake twitter followers using graph centrality measures,” in Contemporary Computing and Informatics (IC3I), 2016 2nd International Conference on. IEEE, 2016, pp. 499-504.

[31] J. Jia, B. Wang, and N. Z. Gong, “Random walk based fake account detection in online social networks,” in Dependable Systems and Networks (DSN), 2017 47th Annual IEEE/IFIP International Conference on. IEEE, 2017, pp. 273-284.

[32] J. Zhang, R. Zhang, J. Sun, Y. Zhang, and C. Zhang, “Truetop: A sybilresilient system for user influence measurement on twitter,” IEEE/ACM Transactions on Networking, vol. 24, no. 5, pp. 2834-2846, 2016.

[33] Z. Gilani, E. Kochmar, and J. Crowcroft, “Classification of twitter accounts into automated agents and human users.”

[34] Z. Gilani, L. Wang, J. Crowcroft, M. Almeida, and R. Farahbakhsh, “Stweeler: A framework for twitter bot analysis,” in Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016, pp. 37-38.

[35] C. Cai, L. Li, and D. Zengi, “Behavior enhanced deep bot detection in social media,” in Intelligence and Security Informatics (ISI), 2017 IEEE International Conference on. IEEE, 2017, pp. 128-130.

[36] N. Chavoshi, H. Hamooni, and A. Mueen, “Debot: Twitter bot detection via warped correlation.” in ICDM, 2016, pp. 817-822.

[37] 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, vol. 9, no. 6, pp. 811-824, 2012.

[38] 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.

[39] N. Chavoshi, H. Hamooni, and A. Mueen, “Identifying correlated bots in twitter,” in International Conference on Social Informatics. Springer, 2016, pp. 14-21.

[40] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, “Who is tweeting on twitter: human, bot, or cyborg?” in Proceedings of the 26th annual computer security applications conference. ACM, 2010, pp. 21-30.

[41] F. Ahmed and M. Abulaish, “A generic statistical approach for spam detection in online social networks,” Computer Communications, vol. 36, no. 10, pp. 1120-1129, 2013.

[42] M. Chakraborty, S. Pal, R. Pramanik, and C. R. Chowdary, “Recent developments in social spam detection and combating techniques: A survey,” Information Processing & Management, vol. 52, no. 6, pp. 1053-1073, 2016.

[43] Kwak, H., Lee, C., Park, H., & Moon, S. “What is Twitter, a social network or a news media?”. In Proceedings of the 19th international conference on World wide web (pp. 591-600). ACM. 2010.

[44] Lee, K., Eoff, B. D., & Caverlee, J. .” Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter”. In ICWSM. 2011

[45] Morstatter, F., Wu, L., Nazer, T. H., Carley, K. M., & Liu, H. “ A new approach to bot detection: Striking the balance between precision and recall”. In Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on (pp. 533-540). IEEE. 2016