IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON INTELLIGENT BIG DATA ANALYTICS FOR INTERNET OF THINGS, SERVICES AND PEOPLE

Received February 23, 2021, accepted March 13, 2021, date of publication April 22, 2021, date of current version May 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074953

Tweet-Based Bot Detection Using Big Data Analytics

ABDELOUAHID DERHAB[1], RAHAF ALAWWAD[2], KHAWLAH DEHWAH[2], NOSHINA TARIQ[3], FARRUKH ASLAM KHAN[1], (Senior Member, IEEE), AND JALAL AL-MUHTADI[1,2]

[1]Center of Excellence in Information Assurance (CoEIA), King Saud University, Riyadh 11653, Saudi Arabia
[2]College of Computer and Information Sciences, King Saud University, Riyadh 11653, Saudi Arabia
[3]Department of Computer Sciences, Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad 44000, Pakistan

Corresponding author: Farrukh Aslam Khan (fakhan@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University, Saudi Arabia, through the Research Group under Grant RGP-214.

ABSTRACT Twitter is one of the most popular micro-blogging social media platforms that has millions of users. Due to its popularity, Twitter has been targeted by different attacks such as spreading rumors, phishing links, and malware. Tweet-based botnets represent a serious threat to users as they can launch large-scale attacks and manipulation campaigns. To deal with these threats, big data analytics techniques, particularly shallow and deep learning techniques have been leveraged in order to accurately distinguish between human accounts and tweet-based bot accounts. In this paper, we discuss existing techniques, and provide a taxonomy that classifies the state-of-the-art of tweet-based bot detection techniques. We also describe the shallow and deep learning techniques for tweet-based bot detection, along with their performance results. Finally, we present and discuss the challenges and open issues in the area of tweet-based bot detection.

INDEX TERMS Social media, Twitter, big data analytics, shallow learning, deep learning, tweet-based bot detection.

I. INTRODUCTION

Nowadays, social media is one of the most popular tools used by people to communicate with one another. It is also largely used by organizations to reach out to customers. In [1], it has been reported that there are 3.5 billion active social media users globally. Facebook, Twitter, LinkedIn, and other social media networks are used by organizations to improve brand visibility and boost their sales. Twitter is one of the most popular social media platforms. It has 340 million active users who are allowed to communicate at a large scale and share their opinions about different topics. Twitter could be targeted by various kinds of attacks. For example, a spear phishing attack in July 2020 led to the hijack of high-profile Twitter accounts [2]. Also, fraudulent accounts could be created to impersonate legitimate users and organizations.

Twitter can also be exploited by botnet, which is a set of malicious accounts that operate under a botmaster, and are controlled by software programs rather than human users. The tweet-based social media bots pose serious security risks to Twitter users. These bots are used to spread fake contents, phishing links, and spams. Although they are not used as bots to launch DDoS attacks, they could be utilized as Command and Control (C&C) infrastructure to coordinate DDoS attacks [3], [4]. They are capable of interacting with human accounts to deceive the users and hijack their accounts. These bots are also used as tools to launch large-scale manipulation campaigns to influence public opinions. According to a study [5], 52% of online traffic is generated by botnets, and the rest is produced by actual users. It is also worthy to note that some bots are found with over 350,000 fake followers. To deal with the above issues, there is a need to develop detection systems that can accurately distinguish between Twitter bot accounts and human accounts. Twitter data represent one of the examples of big data as around 500 million tweets are generated every day, i.e., 6,000 tweets every second [6].

Big data analytics has been widely used in different fields [7]–[11] to process large amount of data, discover hidden patterns, and find correlations among data points. Artificial intelligence techniques are increasingly leveraged by big data analysis. In particular, shallow (conventional) and deep learning techniques have received considerable attention from the academia and industry due to their success in dealing with heterogeneous and complex data, automatic learning

The associate editor coordinating the review of this manuscript and approving it for publication was Yuan Tian.

of models, revealing unseen patterns, identifying dependencies, and getting insights from analyzing data.

Artificial intelligence has been extensively used by Twitter to determine tweet recommendations for users. In fact, deep neural networks are applied on Twitter data to determine the relevant content for users, and hence improve their experience on the platform [12]. Artificial intelligence has played an important role in fighting inappropriate content. In 2017, about 300,000 accounts were suspended and identified with the help of artificial intelligence tools rather than humans.

This review aims at providing an overview of different tweet-based bot detection methods that use shallow and deep learning techniques to distinguish between human accounts and bot accounts. In particular, the main contributions of the paper are the following:

1) A taxonomy, which classifies the state-of-the-art on machine learning techniques for tweet-based bot detection, is presented.
2) A comprehensive review is presented on shallow and deep learning techniques for tweet-based bot detection, which covers the solutions up to year 2020.
3) The challenges and open issues related to tweet-based bot detection are highlighted and discussed.

The rest of the paper is structured as follows: Section II discusses the related surveys on tweet-based bot detection techniques. Section III presents the state-of-the-art related to deep and shallow learning based detection methods, followed by a discussion and analysis in Section IV. Finally, the conclusions are presented in Section V.

II. RELATED SURVEYS

In the literature, there exist some previous surveys that discuss and review existing papers published on social bot and spam detection, similar to this work. However, each one has its own limitations and strengths. Therefore, in this section, we briefly describe each survey and summarize it in Table 1.

Kabakus and Kara [13] provided a short comparative survey of the research work in the field of Twitter spam detection within the year range of 2009-2015. They described different detection methods within four categories: account-based, tweet-based, graph-based, and hybrid-based methods. The account-based methods were shown to leverage the user profile's metadata like followers and following count and other derived features such as age of the account. While in graph-based methods, features like distance and strength of connectivity between users were shown to be used for spam detection. However, in tweet-based methods, the survey mainly focused on detecting spam using URL and its derived features, such as length and domain name. To detect a spam user, posted URLs were analyzed and classified as malicious or benign. Besides this, the authors highlighted overlooked features that were argued to improve the spam detection.

Another comparative survey was presented by Chakraborty *et al.* [14] in the field of multiplatform spam user detection. The authors recognized that different platforms, such as e-mails, blogs, or microblogs, require different techniques and features to achieve accurate detection. Therefore, proposed techniques within the year range of 2011-2015 were classified based on the platform that the dataset lies within. A qualitative comparison was conducted for each group of methods under the same platform. Besel *et al.* [21] observed that the botnet used a URL network shortening services and redirections to obfuscate the actual landing pages. They disclosed that users clicked on these URLs, found the botmaster establishing the Bursty botnet, and registering landing pages on phishing websites. They confirmed that the botmaster is still successful in owning Twitter bot-related services. This study includes a review and insight into Twitter's cyberspace infrastructure, cybercrime operation, and the dark markets.

Alothali *et al.* [15] summarized recent research work in the field of Twitter social botnet detection. They provided an analytical review of each proposed method with its limitations and advantages. The techniques were classified into three main categories, namely graph-based, machine learning-based, and crowdsourcing based techniques. The crowdsourcing technique uses human intelligence to identify various patterns, which is stated to be the most error prone out of the three techniques. It was also shown that machine learning methods and, more specifically, random forest classifiers are the most commonly used for detecting social bots in Twitter users. Latah [16] presented a comprehensive review focusing on malicious social bots' stealthy manner and their detection techniques. The author precisely reviewed detection approaches, which are graph-based, machine learning based, and emerging approaches. Besides, the paper reviewed the strengths and weaknesses of these techniques and the means considered by the bots to avoid detection. Consequently, the paper suggested approaches that may enhance the defense procedures against malicious bots.

One of the challenges faced in evaluating bot detection approaches is that the ground-truth data is insufficient [17]. Detection techniques were compared with different aspects such as several features, the dataset's size, and the data-crawling operation. The datasets were categorized into synthesized data, crawled from online social networks, and gathered from honey profiles that attract social bots. A detailed review of existing datasets used by researchers was studied along with the results and experimental findings. In the end, the paper highlighted the constraints of the detection approaches and proposed some directions for future work. One of the suggestions was to concentrate on detection methodologies for general purposes. Also, it was suggested to build datasets that have different sets of social bots in order to assist in the generalized evaluation of the detection techniques [17]. Guo *et al.* [18] presented a survey on Online Social Deception (OSD). OSD is a serious threat in cyberspace, especially for users that are vulnerable to such cyber attacks. Cybercriminals have exploited social network services (SNSs) to conduct risky OSD activities, such as financial fraud, data threats, or sexual/labor violence. Therefore, OSD identifies and implements proactive responses to

TABLE 1. Summary of existing surveys.

Ref.	Outline	Taxonomy	Type	Open Issues/ Future Directions	Drawbacks
[13]	- Comparative review of Twitter spam detection methods - Introduction of potentially useful and overlooked features	- Account-based - Tweet-based - Graph-based - Hybrid-based	Comparative	No	Up to 2013, Insignificant tweet-based coverage
[14]	- Survey of social spam detection and mitigation techniques - Classification according to the used electronic platform	- E-mail spam - Blog spam - Microblog spam - Bookmarking spam - Social network spam - Review spam - Location search spam - Comment spam - Cross-media spam	Comparative	No	2011-2015 Not enough coverage
[15]	- Comparative review of twitter social bot detection methods	- Graph-based - Crowdsoursing-based - ML-based	Comparative	No	Short survey and limited number of solutions
[16]	- Categorize various attack types at different stages - Detailed discussion of existing social bot detection approaches and present strengths and limitations - Suggest possible countermeasures and strategies for improving current detection techniques	- Graph-based approaches - Machine learning approaches - Emerging approaches	Comparative	Yes	Non-focused tweet-based bot
[17]	- Review of social bots and study of user interaction with bot detection tools	- Nil	Descriptive	Yes	Short description of bot detection methods and non-focused tweet-based bot
[18]	- Survey of online social deception and their corresponding countermeasures	- False information - Luring - Fake identity - Crowdturfing - Human targeted attack	Descriptive	Yes	Non-focused tweet-based bot
[19]	- Review techniques on Twitter spam detection	- Content analysis - User analysis - Tweet analysis - Network analysis - Hybrid analysis	Systematic review	Yes	Focus only on twitter spam , and no performance results
[20]	- Survey on spam URLs detection in Twitter - Performance evaluation of some machine learning classifiers	- Nil	Descriptive	No	Short survey

build credible OSD SNSs. It provided a comprehensive survey of social deceit's multidisciplinary concept focused on various OSD attacks and OSD attack types.

Researchers have recently offered several innovative approaches that have vastly increased the efficiency of spam identification. It also offers an opportunity to perform a thorough analysis on Twitter of numerous spam identification methods. Abkenar *et al.* [19] focused on extensively evaluating the current Twitter spam identification testing techniques. Analysis of the literature review shows that most current approaches depend on algorithms that concentrate on machine learning. Among these algorithms for machine learning, the major differences relate to separate methods of collection of features. Therefore, they suggest a taxonomy focused on multiple approaches and evaluations of functionality collection, namely material analysis, user analysis, tweet analysis, network analysis, and hybrid analysis. Daffa *et al.* [20] discussed the identification of spam URLs in Twitter by presenting the types of harmful activities, detection avoidance strategies, detection functionality, detection techniques, and their limitations. Via machine learning classification based on different published characteristics, they demonstrated the best results. They used four classifiers on a 10713 consumer dataset of Twitter accounts with 5358 labeled as benign and 5355 labeled as spam along with 17 stable features. The features were content-based and user-based features. The outcome revealed that of the four classifiers, the Random Forest classifier with hybrid feature methods achieved the best estimation with 96.4 percent accuracy. In comparison, J48 classifier obtained 94.5 percent accuracy score.

Differently from the above surveys, our review focuses on techniques that employ shallow and deep learning methods for the detection of tweet-based social bots.

III. TWEET-BASED BOT DETECTION

Although the detection of social bots is a challenging task, there are some works that analyzed the characteristics and behavior of bots [14], [15], [22] and offered various features that are recurrent in the majority of works. For example, ver-

ified accounts are guaranteed to be human users. Moreover, the ratio of followers to following and the age of the account are considered discriminative characteristics in detecting bots since bots generally mass-follow and have short life span [16]. The following features are mainly used by tweet-based bot detection techniques to distinguish between tweet-based bots and humans accounts [23]:

- ID: It represents the unique identifier of the tweet.
- User: It represents the user who posted the tweet.
- Created_at: It indicates the UTC time when the tweet is created.
- Text Tweet: It refers to the body of the tweet.
- Length of Tweet: It gives the number of characters in the tweet.
- #Hashtags: It indicates the number of hashtags in the tweet.
- #URLs: It indicates the number of URLs in the tweet
- in_reply_to_status_id: If the tweet is a reply, this feature represents the original tweet's ID.
- in_reply_to_user_id: If the tweet is a reply, this feature represents the author of the original tweet.
- Coordinates: It represents the geographic location of the tweet.
- Favorite_Count: It indicates how many times the tweet has been liked by Twitter users.
- Retweet_Count: It is the number of times the tweet has been retweeted
- Reply Count: It is the number of times the tweet has been replied to.
- Favorited: a boolean feature, which holds true when the tweet is liked by the authenticating user.
- Retweeted: a boolean feature, which holds true when the tweet is retweeted by the authenticating user.
- Possibly_sensitive: a boolean feature, which holds true when the tweet contains a link.

A. TAXONOMY

In this section, we describe machine learning techniques used for tweet-based bot detection. As shown in Fig. 1, the state-of-the-art techniques are classified into two major categories: shallow learning-based detection and deep learning-based detection. According to the learning approach, the shallow detection techniques are further classified into three subcategories: supervised learning, semi-supervised learning, and unsupervised learning. In supervised learning, the learning model is trained with labeled data, so it can predict the output of the new data. An unsupervised learning technique builds the model from unlabelled data. It aims to find structures and patterns within the data itself. The semi-supervised learning techniques use both labeled and unlabeled data to train the model. On the other hand, deep learning-based detection techniques are further classified into two subcategories: generative architecture based techniques and discriminative architecture based techniques. If we have input data x and we want to classify them into labels y, a generative model learns the joint probability distribution $p(x, y)$. On the other hand, the discriminative model learns the conditional probability distribution $p(y|x)$. The deep generative architecture is formed by combining a generative model and a deep neural network. It is generally associated with unsupervised learning. The deep discriminative architecture adopts supervised learning, and is built by combining a discriminative model and a deep neural network to compute and optimize $p(y|x)$. The detailed discussion on each category is given in the rest of the section.

B. DEEP LEARNING-BASED DETECTION METHODS

Recently, deep neural networks have gained noticeable attention from researchers in different fields ranging from computer vision to language processing. It has proven its effectiveness in terms of textual classification. It can process structured data like sentences and automatically produce discriminant features, thus relinquishing handcrafting features, which is expensive and requires extensive knowledge of the data. Therefore, since the overall performance of a classifier relies heavily on the quality of its data, deep neural networks such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) were employed as a feature extractor or classifier for many language processing problems, one of which is the tweet-based bot detection. However, neural networks require a certain form of input, preferably structured data, but most importantly, a numerical vector representing the data. There are several pre-trained word embedding models for that purpose, such as the popular Word2Vec model. Therefore, as a preliminary step, all text tweets are converted into a form accepted by the network using trained models.

The Long Short-Term Memory (LSTM) model is the most popular model for language processing and classification. It is an improved version of the RNN vanilla model that can maintain a memory of the past input for a longer period, preferable for long text input. Hence, most work mentioned in this section employs a variation of LSTM. For example, Kudugunta and Ferrara [24] recognized the limitation of utilizing either tweet metadata or tweet text as a single input. Therefore, a Contextual LSTM was proposed that takes both features for improved bot detection. They used the public dataset Cresci-2017 to train the model to reduce the exhibited imbalance. The Synthetic Minority Oversampling Technique (SMOTE) was used to fill the minority class without fully synthetic data that might affect the performance. Training data of 8,386 users' tweets were tokenized and loaded into the GloVE word embedding model and fed to the model for feature extraction. Besides, tweet metadata such as retweet and reply count were concatenated with the tweet text's features before classifying in the dense layer. This yielded better performance than using tweet metadata only. To prove the strategy's superiority, the model was tested using single and combined features resulting in 96% for both precision and accuracy favoring the proposed method.

Wei and Nguyen [25] proposed a bidirectional RNN to identify bot accounts in Twitter by utilizing the LSTM model.

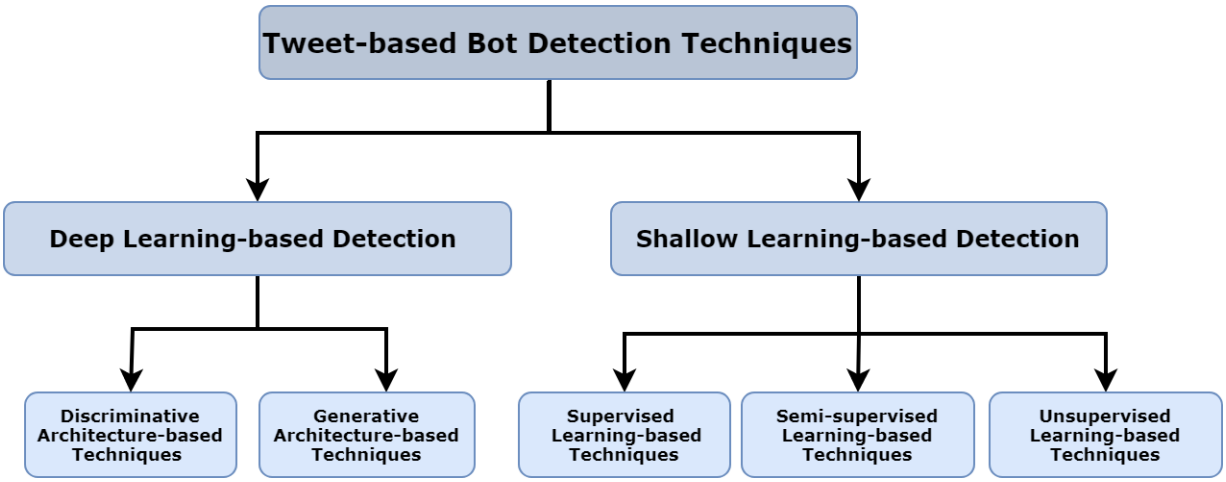

FIGURE 1. Proposed taxonomy for tweet-based bot detection techniques.

The proposed model implemented the bidirectional strategy in which tweet sentences are processed both forward and backward for each layer enabling a better understanding of the overall text context. To train the model, a public dataset Cresci-2017 is used that consists of tweets from 3,474 human accounts and 1,455 bots, resulting in 11.4 million tweets in total. Before training, each tweet was preprocessed and tokenized to fit the word embedding model. A pre-trained GloVe model was used to convert text to numerical vectors that are acceptable by the network. The vectors were fed to a three-layer model with a decreasing dropout layer that was initially set to 0.5. Two subsets of testing datasets, composed of 1,982 and 928 accounts respectively, were used to evaluate their model resulting in precision and accuracy of 93% and 95%, respectively. Mazza *et al.* [26] introduced a novel deep learning model to distinguish bots from humans using their retweet patterns termed RTbust. Before building the model, the authors analyzed the behavioral patterns of bots and humans alike. The analysis demonstrated a distinctive pattern of retweeting in terms of timing, and it was categorized into four patterns. The first is the droplet pattern, which corresponds to normal users in which there exists a fair amount of time between the tweet being posted and the retweet operation. The three remaining patterns belonged to potential bots due to their suspicious and rapid retweeting pattern.

To avoid human annotators, the authors suggested a novel unsupervised variational autoencoder LSTM model and utilized merely the timestamp of tweets and retweets. They collected 9,989,819 retweets by 1,446,250 different users. However, the model aimed to recognize sophisticated and malicious human-like bots, and thus the dataset was reduced to 63,762 users. The retweet timestamps were further reduced to remove the timestamp of non-retweets by employing the run-length encoding technique. However, the suggested model was only the first phase. It was implemented as a feature extractor, where the recognition tasks fell under a clustering algorithm such that the normal behavior is represented by noise. Approximately 1000 users of the dataset were self-annotated to evaluate the model, resulting in 93% precision and 87% accuracy. A combination of LSTM and CNN proposed by Cai *et al.* [27] used a behavioral deep model termed as BeDM to detect bot users in Twitter. They trained the proposed model with a Morstatter public dataset that consists of 5658 users. The tweet messages for each user were converted to a proper form accepted by the CNN using DeepTalk word embedding model. The former was used to extract high-level features to be fed later to LSTM to extract the overall behavior of a user. To evaluate the model, the dataset was split into 90% training and 10% testing datasets, and the results were achieved with 88.4% precision.

By integrating the Deep Q-Learning (DQL) model with Twitter's social attributes to identify social bots, a deep Q-network architecture based on Q-value updates (i.e., state-action value function) was proposed by Lingam *et al.* [30]. Each customer's social character is seen as a state, and the transition from one state to another is seen as an event. Both state-action pairs for the Q-value function are used to construct the state-transition likelihood value among the state-action pairs. For social bot identification in a proposed DQL algorithm, the learning agent chooses a certain learning action in each state with optimum Q-value. The proposed algorithm achieved an average accuracy of 93% by integrating all social parameters.

CNN could be implemented as a stand-alone bot detector if the input dataset were structured as proposed by Färber *et al.* [28]. A simple CNN architecture was used to classify textual tweets into bot or human. The model consisted of two-dimensional convolutional layers, two max-pooling layers, and two dropout layers of 50%. The model was trained using the CLEF-2019 public dataset split into 2873 users for training and 1240 for testing. Crawled tweets for each user were concatenated to form an article of tweets representing that user. A small word embedding

TABLE 2. Summary of deep learning based detection methods.

Ref.	Dataset	Tweets	Training Set	Testing Set	Pre-processing	Features	Classifier	Architecture/ Approach	Accuracy	FP	TP	Precision
[24]	Cresci-2017	11.4 m	8,386	N/A	GloVE SMOTE	Tweet & account Metadata	LSTM	Discriminative /Supervised	96%	N/A	N/A	96%
[25]	Cresci-2017	11.4 m	4,929	2,910	GloVe	Tweet Text&Tweet & Account Meta-data	BiLSTM	Discriminative/ Unsupervised	95%	6%	94%	93%
[26]	Own	446,334	62,762	1,000	Word em-bedding	Tweet Metadata	Autoencoder LSTM	Generative/ unsupervised	87%	N/A	N/A	93%
[27]	Morstatter	5,658	5,092	566	DeepTalk	Tweet Text & metadata	CNN + LSTM	Discriminative/ Supervised	N/A	N/A	N/A	88.4%
[28]	CLEF-2019	N/A	2,873	1,240	Word em-bedding	Tweet Text & metadata	CNN	Discriminative/ Unsupervised	85%	N/A	N/A	97%
[29]	Twitter	20	500	25,817	Spam de-tection	IDs, screen name, location	Bayesian classifica-tion	Supervised	N/A	N/A	N/A	89%
[30]	Twitter(Fake Project, Social Honeypot, User Popularity Band)	9,987,698, 5,613,166, 150,336	N/A	N/A	Socail bot detection	Tweet-based, user profile-based and social graph based attributes	Deep Q-Learning (DQL)	Unsupervised	93%	N/A	N/A	80%
[31]	ASW EC2	6M	N/A	N/A	N/A	Twitter statistics + Category vector + Sentiment + LDA	Graph neu-ral network	Unsupervised	89%	N/A	N/A	N/A
[32]	Bots and Gender Profiling 2019	412,000	288,000 144,000	Bot	human tweet differenti-ation	Bi-LSTM	Deepbot	Unsupervised	79.64%	N/A	N/A	N/A
[28]	CLEF 2019	3110	2873	1240	Twitter bot	N/A	Convolutional neural network	Unsupervised	N/A	N/A	N/A	97.02%
[33]	ISIS dataset	9M	N/A	N/A	N/A	N/A	Deep neural network	Discriminative/ Unsupervised and Semi-supervised	82%	N/A	N/A	90%

layer was employed to vectorize the text and improve the layer's performance. The network was trained in an end-to-end fashion. Several word embedding models were tested, and the proposed layer outperformed, resulting in 85% accuracy and 97% precision of bot detection. Overall, the deep learning method proved to perform well with text classification, as illustrated in Table 2. However, in terms of purely tweet-based bot detection, it still lacked good performance. It was shown that combining tweet features like retweet count or length of the tweet could improve the performance of the detector. However, these features were not structured and could not be extracted directly using LSTM or CNN, which is the key advantage of these models.

Mesnards *et al.* [34] addressed the issue of identifying bots and their effect on people's views in online social networks. The bots' activities in social networks were analyzed and it was identified that they interacted with people more than other bots. A recognition system based on Ising model derived from statistical physics was created. By solving the minimum cuts query, the bots were detected. The authors demonstrated that the Ising model would classify bots with greater precision and far less data than the other state-of-the-art approaches. To quantify the effect on users' views in social networks, they established a deep network-based generalized harmonic influence centrality to assess the impact of opinions on posts in the social network. Similarly Luo *et al.* [32] presented Deepbot. It was configured to use the Bi-LSTM model for tweet research. An open-access framework was built utilizing a database server for the Web interface. This method boosted the accuracy rate by their analytical studies.

Waskale and Jain [35] detected low-quality substances from the perspective of the customer. Expectation-Maximization (EM) measurement has been used to arrange low-quality substances coarsely. With the distinguished highlights, they integrated word analysis and developed a watch-word boycott lexicon to enhance recognition. They marked a large Twitter data set with 100,000 tweets and continuously

conducted low-quality content discovery based on the major highlights and word-level inspection. The results revealed an accuracy of 0.9711 and an F1 of 0.8379 based on continuous execution of woodland classifier in the tweet detection of low-quality substance. Lingam *et al.* [36] created a weighted signed Twitter network graph focused on behavioral similarity and confidence values as weighted edges. To detect related styles of interaction (malicious or not) among Twitter participants, the behavioral similarity is calculated from the point of view of Twitter message resemblance, mutual URL resemblance, concern resemblance, and social resemblance. In contrast, a random walk model calculates the participant's trust value. They developed two algorithms — Social Botnet Community Detection (SBCD) and Deep Autoencoder-dependent SBCD (DA-SBCD) — where the former recognized social botnet communities with malicious behavioral similarities. Simultaneously, the latter more reliably reconstructed and detected social botnet communities in the middle of many forms of malicious behavior. Two Twitter repositories analyzed the efficiency of the proposed algorithms. Concerning uniform reciprocal knowledge, precision, recall, and F-measurement, experimental findings indicated better efficiency than the current schemes. Specifically, the DA-SBCD algorithm was around 90% accurate and displayed up to 8% NMI enhancement.

Shah [37] used machine learning to identify individual Twitter botnet accounts. The dataset in the experiment included 1,321 bots and 1,476 user profile-related non-bots. The author used semantic ranking and user profiles in classifying Twitter bots. Using Google's pre-trained deep learning algorithm, an average semantic score is calculated. Each tweet has a set vector size of 512. The Neural Network model proved to be better than a Multinomial Naive Bayes approach with 67.7% accuracy. Using a convolutional neural network, Farber *et al.* [28] defined Twitter bots based on their written messages. They used different embedding approaches and architectures of convolutional neural networks and compared their efficiency. They achieved a precision of 90.34% on the real test data collection of the CLEF 2019 Bots Profiling Subtask dataset.

Chen [38] suggested a real-time demand identification system for low-quality goods. A preliminary research-based survey is deliberately structured to collect views on numerous types of low-quality material. Primary and indirect characteristics, including new technology, are used to define different types of low-quality content. Word-level research was integrated with defined attributes, creating a blacklist dictionary for keyword detection efficiency. The author marked a comprehensive Twitter 100,000 tweet dataset and conducted real-time low-quality content identification based on significant characteristics and word-level study. Because knowledge travels much faster and wider through Online Social Networks (OSN), rumors and misinformation can be easily propagated through OSNs. Due to the possible damage that false information may do to the public, gossip identification has become a big yet difficult research topic. The author proposed an RNN and Autoencoder (AE) framework to differentiate rumors as deviations from other trustworthy microblogs centered on user interactions to identify the few yet potentially dangerous rumors that may trigger public issues. The reason people still read OSN rumors is that today, particularly for news updates, OSN plays a major role as an information-sharing platform. Orthodox media news was long used to estimate market movement; however, it now helps to use OSN news material to forecast stock market behavior. It picks accounts from China's largest OSN by deleting emotional and Latent Dirichlet Allocation (LDA) components. To forecast Chinese stock market volatility, they input these features and technical indicators into a new RNN-boost model. The research discussed in this study illustrated OSN's advantages and disadvantages, providing methodologies and applications to maximize positive impacts, thus mitigating OSN's future negative impacts.

C. SHALLOW LEARNING-BASED DETECTION METHODS

Shallow learning based techniques have been split into three categories in the literature: (i) supervised, in which the dataset is labeled, (ii) semi-supervised, in which a small amount of labeled data is combined with a large amount of unlabeled data during the training phase, and (iii) unsupervised, in which the model does not require labeling but only the ground-truth for some of the data is required to evaluate the model. In this section, we present these methods used for tweet-based bot detection and compare them based on several parameters. Table 3 contains the summary of the shallow detection techniques. The detailed analysis is discussed in the following sub-sections:

1) SUPERVISED LEARNING TECHNIQUES

Knauth [39] proposed to combine both tweet-based and account-based features for more accurate bot detection. The bots are assumed to express different behavior than human users. Therefore, behavioral and emotional features were derived from tweet-based features. These include calculating the user data properties such as the minimum, maximum, and mean of tweeting statistically. Several classifiers were employed to evaluate the selected features' consistency, including support vector machines, random forests, logistic regression, and multi-layer perceptron. However, the Adaboost classifier performed best on the public dataset Cresci-2017, with 6708 users for training and 1677 for testing, resulting in 99% for both precision and accuracy. Wang and Paschalidis [70] suggested a botnet identification approach evaluating social node interactions. The approach had two phases: (1) detecting an anomaly in an ''interaction'' graph between various nodes using large results of deviation on the degree distribution, and (2) Detecting community in a social ''correlation'' graph where nodes are connected by edges having strong correlated communication. The proposed method is applied to real botnet traffic and the performance is compared with existing community detection approaches.

TABLE 3. Summary of shallow learning based detection methods.

Ref.	Dataset	Tweets	Training Set	Testing Set	Pre-processing	Features	Classifier	Approach	Accuracy	FP	TP	Precision
[39]	Cresci-2017	N/A	6708	1677	Tokenization	Tweet & account metadata	Adaboost	Supervised	99'%	N/A	N/A	99%
[40]	Own Libya Own Arab Honey-pot	1726359 725179	94,535 6285	N/A	LDA reduction	Tweet text & Metadata	BoostOR	Supervised	N/A	N/A	N/A	75% 71%
[41]	Own	10000	4000	1000	N/A	Tweet Metadata	J48-HP	Supervised	N/A	90	1041	99%
[42]	Own	N/A	N/A	N/A	N/A	Tweet & account metadata	Random For-est	Supervised	N/A	N/A	N/A	95%
[43]	Own	23,100	16,170	6,930	N/A	Tweet metadata	ANN	Supervised	94.4%	N/A	N/A	N/A
[44]	Own (GTR1)	3020	2000	1020	N/A	Tweet metadata & Account metadata	DT, NB, SVM, Multilayer ANN	Supervised	91.39%	0.8	0.91	0.92
[45]	Cresci-2017	13,253,492	N/A	N/A	NLP	Tweet metadata & Account metadata	One-class classification	Supervised	N/A	N/A	N/A	N/A
[46]	Own	1000	N/A	N/A	Entropy Mini-mization Dis-cretization (EMD)	Tweet metadata & Account metadata	Naïve Bayes	Discriminative/ Supervised	90.9%	31	441	N/A
[47]	Cresci-2017	Tweets: 6,888,102 Ac-counts: 14,368	N/A	N/A	NLP feature extraction	Tweet text	LR, ADA Boost, XGBoost, RF	Supervised	95.55%	N/A	N/A	N/A
[48]	Cresci-2017	51,457 tweets	N/A	N/A	Eliminate redun-dancy items, duplicate and specific contrast patterns	Tweet metadata Tweet text	Contrast pattern-based classifier	Supervised	N/A	N/A	N/A	N/A
[49]	Own	4500 ac-counts 160,000 tweets	N/A	N/A	NLP	Tweet metadata Account meta-data	Label spread-ing and label propagation	Semi-supervised	91%	N/A	N/A	88%
[50]	Twitter Social Honey-pot	325 ac-counts 6500 tweets	N/A	N/A	Expectation Maxi-mization (EM) clustering	Tweet metadata Account meta-data	K-NN, DT, NB, LR, SVM, XGBoost	Supervised	91%	N/A	N/A	92%
[51]	Own	25,847 users 500K tweets	N/A	N/A	NLP	Tweet metadata Account meta-data	SVM, NN, and k-NN, DT, NB	Supervised	N/A	N/A	N/A	91.7%
[52]	2016 Presi-dential election	2.3 M	80%	20%	#ImWithHer or #Lock-HerUp	Opinions(DeGroot Model, tweets)	neural network	Unsupervised	92%	N/A	N/A	N/A

The results confirm that the refined modularity measure greatly enhances the detection accuracy.

Morstatter *et al.* [40] proposed a detection model that achieved the perfect tradeoff between recall and precision

TABLE 3. *(Continued.)* **Summary of shallow learning based detection methods.**

Ref.	Dataset	Tweets	Training Set	Testing Set	Pre-processing	Features	Classifier	Approach	Accuracy	FP	TP	Precision
[54]	Real-world Twitter datasets	591,768	147,387	N/A	Behaviors detection	Accounts, tweets, follower, friends	GenBot	Unsupervised	N/A	N/A	N/A	N/A
[55]	Brexit dataset Brexit Remain,and Pro-NS2	51 million	N/A	N/A	right-leanings and left-leaning	Kolmogorov-Smirnov goodness-of-fit	N/A	Supervised	N/A	N/A	N/A	N/A
[56]	Own database	50,000	N/A	N/A	news, best, awesome	bit.ly, ift.tt, ow.ly, goo.gl, tinyurl.com, dlvr.it, dld.bz, viid.me and ln.is.	Clustering algorithm	Unsupervised	N/A	N/A	N/A	N/A
[57]	Star Wars bots	2.8 million	N/A	N/A	N/A	Detection of Star Wars botnet	Naive Bayes	Supervised	N/A	N/A	N/A	N/A
[58]	Real Users' data, Star Wars bots	350,000	N/A	N/A	N/A	Star Wars detection	Naive Bayes	Supervised	N/A	N/A	N/A	>99%
[59]	Twitter	40 million	N/A	N/A	Human, cyborg, bot	Entropy, spam detection, and decision maker	Naive Bayes	Supervised	96%	N/A	N/A	55%
[34]	U.S. presidential debate, Brexit, Gilets Jaunes	1.3M	80%	20%	"pizzagate", blm" and "black-livesmatter"	Impact of user opinion	Deep neural network	Unsupervised	83%	N/A	N/A	N/A
[60]	Sina Weibo	24,705	864	1592	N/A	Active learning	Class-biased multi-label emotion classification	Supervised	N/A	N/A	N/A	N/A
[61]	Yelp	100	90	10	User ID, User name, Friend count, Follower Count, Favorite count, Account age, User location	User Behavioural Profile	k-Nearest Neighbour, RF, ensemble	Supervised	95%	N/A	N/A	N/A
[62]	Fake Project	563,693	N/A	N/A	Trustworthy path identification	Direct trust and indirect trust	Bayesian theory, Dempster-Shafer	Supervised	85%	N/A	N/A	N/A
[63]	Twitter	N/A	N/A	N/A	N/A	community-based features with other feature categories, including metadata, content, and interaction	Random Forest, Decision Tree Bayesian Network	Supervised	N/A	N/A	N/A	N/A
[64]	N/A	60,572	75%	25%	credibility analysis	hashtags #cyclonegaja, TNSDMA, #SaveDelta	NB, SVM, and RF algorithms	Supervised	87%	N/A	N/A	N/A

in classifying tweets as bots. The authors assumed that bots could not generate original content but rather relied on retweeting and reposting topics. Therefore, they focused on the heuristic features such as the ratio of retweets, the average length of tweets, and the average time between tweets. To improve the performance, the tweet content was employed

TABLE 3. *(Continued.)* **Summary of shallow learning based detection methods.**

Ref.	Dataset	Tweets	Training Set	Testing Set	Pre-processing	Features	Classifier	Approach	Accuracy	FP	TP	Precision
[65]	Twitter	3.5 million	274,820	31,171	Emotions, sentiment	Cox regression technique (eemotionnegative, positive))	Zero-truncated negative binomial (ZTNB) regression	Supervised	N/A	N/A	N/A	N/A
[66]	Twitter	77,033	35,343	15,147	Spam detection	The collector service, database, and analyzer service	Naive Bayes	Supervised	94.3%	N/A	N/A	N/A
[67]	Twitter	2,031,081	227,196	10,280	Spam detection	Follower ratio, Account reputation, Hashtag position aggregate, URL ratio	Random forest	Unsupervised	96.30%	4.3%	96.2%	96.2%
[68]	Twitter real dataset	600 million	N/A	N/A	Class imbalance of spam detection	account_age, no_follower, no_following, no_userfavourites, no_lists, no_tweets, no_hashtag, no_retweets, no_urls, no_char, no_digits	heterogeneous stacking-based ensemble learning	Supervised	N/A	0.03	0.70	70%
[37]	Kaggle	N/A	70%	30%	N/A	name, status, description, follower count, listed count, screen name, verification status, and average semantic score	Deep Neural Network	Unsupervised	67.7%	N/A	N/A	85%
[69]	Koobface, Twitterbot, TWebot, Yazanbot, Fix-Nazbot, Wbbot, Fbbot	4.2M	N/A	N/A	infection, predefined host behaviors, Command and Control	host behavior monitoring and analysis and identification	Behavior detection	Unsupervised	N/A	29.6%	4.5%	N/A

in the form of tweet topic features. It was extracted from the tweet text and fed to Latent Dirichlet Allocation (LDA) to reduce the dimensionality. The authors assumed that bots were not created to have various interests in several topics. However, they were focused on a certain topic they were made for. Both topic and heuristic features were fed to several classifiers to form a boosting detector. Two boosting techniques were implemented: Adaboost and the novel Boosting through Optimizing Recall (BoostOR). The latter focuses on the tradeoff between recall and precision through bot weight-updates rather than misclassified updates. This ensures better recall performance since only misclassified bots are reentered with higher weights. In order to test and evaluate both models, two datasets were collected. The first was the Arab Spring activity in Libya, which contained 94,535 users with only 7% bot accounts that were labeled using a simple technique of recrawling and checking if they still exist after a while. The second was Arabic Honeypot Dataset, which was labeled using honeypot and contained 6,285 users with 49% bot accounts. The results were achieved with 75% precision for both models on the Libya dataset, and 79% and 71% precision for Adaboost and BoostOR, respectively. However, in terms of a better tradeoff, BoostOR is favored with a higher F-measure score for both datasets.

Lundberg *et al.* [41] presented a Twitter bot detection method that utilizes the tweet metadata and it is, therefore, language-independent. Tweets were collected using Twitter API to form 5,000 multilingual tweets from three languages: English, Swedish, and Finish. However, after the first experiment, the English tweets were dropped due to their apparent easier classification. The method trained several decision tree classifiers over 4,000 tweets, but the J48-HP classifier outperformed all with over 99% precision for 1000 tweets. Haidermota *et al.* [42] experimented with classical classifiers

such as Logistic Regression, Naïve Bayes, and Random Forest to detect bot users from the Twitter dataset. The features extracted contained both account and tweet metadata such as URL ratio within a tweet, entropy tweeting level, etc. All classifiers but Random forest performed poorly. Therefore, they were implemented with the Adaboost method, which improved all detectors' overall performance. Still, the Random Forest must take the lead with 95% precision for the test data. Kiran *et al.* [43] proposed to use an artificial neural network to detect Twitter bots from their account and tweet metadata. The features used were Friendship Ratio, Favorite Ratio, and Tweet Ratio collected from 231 users. The dataset was divided into 70% training and 30% testing to evaluate the ANN model. The results achieved were considerably high with accuracy in the range of 94.4% to 95.80%.

Alarifi *et al.* [44] gathered and labeled a dataset containing human, bot, and hybrid accounts (both human and bots post tweets). They used four supervised machine learning techniques: SVM, Bayesian network, decision tree, and multi-layer artificial neural network. Authors used tweet features that considered the quantity of links, hashtags, mentions, or characters per tweet. Besides, they used novel features for the detection process, such as the number of pictures and Tweeting two or more tweets simultaneously. The paper addressed two training scenarios, a two-class classification, which considers human and bot accounts, and a three-class classification, which considers hybrid accounts in addition to human and bot accounts. At both classification scenarios, the Bayesian network and random forest outperformed other classifiers with an accuracy of 90.90% and 91.39%, respectively. However, all other classifiers performed better in detecting bots and humans but faced difficulty in detecting hybrid accounts. Moreover, the authors built a browser plug-in called Twitter Sybils Detector (TSD), which notifies the user before accessing Sybil accounts. Ji *et al.* [69] suggested a new approach to identify social bot activity. They developed a behavioral tree-based solution for social bot identification. They paired it with the prototype repository for producing identification results after building the suspicious behavior tree. To assess the results, they gathered actual social botnet traces. The findings revealed a 29.6% false positives and a 4.5% false-negative rate.

Another study in which a Naïve Bayesian classifier achieved high accuracy (90.9%) was conducted by Ersahin *et al.* [46]. A Supervised discretization technique, namely Entropy Minimization Discretization (MDE), was applied to numerical data. This study considered some derived features by taking the average of hashtags, mentions, and URLs in the last 20 tweets. Another study was performed by Wang [51] using supervised learning. Wang used support vector machines, neural network, k-nearest neighbors, decision tree, and naïve Bayes. The author collected a dataset containing 25,847 user accounts and around 500K tweets. Graph-based and content-based features were extracted (only three from each). Graph-based features were a count of followers, count of friends, and follower's ratio. Content-based features included duplicate tweets, count of links, and count of mentions/replies. Experimental evaluation showed that Naïve Bayes had the best performance with 91.7% Precision, Recall, and F-measure. A study conducted by Alom *et al.* [50] utilized seven different supervised techniques to detect spam users, namely: Decision Tree, Support Vector Machine, Random Forest, k-Nearest Neighbor, Naïve Bayes, Logistic Regression, and eXtreme Gradient Boosting (XGBoost). The authors used the Twitter Social Honeypot dataset. They proposed a set of features of two types: graph-based features and content-based features.

Graph-based features can be obtained as the social graphs that Twitter allows users to build. These graphs represent the relationships between users. For example, in triangle count of the user network, the triangle represents the adjacency between nodes; a high number of triangles means the user is genuine. Content-based features are derived from tweet text. For example, in Unique URL ratio, spammers try to spread a malicious URL for a specific site as much as possible to increase the probability that a legitimate user visits this site. The more uniqueness (high rate URL ratio) means the user is legitimate. Authors compared the proposed features with other studies and found that it is effective for spam user detection on Twitter. Random Forest algorithm outperformed other classifiers with an accuracy of 91%. Most researchers use profile-based features to detect malicious accounts. On the other hand, Pakaya *et al.* [47] argued that the way malicious and real users post tweets can be distinguished. Therefore, the authors developed a classification model based on account tweets only. The models used were Logistic Regression, ADA Boost, XGBoost, and Random Forest. The proposed approach was to concatenate the tweets to be in a single document. The research conducted two scenarios: first, map all tweets to the account by filling a long document with a single account. The second is the same, however, it only considered the first 100 tweets and does not include retweets to preserve the classifier from length bias. The following NLP feature extraction methods were used: tf-idf, bigram, and Word2Vec. Moreover, they built a multi-class model based on tweet features to detect the different types of malicious accounts (spambots and fake followers). Results showed that the best algorithm was the XGBoost, as it performed better in the detection between malicious and human accounts with an accuracy of 95.55% using tf-idf features. Also, XGBoost performed better in multi-class classification using Word2Vec features with an accuracy of 95.2%. There are different understandable classifiers used to solve other world issues [71]–[73]. Among them all, contrast pattern-based classification is outstanding; this is because it produces a demonstration that is easily understandable by experts. Moreover, it provides more accurate results than the popular supervised learning techniques [74]. El-Mawass *et al.* [75] studied how to utilize the previous controlled classification systems' success to identify spammers. In a probabilistic graphic model framework, they proposed classifier output as a prior belief. This

method makes spreading views easy to like social accounts. Using a map of similar people, Markov Random Field used several state-of-the-art classifiers to compute past beliefs. Loopy Anticipation Propagation was also used for subsequent market predictions. A new manually named Twitter dataset analyzed the system.

Despite the fact that several models were built for detecting bot behavior, most of the suggested classifiers were not understandable. Nowadays, researchers aim not just to achieve high accuracy but also to understand the model and results of the classifier by area experts. Therefore, Loyola-González *et al.* [48] utilized a contrast pattern-based classifier for the bot detection problem in Twitter. Furthermore, they proposed a new feature model that integrates tweet content sentiment analysis and tweet account usage. Results showed that the proposed model improved the performance in different classifiers. The correlation analysis showed that sentiment analysis is appropriate and has a good impact on bot detection. Rodríguez-Ruiz *et al.* [45] conducted a bot detection model using one-class classification approach. The authors used part of Cresci-2017 datasets. One-class classification is a part of supervised learning techniques in which the training dataset consists of instances of one class only. The classifiers used were: Bagging-RandomMiner, Bagging-TPMiner, One-Class K-means with Randomly projected features Algorithm (OCKRA), and one-class versions of Naïve Bayesian and Support Vector Machine. They first measured the performance of some binary classifiers such as Adaboost, SVM, and Naïve Bayes. Then, they classified the same dataset with a one-class classifier. The model achieved AUC above 89%.

Wang [29] suggested a system for solving spam tweets on Twitter, where users and tweets, along with the message content, identify the tweets. The author used three separate machine learning algorithms for assessment: the support vector machine, neural network, and random forest. The Naive Bayes classification provided approximately 80 percent accuracy. Devi and Karthika [64] focused on the authenticity of natural disaster-related telephone tweets and recognized tweets distributing false data. They suggested an evaluation of their legitimacy to identify tweets according to their credibility. They contrasted tweets' success with sophisticated machine-learning algorithms such as SVM, Naive Bayes, and Random Forest classification. On the real-time dataset obtained during the Gaja cyclone occurrence, the experimentation was carried out. The Random Forest classification was provided with 87% accuracy as the most appropriate algorithm for credibility evaluation.

El Hjouji *et al.* [52] gave a description of the impact of artificial users, or bots, on a social network. They modeled the opinions using a variant of the best known DeGroot model, which connects beliefs with the network's structure. They saw a strong link between viewpoints centered on this network model and Twitter users' tweets discussing the 2016 US presidential election. Using a bot detection algorithm, they identified bot accounts that contained less than one percent of the network and figured out that the bots supporting Donald Trump were twice the number of bots that supported Hillary Clinton. Bots were removed from the network and the opinions utilizing the network construct were recalculated. Bot activity analysis indicates that bots create a huge shift in the opinions. The discrepancy in opinion change is because the Clinton bots suggest that a small number of strongly active bots in a social network may have a disproportionate effect on beliefs. To characterize Twitter bot accounts and assess their prominence in the current online debate, Efthimion *et al.* [76] suggested numerous bot identification algorithms. A machine learning algorithm using a variety of bot detection features was introduced. With as low as a 2.25 percent misclassification score, it is efficient at identifying bots.

Twitter is one of the most popular entertainment and news updating source. However, owing to its 280-character cap and automatic shortening of URLs, computer attackers are constantly targeting for drive-by-download assaults where a user's system is compromised by visiting a web page [65]. Cybercriminals utilize regular processes to recruit large quantities of people to hack and distribute malware using common hashtags to generate misleading messages to attract fraudulent websites. Javed *et al.* [65] discussed a drive-by-download attack that was conducted in an appealing tweet and used as a clickbait to draw traffic to a tricky website. Two questions in this article were raised: ''Why are any malignant tweets retweeted?'' and ''Is empathy a viral tweet drive?'' They tweeted about seven separate athletic competitions over three years and found drive-by-download attack tweets. They generated data samples from existing malicious (N = 105, 642) and normal (N = 169, 178) to forecast survival information flow. They defined the size as the sum of retweets and survival as the tweet duration in the study window. They focused on measuring their dependent size measure and findings relative to other predictive models and picked the Zero-Truncated Negative Binomial (ZTNB) regression. To model information flows' survival, they used Cox regression to estimate relative danger rates for independent steps. The results show that variables for aggressive and peaceful tweets are statistically important for information flow size and durability. Healthy emotions and optimistic tweets predict benign dataset scale and longevity. Additionally, data flows for malicious data samples associate negative emotions, particularly fear.

A complementary unlabeled resource sampling strategy is suggested by Kang *et al.* [60], measuring a probabilistic gap between a transient corpus expected distribution of the emotion mark and a standardized distribution. Unplanned examples in which model ambiguity for multi-label emotional projections, syntactic representativeness for the other unlabeled examples, and diversity for a high-quality sampling of the labeled examples are tested and often presented with quality evaluations. Combining community-driven features with other feature categories, including text, metadata, interaction-based features, and automated spammers, Fazil and Abulaish [63] presented a hybrid approach. Based on their interactions with their followers, they characterized

users to prevent features associated with their activity. However, it is difficult to avoid followers-based features. A real dataset of benign users and spammers was used, including six newly-defined features and two redefined features, and three classifiers were identified: decision trees, random forest, and Bayesian networks, for learning. Using Twitter, Al-Dayil and Dahshan [77] suggested an identification strategy for social networking related mobile botnets, to identify tweets induced by bots and distinguish those against tweets created by users or by user-approved applications. The proposed approach incorporated the connection between tweeting and user behavior, such as clicking, and an Artificial Immune System (AIS) tracker. The tracker generates a tweet signature and checks it with a dynamically modified bot activity signature from a signatures library. The test results indicated a 95% precision detection rate in identifying bot tweets.

To deal with the problem of class imbalance of spam detection in social networks, Zhao *et al.* [68] proposed a heterogeneous stacking-based ensemble learning framework, which consists of two main modules: a base module and a combined module. In the base module, they trained six separate base classifiers to generate meta-data with new features, which are fed to the combined module. In this module, they incorporated cost-effective deep neural network to train a meta-classifier. The evaluation results show that the framework succeeds in improving the spam detection rate on imbalanced datasets. Heredia [78] applied machine learning techniques, especially ensemble methods and feature selection, to analyze the content of social media and detect the malicious users. They studied the influence of social bots on public opinion, and the accuracy of using Twitter as a polling source. To this end, they investigated the effectiveness of Twitter to predict the 2016 presidential election in the United States using social bots. The outcomes of the 2016 US Campaign proclaimed Trump the victor with 306 voters (56.88%) in 30 states, leaving Clinton with 232 voters (43.12%) all over the 20 states and the District of Columbia (DC). Trump had 139 (47.93%) voters, and Clinton had 151 (52.07%) voters while restricting voters to the 21 chosen states. Savyan and Bhanu [61] proposed a framework named User Behavior Analytics based Compromised Account Detection (UbCadet). The profile of a Twitter user is built using similarity of twitter text, similarity of hashtag, tweeting period, and geo-location of the user. Based on this profile, the tweet patterns of each user are computed and are fed to the K-Nearest Neighbor learning algorithm to be classified as normal or malicious. Evaluation results show that UbCadet is able to detect more than 90% of malicious tweets.

2) SEMI-SUPERVISED LEARNING TECHNIQUES

Semi-supervised learning techniques are useful in determining patterns in a large amount of data, in platforms such as social networks, in which labeling is a costly and time-consuming task. Alharthi *et al.* [49] developed a semi-supervised technique to label their dataset and classify Twitter accounts into spam or genuine. They targeted Arab spammers' accounts and figured out if they behave like Botnet or show a software behavior. They applied Label propagation and Labeled spreading algorithms and had a good performance with an accuracy of 91%. To differentiate Pathogenic Social Media (PSM) from regular users during a brief span of the actions, Shaabani *et al.* [33] used a causative attribution system coupled with graph-based indicators. The findings on a true Twitter dataset accentuated the value of the suggested approaches. The proposed solution increased the F1 score by 0.28 relative to current methods with 0.90 precision and an F1 score of 0.63.

Kabakus and Kara [66] discussed that spammers and legitimate users are aware of the popularity and advanced APIs offered by Twitter for programming reading and writing of Twitter information. As Twitter has many special features, it is not easy to use standard spam detection tools specifically on Twitter to identify spam. Therefore, the paper proposes a mechanism for spam detection, which is explicitly developed for Twitter, namely TwitterSpamDetector. To identify spam on Twitter, TwitterSpamDetector uses Twitter-specific characteristics. 77,033 messages have been shared on Twitter's API by 50,490 users. Naive Bayes uses the special features of Twitter that specifically classify the spammers of legitimate users for TwitterSpamDetector preparation. According to the assessment findings, the precision and sensitivity of the TwitterSpamDetector were computed as 0.943 or 0.913.

Kouvela *et al.* [79] proposed Bot-Detective, a web service that considers both successful detection of bot users and data interpretability. They proposed a new explanatory bot identification solution that is a comprehensible, transparent, and AI-driven bot recognition model on Twitter. They implemented a freely available site recognition tool that provides an expandable ML framework and user feedback functionality within a powerful crowd-sourcing process. Through exploiting Twitter's rules and existing tools, a freshly created annotated dataset was developed and used in the proposed service. The lack of facts is one of Twitter's greatest practical hurdles. Echeverria and Zhou [58] noticed that the Star Wars botnet totaled over 350k bots. These bots have several unique features, revealing deep vulnerabilities of existing bot detection approaches. Their research has critical implications in cryptography, not just because the botnet's size is broader than any previously studied botnet, but also because it has been hidden since its creation in 2013. They argued that more analysis was needed to better analyze potential security risks to the Twitter community that could pose a huge, covert botnet attack.

To identify social botnets in a Twitter-like SNS, Dorri *et al.* [80] resolved these limitations by introducing SocialBotHunter. This semi-supervised collaborative recognition strategy integrates a social graph's systemic details with the knowledge of users' social actions in a unified way. The findings revealed that SocialBotHunter could reliably detect the social bots involved. SocialBotHunter calculates an initial anomaly score for a user by producing a feature vector for every user on the basis of their social behavior.

A random binary variable is then associated with each user and the social interactions are modeled among all the users as a pairwise Markov random field (MRF) defining a joint probability distribution. Then, belief propagation is applied to MRF for revising anomaly scores so that social botnets are detected.

Integrating a trust model (consisting of two criteria, such as direct trust and indirect trust) suggests a social botnet identification algorithm to define a dedicated route in the social networking site (like Twitter). Besides, using Bayesian philosophy, Lingam *et al.* [62] used the confidence value of direct relationships between respondents (i.e., direct trust) to calculate trust. The trust value of adjacent respondents (i.e., indirect trust) is determined using the Dempster-Shafer theory. Trust performance was enhanced by combining these two criteria for identifying social bots from respondents. To demonstrate their social botnet identification algorithm's effectiveness, tests were carried out using The Fake Project dataset (collected from Twitter). Echeverria *et al.* [57] observed the Star Wars bots' odd actions that the bots were generated in bursts and only tweeted in the first few minutes after creation. They discovered a bigger Twitter botnet with over 500,000 bots, the Bursty botnet. The research revealed that the Bursty botnet was explicitly responsible for a significant interactive spamming attack in 2012. Most bot detection techniques focused on the ''common'' attribute supported among all bots. However, their uncovered botnets do not exhibit all of these qualities; instead, they were recognized by previously unknown different, peculiar tweeting habits.

Chu *et al.* [59] focused on the individual, bot, and cyborg accounts grouping on Twitter. For over 500,000 profiles, they carried out a series of significant calculations. The variations in messaging conduct, tweet information, and profile attributes were found among humans, bot, and cyborg accounts. Relying on the calculation effects, they suggested a classification scheme that consists of entropy constituent, spam identification element, profile attributes, and decision-maker. It uses the set of attributes from an anonymous consumer to decide if a person is a bot, a human, or a cyborg. The work proposed by Elhadad *et al.* [81] intends to support the continuing research initiatives to tackle COVID-19 awareness. A set of COVID-19 Twitter (COVID-19-FAKES) data from February to March 2020 is used, which is automatically labeled for bilingual (Arabic/English) data. They used the information shared on the official UNICEF, WHO, and UN web pages, official Twitter accounts, and pre-checked facts from COVID-19 web pages as a basis of reliable knowledge to create the ground database. Then, in a COVID-19-FAKES dataset (i.e., tweets), thirteen distinct machine learning algorithms are annotated using seven different extraction methods.

3) UNSUPERVISED LEARNING TECHNIQUES

Most of the bot detection approaches for Twitter accounts depend on supervised techniques. However, some works apply the unsupervised approach and are tested on labeled datasets to measure the performance. Miller *et al.* [82] used two modified unsupervised stream clustering algorithms, namely StreamKM++ and DenStream, which cluster normal Twitter accounts and consider the outliers as spam accounts. Both StreamKM++ and DenStream achieve 99% recall. When the two algorithms are combined, 100% recall and 2.2% false positive rate are achieved. Based on the similarities between spam accounts, Adewole *et al.* [67] proposed a new approach to detect Twitter spammers. To this end, they applied Principal Component Analysis (PCA) and tuned K-means on 200,000 accounts, randomly selected from more than 2 million tweets to identify spammer clusters. The generated clusters are used as a ground-truth to train three classifiers: Random Forest, MLP, and SVM, which reached an accuracy of 96.30%, 96.00%, and 95.60% respectively.

Farkas and Bastos [53] analyzed the content of fake accounts posted by the Internet Research Agency (IRA). They used a database of 4,539 tweets that are posted between 2012 and 2017, in addition to manually coded 2,501 tweets. Using 19 control variables, tweets were annotated to examine whether IRA operations fit classic propaganda models. The findings reveal that some user accounts are configured to perform specific tasks with focus on daily US news and distributing controversial news about different national matters. Shevtsov *et al.* [83] analyzed the behavior of Greek-speaking Twitter accounts over 36 months. They applied Concurrent Content Injection Detection (CCID) [84] to identify botnets, i.e., accounts posting almost similar tweets nearly simultaneously. They discovered that 1,850 accounts exhibit this behavioral pattern, which means that they are controlled by the same software.

The 'Star Wars' botnet on Twitter, which comprises more than 350,000 bots that tweet random quotes solely from Star Wars novels, was extracted, revealed, and examined in [85]. The botnet has one form of bot that shows precisely having similar characteristics all across the botnet.

In order for a propaganda campaign to succeed, Smith [86] revealed that customized information boosted an anticipated emotional response from a target audience. They examined automated software program usages, such as sock puppets, bots, and cyborgs. During the election, bots may generate popularity for fake news tweets commenting on political topics. They discussed that fake news topics improve the likelihood that news media reported these subjects. To explain how the contents of social botnet tweets vary from normal users in a single dataset, Abokhodair *et al.* [87] considered one Twitter social botnet to understand how social botnets influence the discussions. They analyzed almost 3,000 English and Arabic tweets that were posted by the Syrian social bot through 35 weeks before its shutdown. They found out that the development, actions, and content of this botnet do not conform to the standard concepts of botnets.

Pantic and Husain [88] proposed a system that uses Twitter as Covert botnet command and control. The system-generates plausible cover messages based on a tweet's required length, determined by an encoding map that utilizes a secret message

structure. The encoding is built based on input symbol frequencies and posting frequencies. They also proposed a technique to create Twitter account names based on Markov chains. If the current botmaster account is not reachable, the bots would be connected to new accounts. The experiments are conducted using 7.3M tweets from 3.7K validated accounts. The efficacy and usability of the system are evaluated using Emulab and Amazon's Mechanical Turk, and the obtained results are promising. Yang *et al.* [89] investigated the features that are required to detect Twitter spammers that employ evasion tactics. They evaluated the robustness of 24 detection features that are already used in the literature along with the proposed ones. After analyzing more than 14 million tweets and 500,000 Twitter accounts, high detection rate (i.e., 0.85), and low false positive rate (i.e., 0.01) are achieved.

IV. DISCUSSION AND OPEN ISSUES

Although many solutions have been proposed in the literature to detect tweet-based bots, some challenges still need further investigation. In this section, we discuss the main open issues related to tweet-based bot detection.

- *Feature selection:* Generally, the quality of the classifiers relies on the quality of selected features. Twitter bot detection problem is not an exception to this rule. As proven by the previous work, there is no standard set of features that can guarantee good performance, but rather each literature study introduced some set of features that were believed to be the ideal for their chosen classifier. For example, in [41], [43], the authors employed the tweet's metadata like the retweet count or favorite count to produce lightweight classifiers that avoid overfitting. However, the choice of features to include is crucial, as pointed out by [24], in which account metadata was combined with tweets to produce a high accuracy of 96%. In contrast, both [26] and [47] opted to utilize one feature only. The method in [26] used the timestamp of retweets. It analyzed the behavior of the bot by monitoring the retweeting pattern, which yielded good results since the bot (including social bots) has a hazy concept of time. Similarly, in [28], the pattern of tweeting was monitored. However, it was combined with more features to improve the performance. Deep learning methods do not require manually selected features as it can be done automatically, which is their main advantage. It allows developers to work with datasets without prior knowledge, though providing the features will certainly improve the performance, as shown in [24].
- *Dataset labeling:* One of the challenges of tweet-based bot detection lies in the data labeling techniques. These techniques differ, hence their results also differ. This issue is mainly due to human annotators' remaining popularity, where they are prone to error, especially in detecting social bots. However, different techniques were employed for labeling, most common being a labeling tool such as Botmeter or Honeypot labeler, as used in [40]. Another simple and less effective technique was proposed by the same authors in [82] in which recrawling the users is performed after a while along with trusting the Twitter bot detector to suspend the accounts. Therefore, accounts that no longer exist are perceived as bots. However, this is avoided in unsupervised approaches [26], [82] in which only the ground truth are labeled by humans, and due to its small size, the labeling quality can be controlled.
- *Dataset balancing:* Human accounts are the majority of bot datsets as in [40], [90], or they represent an acceptable ratio as in [39]. To capture the true performance of the detection, precision and recall were utilized in place of accuracy. However, some literature [24] decided to handle it before detecting and producing synthetic data. Synthetic data are usually avoided as it can cause the model to deduce and learn unrepresentative features. Therefore, the literature mentioned various techniques like SMOTE to produce synthetic data close to existing bot characteristics neatly. This resulted in balanced data and better performance.
- *False positives:* Generally, most mentioned literature has a good performance. However, there is still an issue with misclassifying human accounts (i.e., incurring false positives), which do not improve the recall performance for bot detection. This could cause problems if adopted by platforms such as Twitter since harmless human users are suspended.
- *Detection evasion:* Some bot accounts have little or no information related to botnet campaigns, and hence they can evade detection [91], especially if they are inactive. Although each account has inherently related details on each social networking platform, it is very simple to join other spammers, i.e., those who only consume content but do not interact. The only answer to this issue is to implement real-time detection. In some spam campaigns, the tweets are changing over time in order to evade detection. This issue is called the spam drift. If the detection techniques are not trained with the updated tweets, they will incur poor detection performance. Although some works [59], [82], [92], [93] dealt with the spam drift issue, this is still one of the open challenges for the research community.
- *Streaming big data analysis:* Twitter data stream is an example of big data. Hence, machine learning techniques have to deal with real-time data and the challenging characteristics of big data, i.e., high volume, high velocity, and high volatility. Most of the research works focus on offline detection, and hence the machine learning detection techniques should be designed to be online and scalable, in order to deal with continuous big Twitter data stream.
- *Language diversity support*: The majority of detection tools and algorithms are designed to deal with the English language, and there are few research works

that consider other languages. As each language has its own syntactic and semantic features, it is important to develop tools and algorithms or extend the existing ones to consider language diversity. Another approach that could be further investigated is to use language-agnostic features [39] in tweet-based bot detection.

- *Botnet infrastructure detection:* The issue about using online platforms, such as social networks, as botnet infrastructures for Command and Control (C&C) is also important [94]. Treating the current versions of social media platforms as a starting point, the researchers foresaw developing C&C approaches and discussed countermeasures based on social networks. If the C&C is taken down, the bots will be inactive. Hence, instead of analyzing a large set of Twitter accounts to identify whether they are humans or bots, the research efforts could go towards detecting and taking down the botnet infrastructure, which can be achieved by identifying the C&C servers, DNS servers, and IP addresses that are used in building this infrastructure.

V. CONCLUSION

Twitter is one of the most popular social media platforms that allows connecting people and helps organizations reaching out to customers. Tweet-based botnet can compromise Twitter and create malicious accounts to launch large-scale attacks and manipulation campaigns. In this review, we have focused on big data analytics, especially shallow and deep learning to fight against tweet-based botnets, and to accurately distinguish between human accounts and tweet-based bot accounts. We have discussed related surveys, and have also provided a taxonomy that classifies the state-of-the-art tweet-based bot detection techniques up to 2020. In addition, the shallow and deep learning techniques are described for tweet-based bot detection, along with their performance results. Finally, we presented and discussed the open issues and future research challenges.

ACKNOWLEDGMENT

The authors extend their sincere appreciation to the Deanship of Scientific Research at King Saud University, Saudi Arabia, for funding this work through the Research Group No. RGP-214.

REFERENCES

[1] M. Mohsin. (2020). *10 Social Media Statistics You Need to Know in 2021*. [Online]. Available: https://www.oberlo.com/blog/social-media-marketing-statistics

[2] I. Arghire. (2020). *Twitter Hack: 24 Hours From Phishing Employees to Hijacking Accounts*. https://www.securityweek.com/twitter-hack-24-hours-phishing-employees-hijacking-accounts

[3] *The Rise of Social Media Botnets*. Accessed: Feb. 21, 2021. [Online]. Available: https://www.darkreading.com/attacks-breaches/the-rise-of-social-media-botnets/a/d-id/1321177

[4] M. Imran, M. H. Durad, F. A. Khan, and A. Derhab, "Toward an optimal solution against denial of service attacks in software defined networks," *Future Gener. Comput. Syst.*, vol. 92, pp. 444–453, Mar. 2019.

[5] M. S. Savell. (2018). *Protect Your Company's Reputation From Threats by Social Bots*. [Online]. Available: https://zignallabs.com/blog/protect-your-companys-reputation-from-threats-by-social-bots/

[6] S. Aslam. (2021). *Twitter by the Numbers: Stats, Demographics & Fun Facts*. [Online]. Available: https://www.omnicoreagency.com/twitter-statistics/

[7] A. Aldweesh, A. Derhab, and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowl.-Based Syst.*, vol. 189, Feb. 2020, Art. no. 105124.

[8] S. Mahdavifar and A. A. Ghorbani, "Application of deep learning to cybersecurity: A survey," *Neurocomputing*, vol. 347, pp. 149–176, Jun. 2019.

[9] E. B. Karbab, M. Debbabi, A. Derhab, and D. Mouheb, "MalDozer: Automatic framework for Android malware detection using deep learning," *Digit. Invest.*, vol. 24, pp. S48–S59, Mar. 2018.

[10] F. A. Khan, A. Gumaei, A. Derhab, and A. Hussain, "A novel two-stage deep learning model for efficient network intrusion detection," *IEEE Access*, vol. 7, pp. 30373–30385, 2019.

[11] A. Derhab, A. Aldweesh, A. Z. Emam, and F. A. Khan, "Intrusion detection system for Internet of Things based on temporal convolution neural network and efficient feature engineering," *Wireless Commun. Mobile Comput.*, vol. 2020, pp. 1–16, Dec. 2020.

[12] B. Marr. (2020). *How Twitter Uses Big Data and Artificial Intelligence (AI)*. [Online]. Available: https://www.bernardmarr.com/default.asp?contentID=1373

[13] A. T. Kabakus and R. Kara, "A survey of spam detection methods on Twitter," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 3, pp. 29–38, 2017.

[14] M. Chakraborty, S. Pal, R. Pramanik, and C. R. Chowdary, "Recent developments in social spam detection and combating techniques: A survey," *Inf. Process. Manage.*, vol. 52, no. 6, pp. 1053–1073, Nov. 2016.

[15] E. Alothali, N. Zaki, E. A. Mohamed, and H. Alashwal, "Detecting social bots on Twitter: A literature review," in *Proc. Int. Conf. Innov. Inf. Technol. (IIT)*, Nov. 2018, pp. 175–180.

[16] M. Latah, "Detection of malicious social bots: A survey and a refined taxonomy," *Expert Syst. Appl.*, vol. 151, Aug. 2020, Art. no. 113383.

[17] K. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots," *Hum. Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 48–61, Jan. 2019.

[18] Z. Guo, J.-H. Cho, I.-R. Chen, S. Sengupta, M. Hong, and T. Mitra, "Online social deception and its countermeasures: A survey," *IEEE Access*, vol. 9, pp. 1770–1806, 2021.

[19] S. B. Abkenar, M. H. Kashani, M. Akbari, and E. Mahdipour, "Twitter spam detection: A systematic review," 2020, *arXiv:2011.14754*. [Online]. Available: http://arxiv.org/abs/2011.14754

[20] W. Daffa, O. Bamasag, and A. AlMansour, "A survey on spam URLs detection in Twitter," in *Proc. 1st Int. Conf. Comput. Appl. Inf. Secur. (ICCAIS)*, Apr. 2018, pp. 1–6.

[21] C. Besel, J. Echeverria, and S. Zhou, "Full cycle analysis of a large-scale botnet attack on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 170–177.

[22] S. C. Woolley and P. N. Howard, *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford, U.K.: Oxford Univ. Press, 2018.

[23] *Data Dictionary: Standard V1.1*. Accessed: Feb. 21, 2021. [Online]. Available: https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet

[24] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Inf. Sci.*, vol. 467, pp. 312–322, Oct. 2018.

[25] F. Wei and U. T. Nguyen, "Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings," in *Proc. 1st IEEE Int. Conf. Trust, Privacy Secur. Intell. Syst. Appl. (TPS-ISA)*, Dec. 2019, pp. 101–109.

[26] M. Mazza, S. Cresci, M. Avvenuti, W. Quattrociocchi, and M. Tesconi, "RTbust: Exploiting temporal patterns for botnet detection on Twitter," in *Proc. 10th ACM Conf. Web Sci. (WebSci)*, 2019, pp. 183–192.

[27] C. Cai, L. Li, and D. Zengi, "Behavior enhanced deep bot detection in social media," in *Proc. IEEE Int. Conf. Intell. Secur. Informat. (ISI)*, Jul. 2017, pp. 128–130.

[28] M. Färber, A. Qurdina, and L. Ahmedi, "Identifying Twitter bots using a convolutional neural network," in *Proc. CLEF Working Notes*, 2019.

[29] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Proc. Int. Conf. Secur. Cryptogr. (SECRYPT)*, 2010, pp. 1–10.

[30] G. Lingam, R. R. Rout, and D. V. L. N. Somayajulu, "Adaptive deep Q-learning model for detecting social bots and influential users in online social networks," *Int. J. Speech Technol.*, vol. 49, no. 11, pp. 3947–3964, Nov. 2019.

[31] E. Nasca, "Text-based and graph-based analysis for fake news detection on social media," M.S. thesis, Scuola di Ingegneria Industriale e dell'Informazione, Politecnico di Milano, Milan, Italy, 2019. [Online]. Available: https://www.politesi.polimi.it/bitstream/10589/149908/3/thesis.pdf

[32] L. Luo, X. Zhang, X. Yang, and W. Yang, "Deepbot: A deep neural network based approach for detecting Twitter bots," in *Proc. IOP Conf., Mater. Sci. Eng.*, vol. 719, no. 1. Bristol, U.K.: IOP, 2020, Art. no. 012063.
[33] E. Shaabani, A. S. Mobarakeh, H. Alvari, and P. Shakarian, "An end-to-end framework to identify pathogenic social media accounts on Twitter," in *Proc. 2nd Int. Conf. Data Intell. Secur. (ICDIS)*, Jun. 2019, pp. 128–135.
[34] N. G. des Mesnards, D. S.Hunter, Z. El Hjouji, and T. Zaman, "Detecting bots and assessing their impact in social networks," 2018, *arXiv:1810.12398*. [Online]. Available: http://arxiv.org/abs/1810.12398
[35] M. T. S. M. Waskale and P. Jain, "Rumors detection on Twitter using machine learning techniques," *Int. J. Sci. Res. Eng. Trends*, vol. 5, no. 3, May/Jun. 2019. [Online]. Available: https://ijsret.com/wp-content/uploads/2019/05/IJSRET_V5_issue3_261.pdf
[36] G. Lingam, R. R. Rout, D. Somayajulu, and S. K. Das, "Social botnet community detection: A novel approach based on behavioral similarity in Twitter network using deep learning," in *Proc. 15th ACM Asia Conf. Comput. Commun. Secur.*, Oct. 2020, pp. 708–718.
[37] B. Shah, "Botnet account detection on Twitter using deep learning," Ph.D. dissertation, California State Univ., Sacramento, Sacramento, CA, USA, 2020.
[38] W. Chen, "Towards better prediction and content detection through online social media mining," Ph.D. dissertation, School Comput. Sci. Eng., Nanyang Technol. Univ., Singapore, 2018. [Online]. Available: https://dr.ntu.edu.sg/bitstream/10356/75925/1/thesis.pdf
[39] J. Knauth, "Language-agnostic Twitter-bot detection," in *Proc. Natural Lang. Process. Deep Learn. World*, Oct. 2019, pp. 550–558.
[40] 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 *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2016, pp. 533–540.
[41] J. Lundberg, J. Nordqvist, and M. Laitinen, "Towards a language independent Twitter bot detector," in *Proc. DHN*, 2019, pp. 308–319.
[42] M. Haidermota, A. Pansare, and D. Mitra, "Classifying Twitter user as a bot or not and comparing different classification algorithms," *Int. J. Adv. Res. Comput. Sci.*, vol. 9, no. 3, p. 29, 2018.
[43] K. Kiran, C. Manjunatha, T. S. Harini, P. D. Shenoy, and K. R. Venugopal, "Identification of anomalous users in Twitter based on user behaviour using artificial neural networks," in *Proc. IEEE 5th Int. Conf. Converg. Technol. (I2CT)*, Mar. 2019, pp. 1–5.
[44] A. Alarifi, M. Alsaleh, and A. Al-Salman, "Twitter turing test: Identifying social machines," *Inf. Sci.*, vol. 372, pp. 332–346, Dec. 2016.
[45] J. Rodríguez-Ruiz, J. I. Mata-Sánchez, R. Monroy, O. Loyola-González, and A. López-Cuevas, "A one-class classification approach for bot detection on Twitter," *Comput. Secur.*, vol. 91, Apr. 2020, Art. no. 101715.
[46] B. Erşahin, Ö. Aktaş, D. Kılınç, and C. Akyol, "Twitter fake account detection," in *Proc. Int. Conf. Comput. Sci. Eng. (UBMK)*, Oct. 2017, pp. 388–392.
[47] F. N. Pakaya, M. O. Ibrohim, and I. Budi, "Malicious account detection on Twitter based on tweet account features using machine learning," in *Proc. 4th Int. Conf. Informat. Comput. (ICIC)*, Oct. 2019, pp. 1–5.
[48] O. Loyola-González, R. Monroy, J. Rodríguez, A. López-Cuevas, and J. I. Mata-Sánchez, "Contrast pattern-based classification for bot detection on Twitter," *IEEE Access*, vol. 7, pp. 45800–45817, 2019.
[49] R. Alharthi, A. Alhothali, and K. Moria, "Detecting and characterizing arab spammers campaigns in Twitter," *Procedia Comput. Sci.*, vol. 163, pp. 248–256, Jan. 2019.
[50] Z. Alom, B. Carminati, and E. Ferrari, "Detecting spam accounts on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 1191–1198.
[51] A. H. Wang, "Detecting spam bots in online social networking sites: A machine learning approach," in *Proc. IFIP Annu. Conf. Data Appl. Secur. Privacy*. Berlin, Germany: Springer, 2010, pp. 335–342.
[52] N. G. des Mesnards, D. S. Hunter, Z. El Hjouji, and T. Zaman, "Detecting bots and assessing their impact in social networks," 2018, *arXiv:1810.12398*. [Online]. Available: http://arxiv.org/abs/1810.12398
[53] J. Farkas and M. Bastos, "IRA propaganda on Twitter: Stoking antagonism and tweeting local news," in *Proc. 9th Int. Conf. Social Media Soc.*, Jul. 2018, pp. 281–285.
[54] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Emergent properties, models, and laws of behavioral similarities within groups of Twitter users," *Comput. Commun.*, vol. 150, pp. 47–61, Jan. 2020.
[55] M. Ahmadi, "Hidden fear: Evaluating the effectiveness of messages on social media," M.S. thesis, Arizona State Univ., Tempe, AZ, USA, 2020. [Online]. Available: https://repository.asu.edu/attachments/227442/content/Ahmadi_asu_0010N_20017.pdf
[56] Z. Chen and D. Subramanian, "An unsupervised approach to detect spam campaigns that use botnets on Twitter," 2018, *arXiv:1804.05232*. [Online]. Available: http://arxiv.org/abs/1804.05232
[57] J. Echeverria, C. Besel, and S. Zhou, "Discovery of the twitter bursty botnet," in *Data Science for Cyber-Security*. Singapore: World Scientific, 2017. [Online]. Available: https://www.worldscientific.com/series/icpsst
[58] J. Echeverría and S. Zhou, "The 'star wars' botnet with >350k Twitter bots," 2017, *arXiv:1701.02405*. [Online]. Available: http://arxiv.org/abs/1701.02405
[59] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Detecting automation of Twitter accounts: Are you a human, bot, or cyborg?" *IEEE Trans. Dependable Secure Comput.*, vol. 9, no. 6, pp. 811–824, Nov. 2012.
[60] X. Kang, X. Shi, Y. Wu, and F. Ren, "Active learning with complementary sampling for instructing class-biased multi-label text emotion classification," *IEEE Trans. Affect. Comput.*, early access, Nov. 16, 2020, doi: 10.1109/TAFFC.2020.3038401.
[61] P. Savyan and S. M. S. Bhanu, "UbCadet: Detection of compromised accounts in Twitter based on user behavioural profiling," *Multimedia Tools Appl.*, vol. 79, nos. 1–2, pp. 1–37, Mar. 2020.
[62] G. Lingam, R. R. Rout, and D. V. L. N. Somayajulu, "Detection of social botnet using a trust model based on spam content in Twitter network," in *Proc. IEEE 13th Int. Conf. Ind. Inf. Syst. (ICIIS)*, Dec. 2018, pp. 280–285.
[63] M. Fazil and M. Abulaish, "A hybrid approach for detecting automated spammers in Twitter," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 11, pp. 2707–2719, Nov. 2018.
[64] P. S. Devi and S. Karthika, "#CycloneGaja-rank based credibility analysis system in social media during the crisis," *Procedia Comput. Sci.*, vol. 165, pp. 684–690, Jan. 2019.
[65] A. Javed, P. Burnap, M. L. Williams, and O. F. Rana, "Emotions behind drive-by download propagation on Twitter," *ACM Trans. Web*, vol. 14, no. 4, pp. 1–26, Sep. 2020.
[66] A. T. Kabakus and R. Kara, "'TwitterSpamDetector': A spam detection framework for Twitter," *Int. J. Knowl. Syst. Sci.*, vol. 10, no. 3, pp. 1–14, Jul. 2019.
[67] K. S. Adewole, T. Han, W. Wu, H. Song, and A. K. Sangaiah, "Twitter spam account detection based on clustering and classification methods," *J. Supercomput.*, vol. 76, no. 7, pp. 4802–4837, Jul. 2020.
[68] C. Zhao, Y. Xin, X. Li, Y. Yang, and Y. Chen, "A heterogeneous ensemble learning framework for spam detection in social networks with imbalanced data," *Appl. Sci.*, vol. 10, no. 3, p. 936, Jan. 2020.
[69] Y. Ji, Y. He, X. Jiang, and Q. Li, "Towards social botnet behavior detecting in the end host," in *Proc. 20th IEEE Int. Conf. Parallel Distrib. Syst. (ICPADS)*, Dec. 2014, pp. 320–327.
[70] J. Wang and I. C. Paschalidis, "Botnet detection using social graph analysis," in *Proc. 52nd Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2014, pp. 393–400.
[71] J. Rodríguez, M. A. Medina-Pérez, A. E. Gutierrez-Rodríguez, R. Monroy, and H. Terashima-Marín, "Cluster validation using an ensemble of supervised classifiers," *Knowl.-Based Syst.*, vol. 145, pp. 134–144, Apr. 2018.
[72] Y. Martínez-Díaz, H. Méndez-Vázquez, L. López-Avila, L. Chang, L. E. Sucar, and M. Tistarelli, "Toward more realistic face recognition evaluation protocols for the YouTube faces database," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 413–421.
[73] Y. Martínez-Díaz, N. Hernández, R. J. Biscay, L. Chang, H. Méndez-Vázquez, and L. Enrique Sucar, "On Fisher vector encoding of binary features for video face recognition," *J. Vis. Commun. Image Represent.*, vol. 51, pp. 155–161, Feb. 2018.
[74] O. Loyola-González, M. A. Medina-Pérez, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, R. Monroy, and M. García-Borroto, "PBC4cip: A new contrast pattern-based classifier for class imbalance problems," *Knowl.-Based Syst.*, vol. 115, pp. 100–109, Jan. 2017.
[75] N. El-Mawass, P. Honeine, and L. Vercouter, "SimilCatch: Enhanced social spammers detection on Twitter using Markov random fields," *Inf. Process. Manage.*, vol. 57, no. 6, Nov. 2020, Art. no. 102317.
[76] P. G. Efthimion, S. Payne, and N. Proferes, "Supervised machine learning bot detection techniques to identify social Twitter bots," *SMU Data Sci. Rev.*, vol. 1, no. 2, p. 5, 2018.
[77] R. A. Al-Dayil and M. H. Dahshan, "Detecting social media mobile botnets using user activity correlation and artificial immune system," in *Proc. 7th Int. Conf. Inf. Commun. Syst. (ICICS)*, Apr. 2016, pp. 109–114.
[78] B. Heredia, "Machine learning algorithms for the analysis of social media and detection of malicious user generated content," Ph.D. dissertation, Dept. Comput. Elect. Eng. Comput. Sci., Florida Atlantic Univ., Boca Raton, FL, USA, 2018.

[79] M. Kouvela, I. Dimitriadis, and A. Vakali, "Bot-detective: An explainable Twitter bot detection service with crowdsourcing functionalities," in *Proc. 12th Int. Conf. Manage. Digit. EcoSyst.*, Nov. 2020, pp. 55–63.
[80] A. Dorri, M. Abadi, and M. Dadfarnia, "SocialBotHunter: Botnet detection in Twitter-like social networking services using semi-supervised collective classification," in *Proc. IEEE 16th Int. Conf Dependable, Autonomic Secure Comput., 16th Int. Conf Pervas. Intell. Comput., 4th Int. Conf Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2018, pp. 496–503.
[81] M. K. Elhadad, K. F. Li, and F. Gebali, "COVID-19-FAKES: A Twitter (Arabic/English) dataset for detecting misleading information on COVID-19," in *Proc. Int. Conf. Intell. Netw. Collaborative Syst.* Cham, Switzerland: Springer, 2020, pp. 256–268.
[82] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, Mar. 2014.
[83] A. S. A. Shevtsov, M. Oikonomidou, D. Antonakaki, P. Pratikakis, A. Kanterakis, S. Ioannidis, and P. Fragopoulou, "Discovery and classification of Twitter bots," 2020, *arXiv:2010.15393*. [Online]. Available: http://arxiv.org/abs/2010.15393
[84] Q. Cao, X. Yang, J. Yu, and C. Palow, "Uncovering large groups of active malicious accounts in online social networks," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2014, pp. 477–488.
[85] J. Echeverria and S. Zhou, "Discovery, retrieval, and analysis of the 'Star Wars' botnet in Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Jul. 2017, pp. 1–8.
[86] G. Smith, "Modern day propaganda: Characteristics of fake news and psychological effects on the public," Ph.D. dissertation, Utica College, Utica, NY, USA, 2018.
[87] N. Abokhodair, D. Yoo, and D. W. McDonald, "Dissecting a social botnet: Growth, content and influence in Twitter," in *Proc. 18th ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, Feb. 2015, pp. 839–851.
[88] N. Pantic and M. I. Husain, "Covert botnet command and control using Twitter," in *Proc. 31st Annu. Comput. Secur. Appl. Conf. (ACSAC)*, 2015, pp. 171–180.
[89] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving Twitter spammers," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 8, pp. 1280–1293, Aug. 2013.
[90] K. Weller, A. Bruns, J. Burgess, M. Mahrt, and C. Puschmann, *Twitter and Society [Digital Formations]*, vol. 89. Bern, Switzerland: Peter Lang, 2014.
[91] J. Fields, "Botnet campaign detection on Twitter," 2018, *arXiv:1808.09839*. [Online]. Available: http://arxiv.org/abs/1808.09839
[92] S. Liu, J. Zhang, and Y. Xiang, "Statistical detection of online drifting Twitter spam," in *Proc. 11th ACM Asia Conf. Comput. Commun. Secur.*, May 2016, pp. 1–10.
[93] S. Sedhai and A. Sun, "Semi-supervised spam detection in Twitter stream," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 169–175, Mar. 2018.
[94] E. J. Kartaltepe, J. A. Morales, S. Xu, and R. Sandhu, "Social network-based botnet command-and-control: Emerging threats and countermeasures," in *Proc. Int. Conf. Appl. Cryptogr. Netw. Secur.* Berlin, Germany: Springer, 2010, pp. 511–528.

ABDELOUAHID DERHAB received the Engineer's, M.Sc., and Ph.D. degrees in computer science from the University of Sciences and Technology Houari Boummediene (USTHB), Algiers, in 2001, 2003, and 2007, respectively. He was a Computer Science Engineer and a full-time Researcher with the CERIST Research Center, Algeria, from 2002 to 2012. He was an Assistant Professor with King Saud University, from 2012 to 2018. He is currently an Associate Professor with the Center of Excellence in Information Assurance (COEIA), King Saud University. He is also a Cyber-Security Policy Analyst with Global Foundation for Cyber Studies and Research (GFCYBER). He is the author of more than 100 papers in different peer-reviewed journals and conferences. His research interests include malware analysis, networks security, intrusion detection, mobile security, the Internet of Things, smart grid, blockchain, and cyber security policies. He also served as the workshop chair, the technical committee chair, and a reviewer for many journals and international conferences.

RAHAF ALAWWAD is currently pursuing the master's degree with the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Her research interests include cyber security and computer networking.

KHAWLAH DEHWAH is currently pursuing the master's degree with the College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. Her research interests include online social networks, cyber security, and computer networks.

NOSHINA TARIQ received the M.S. and Ph.D. degrees in computer science from the Department of Computer Science, National University of Computer and Emerging Sciences, Islamabad, Pakistan. She is currently working as an Assistant Professor with the Shaheed Zulfikar Ali Bhutto Institute of Science and Technology (SZABIST), Islamabad. Her research interests include the Internet of Things, fog computing, cyber security, blockchain, and machine learning.

FARRUKH ASLAM KHAN (Senior Member, IEEE) received the M.S. degree in computer system engineering from the GIK Institute of Engineering Sciences and Technology, Pakistan, in 2003, and the Ph.D. degree in computer engineering from Jeju National University, South Korea, in 2007. He also received professional trainings from the Massachusetts Institute of Technology, New York University, IBM, and other institutions. He was the Founding Director of the Wireless Networking and Security (WiNGS) Research Group, National University of Computer and Emerging Sciences, Islamabad, Pakistan. He is currently working as a Professor with the Center of Excellence in Information Assurance, King Saud University, Riyadh, Saudi Arabia. He has published more than 110 research papers in refereed international journals and conferences. He has supervised/co-supervised five Ph.D. students and 18 M.S. thesis students. His research interests include cybersecurity, wireless sensor networks and e-health, bio-inspired and evolutionary computation, and the Internet of Things. He is also a Fellow of the British Computer Society (BCS). He is on the panel of reviewers of over 40 reputed international journals and numerous international conferences. He has co-organized several international conferences and workshops. He serves/served as an Associate Editor for prestigious international journals, including IEEE Access, *PLOS One*, *Neurocomputing* (Elsevier), *Ad Hoc and Sensor Wireless Networks*, *KSII Transactions on Internet and Information Systems*, *Human-Centric Computing and Information Sciences* (Springer), and *Complex & Intelligent Systems* (Springer).

JALAL AL-MUHTADI received the M.S. and Ph.D. degrees in computer science from the University of Illinois at Urbana–Champaign, USA. He is currently the Director and a Cybersecurity Consultant of the Center of Excellence in Information Assurance (CoEIA). He is also an Associate Professor with the Department of Computer Science, King Saud University. He has over 50 scientific publications in the areas of cybersecurity, information assurance, privacy, and the IoT security.