JOURNAL OF CYBERSECURITY

Journal of Cybersecurity, 2023, 1–12
https://doi.org/10.1093/cybsec/tyac015
Research Paper

Research Paper

Classifying social media bots as malicious or benign using semi-supervised machine learning

Innocent Mbona * **and Jan H.P. Eloff**

Department of Computer Science, University of Pretoria, Pretoria 0002, South Africa

*Correspondence address. E-mail: u15256422@tuks.co.za

Received 13 December 2021; revised 25 August 2022; accepted 31 October 2022

Abstract

Users of online social network (OSN) platforms, e.g. Twitter, are not always humans, and social bots (referred to as bots) are highly prevalent. State-of-the-art research demonstrates that bots can be broadly categorized as either malicious or benign. From a cybersecurity perspective, the behaviors of malicious and benign bots differ. Malicious bots are often controlled by a botmaster who monitors their activities and can perform social engineering and web scraping attacks to collect user information. Consequently, it is imperative to classify bots as either malicious or benign on the basis of features found on OSNs. Most scholars have focused on identifying features that assist in distinguishing between humans and malicious bots; the research on differentiating malicious and benign bots is inadequate. In this study, we focus on identifying meaningful features indicative of anomalous behavior between benign and malicious bots. The effectiveness of our approach is demonstrated by evaluating various semi-supervised machine learning models on Twitter datasets. Among them, a semi-supervised support vector machine achieved the best results in classifying malicious and benign bots.

Key words: Benford's law, benign bots, cybersecurity, feature selection, online social networks, semi-supervised machine learning, social bots, malicious bots

Introduction

Owing to reasons such as the simplicity of registering new accounts, online social networks (OSNs) such as Facebook and Twitter are constantly growing their user base [1]. When creating an account, identity-related data such as name and email address are required, which are used to create a unique user account [2]. Authentication methods, such as one-time password and completely automated public turing test to tell computers and humans apart (CAPTCHA) [3], are applied to block nonhuman users [4, 5]. Sometimes, users falsify identity-related information for privacy or malicious reasons such as cyberbullying [6]. Whenever a user provides forged identity-related information, that account is deemed fake [7]. Cresci et al. [4] deemed an OSN account to be fake if it is deceptive about its personal information (owner of an account), content (e.g. tweets), and followers.

Fake accounts, also known as Sybil accounts [8, 9], can be controlled either by a human or a social bot [4, 10]. A social bot can be defined as a social media account controlled by a computer program via an application programming interface (API) [11, 12]. Social bots can be used for either legitimate or malicious activities, such as internet trolling and fraud [11, 13, 14]. In this study, we focus on identifying bots as malicious or benign. The fact is that malicious bots are a constant threat to cybersecurity [6, 15], as they often engage in multiple cybercrimes, including social engineering attacks (SEAs). A SEA employs deception techniques to persuade users to unwittingly make security mistakes [16], such as clicking a malicious uniform resource locator (URL). Abreu et al. [17] investigated phishing SEAs executed by malicious bots on Twitter and discovered that malicious bots often execute phishing attacks by posting tweets embedded with malicious URLs. Akyon and Esat Kalfaoglu [18] discovered malicious bots that engaged in the dissemination of fake news, fake profiles, and content referred to as fake social engagements. Fake social engagement is a SEA that seeks to mislead users of OSNs [17, 18], e.g. an account that purchases followers through black markets to influence public opinion [18, 19]. Fake social engagements caused by malicious bots on OSNs are a long-standing problem that has severe consequences for cybersecurity [6]. Consider, for instance, the case of the Venezuela elections that were influenced by malicious bots spreading forged information through Twitter [20]. It is a severe threat to democratic

governments when computer programs can influence critical matters such as free and fair elections. Therefore, it is crucial to detect bots that engage in trolling activities or inflate their popularity to influence public opinion [18].

In early April 2021, over 533 million individuals' Facebook data were leaked online, which is a recent case of using malicious bots [21]. These personal data included phone numbers, email addresses, and locations scraped across different countries, most notably, the USA and the UK. The main question is, how were these massive data illegally obtained by cybercriminals? Various theories abound as investigations are still ongoing. However, there is overwhelming evidence that suggests that cybercriminals discovered a vulnerability in Facebook's developer API. This vulnerability was then exploited by the cybercriminals who scraped the massive data. A Facebook developer API is an application (bot) that enables certain activities such as following friends and posting content to be automated, and there are firm restrictions on the number of activities that can be automated. This was not the first incident of massive data leaked on Facebook. Facebook stated that the currently leaked data were part of the Cambridge Analytica data scandal [15], which was previously reported and resolved on August 2019. The leaked personal data can be exploited for multiple malicious activities, such as SEAs, scamming, hacking, and fraud [16, 17]. Users of OSNs are encouraged to visit the *HaveIBeenPwned* (https://haveibeenpwned.com) website to check if their email address or phone number has ever been leaked on dark websites. The above discussion further motivates the need to study the behavior of social bots and classify them as either malicious or benign, based on features found on OSNs.

Study motivation

There are various types of bots found in OSNs, such as Twitter and Facebook, which can be broadly categorized as either benign or malicious based on their activities [11, 13, 14]. Often, bot activities, such as posting content, are time asynchronous [19]. Both malicious and benign bots can operate under automation; however, their behavioral activities (i.e. features) and intent may differ [11, 12, 22]. Particularly, benign bots such as news update bots share breaking news, whereas malicious bots such as spamming bots spread spam [11]. These characteristic similarities and differences of social bots can cause benign or malicious bots to be misclassified [11]. Most previous studies focus on identifying features that assist in the correct classification of human users and malicious bots, with very little research on identifying features that can assist in the correct classification of malicious and benign bots.

In Table 1, we highlight the similarities and differences between benign and malicious social media bots. Although malicious bots have been extensively studied in the literature, benign bots have not received much attention. This is mostly because benign bots perform useful services such as sports updates that do not pose any cybersecurity threats, whereas activities such as web scraping performed by malicious bots pose severe cybersecurity threats. Radware [23] performed a network traffic analysis on their Web in 2018 and identified that humans, benign bots, and malicious bots contribute ~74%, ~17%, and ~9% of network traffic, respectively. It is well-established that malicious bots often mimic human behavior to avoid detection [24, 25], benign bots can also mimic human behavior. Specifically, Stieglitz et al. [12] highlighted benign bot accounts that mimic human behavior; however, some of these accounts declare upfront on their profiles that they are controlled by a bot. In addition, not all bot accounts declare this information [17, 18]; hence, this study proposes using behavioral features found on OSNs to distinguish between benign and malicious bots, which is crucial, and research in this area is inadequate [11, 23, 26].

Research questions and goals

The major contributions of this study stem from the following research questions.

(i) Can the same features used in previous studies to successfully distinguish between malicious bots and humans be useful in classifying benign and malicious bots?
(ii) What features found in the metadata of OSNs indicate anomalous behavior between benign and malicious bots?
(iii) Can semi-supervised machine learning (ML) models be used to classify malicious and benign bots, given a limited labeled dataset of such bots?

The remainder of this article is organized as follows. "State-of-the-Art" presents the state-of-the-art on benign and malicious bots. "Related Study" presents related studies. "Datasets and Methodology" describes the implementation of various feature selection (FS) methods on the benign and malicious bot Twitter dataset. "Classification of Benign and Malicious Bots" presents the results of four semi-supervised ML (SSML) models implemented to classify benign and malicious bots. "Results and Discussions" presents a discussion of the findings of this study. Finally, "Conclusions" concludes this study and describes a future study.

State-of-the-Art

The topic of profiling benign and malicious bots is still fairly new, so the terminology relating to these bots is diverse [11, 12]. For instance, most authors use the term social bot for "any social media account, i.e. controlled by a computer program that seeks to mimic human behavior" [14], whereas some use the same terminology when referring to a "harmful computer program social media account" [13]. The most accepted definition of benign and malicious bots from a cybersecurity perspective is that benign bots are bots that perform useful services [13, 27], such as aggregating content and offering news updates [12, 22]. Bots that engage in activities such as spamming, dissemination of fake news, and web scraping to steal user information are called malicious [11, 23]. Bots on OSNs were originally designed to assist humans to automate mundane activities, such as posting content. Unfortunately, the capabilities of bots can be hijacked for malicious activities [11] such as fake social engagements that damage the integrity of OSNs, which negatively impacts the end-user experience. Both benign and malicious bots are automated, but their behavioral features may differ [11, 12, 22]. For example, tweets from benign bots are retweeted more often than those from malicious bots [11]. There is sufficient evidence from the literature that most social bots are benign and only a few are malicious [11, 12]. Given that most bots are benign, we treat these as positive cases and malicious bots as negative cases (anomalies). We use the same dataset from [11] to identify features indicative of anomalous behavior between benign and malicious bots, where consumption[2] and broadcast[3] bots are treated as benign bots, and spambots are treated as malicious bots.

A plethora of literature exists on the classification of human and malicious bot accounts on OSNs [12–14]. Ferrara et al. [28, 29] designed a tool known as *botometer*, which predicts the likelihood of a

2 Consumption bots aggregate and distribute information such as trending topics.
3 Broadcast bots aim to provide benign news updates.

Table 1: Highlight: similarities and differences between benign and malicious social media bots

Benign bots	Malicious bots
Can run automated tasks, such as posting tweets	Can run automated tasks, such as posting tweets
Perform benign activities, such as posting news updates	Perform malicious activities, such as posting fake news and spreading spam
Contribute to network traffic	Contribute to network traffic
Often create real verified profiles	Often create fake unverified profiles
Often controlled by a single user	Often controlled by a botmaster[1] that can form a botnet

[1]A botmaster also known as a puppetmaster is a bot account that controls activities of at least one other bot account [24, 60].

Twitter account being used by a human or bot. Botometer can examine >1150 features extracted from users' account metadata of friends, content, sentiment, network, and timing. These features are subsequently used as inputs into botometer; afterward, a supervised random forest (RF) ML algorithm is implemented. A high score indicates a high probability of an account being a bot, and *vice versa*. Further, complete automation probability is computed using Bayes' theorem [28, 29] to account for the imbalanced dataset, i.e. numerous human accounts and few bot accounts. A botometer is used mostly by academics and practitioners. Its main limitation is that it does not classify bot accounts as either benign or malicious, and we are not aware of any tool that caters to this specific case. Botometers assume bots to be spam or self-promoting malicious bots [28, 29]. The contrasting behavioral patterns of benign and malicious bots can cause misclassification of these bots if features that differentiate them are not carefully selected [11]. Ferrara et al. [28, 29] argued that using public user metadata, such as the number of tweets and retweets, is vital for classifying human and bot accounts because OSNs (including Twitter) do not disclose personal user information in public. Hence, in most studies (including this study), public metadata is used to investigate behavioral features of benign and malicious bots. The major research issue here is whether the same criteria that have previously been used to successfully distinguish between malicious bots and humans can be used to distinguish between benign and malicious bots. In this study, we look into this issue.

Related Study

In this section, we sample relevant studies related to the profiling of social bots to lay a foundation for benign and malicious bot behaviors. Further, we highlight the approach, dataset type, and performance of existing bot detection methods.

Social media bots

Oentaryo et al. [11] identified different types of benign bots (broadcast and consumption bots) and investigated if benign bots could be differentiated from malicious bots (spam-promotion and spam-trick bots) and human Twitter accounts. The authors claimed they are the first to focus on classifying social media bots as either benign or malicious. They considered 484 benign bots and 105 malicious bots; this disproportion in data sizes implies an imbalanced dataset. Supervised ML (SML) models that include naive Bayes, logistic regression (LR), and RF were implemented to classify benign bots, humans, and malicious bots. The LR produced the best results with an F_1 score of 74%, where tweet, retweet, hashtag, mention, and URL features were identified as significant features for classifying benign and malicious bots using Shannon entropy. Features such as tweets and retweets are significant features for differentiating between humans and malicious bots in many previous studies, especially [28, 29].

Freeman and Hwa [2] proposed using SML techniques to classify a cluster of accounts as either fake or legitimate. The fake accounts were assumed to be created by either humans or bots. The authors used data from LinkedIn that contained clearly labeled data of fake and legitimate accounts. The RF model achieved the best results of 98% area under curve (AUC), whereby features such as account description and email address were deemed to be most significant using the Gini importance index. Akyon and Esat Kalfaoglu [18] investigated fake social engagements created by bot and fake accounts on Instagram. The common characteristics of fake accounts include a low number of followers_count, coupled with a high number of following_count, as well as strange names and profile pictures. Support vector machine (SVM), neural networks (NN), LR, and Bayesian-based ML models were implemented to classify fake and bot accounts. The NN model achieved the highest F_1 score of 89%. Features such as followers_count and following_count were identified to be significant in their classification problem. Cresci et al. [4] investigated various types of spam bot accounts found on Twitter. They evaluated the performance of human judgment and existing approaches in differentiating between legitimate (human), spam bot, and traditional spam bot accounts. Human judgment performed poorly (accuracy of 24%) in classifying spam bot accounts and approaches such as botometer were one of the good performing techniques with a recall of 95%. Varol et al. [28] used >1000 features extracted from user network patterns, activity time series, friends, sentiment, and tweet content to classify Twitter accounts as either humans or malicious bots. The RF ML model was implemented to classify human and malicious bot accounts, realizing an AUC of 94%. The authors also estimated that between 9% and 15% of Twitter accounts exhibit bot behavior. Van der Van Der Walt and Eloff [30] investigated fake accounts created by humans and bots in Twitter. They implemented various SML models, whereby RF was best performing with an accuracy of 87%. Further, they concluded that features used to detect bot accounts are not equally effective in detecting fake human accounts. Khaled et al. [31] also investigated the detection of fake bot (Sybil) Twitter accounts using a hybrid SML model of SVM and NN—SVM-NN—realizing accuracy of 98%.

Gilani et al. [32, 33] gathered a large-scale dataset of human and bot users on Twitter. Approximately 65 million tweets were collected; after which, this dataset was categorized into four groups of popularity using the follower_count feature as an indication of popularity. The first group comprised well-known accounts (celebrities and big brands) with a very high follower_count of >9 million. The second group was very popular accounts with follower_count ranging from 900 000 to 1.1 million. The third group comprised average popular accounts with follower_count ranging between 90 000 and 110 000. The fourth group was ordinary users with the smallest number of followers_count of <90 000. In each group, a certain number of bots and human accounts were considered. Further, in all groups, features such as status_count and retweet_count closely followed a Gaussian distribution. In each group, 11 features, such

Table 2: Summary: of well-known existing approaches to detect bots on social media platforms (LinkedIn, Instagram, and Twitter)

Author	Dataset type	Significant features used	Method used	Performance	Approach
Oentaryo et al. [11]	Humans, benign bots, and malicious bots	Tweet, retweet, hashtag, mention, and URL features	LR	F_1 score = 74%	Supervised learning
Freeman and Hwa [2]	Fake and legitimate accounts	Username, email address, and IP address	RF	AUC = 98%	Supervised learning
Cresci et al. [4]	Human, spam bots, and traditional spam bots	Not explicitly defined	Botometer (RF) Human judgment	Recall = 95% Accuracy = 24%	Supervised learning Human judgment
Gilani et al. [33]	Humans and bots	Account age, tweets, retweets, replies and mentions, URL_count, content uploaded, likes per tweet, retweets per tweet, favorites, friend–follower ratio, and activity source count	Botometer (RF)	Accuracy = 48%	Supervised learning
Varol et al. [28]	Humans and malicious bots	User network patterns, activity time series, friends, sentiment, and tweet content	RF	AUC = 94%	Supervised learning
Khaled et al. [31]	Humans and fake bots	Statuses_count, followers_count, friends_count, favorites_count, listed_count, geo_enabled, and profile features	SVM-NN	Accuracy = 94%	Supervised learning
Rodríguez-Ruiz et al. [27]	Humans, spam bots, and traditional spam bots	Retweets, replies, favorites, hashtag, URL_count, mentions, inter-time, friend–follower ratio, listed, uniqueHashtags, uniqueMentions, and uniqueURL	OCSVM	AUC = 89%	Supervised learning
Van der Walt and Eloff [30]	Humans and bots	Account age, duplicate_profile, followers_count, friends_count, geo_enabled, has_image, has_name, listed_count, status_count, and username_length	RF	Accuracy = 87% and F_1 score = 50%	Supervised learning
Akyon and Esat Kalfaoglu [18]	Humans and bots	Follower_count, following_count, highlight reel, external URL, tag number, and hashtag number	SVM	Precision = 91%	Supervised learning
Dorri et al. [24]	Human and spam bots	Followers–following ratio, tweets, account age, mentions, URL_count, and spamword_count	OCSVM	Recall = 99%	Semi-supervised learning
Shi et al. [35]	Humans and malicious bots	Likes_count, comments_count, friends_count, and sharing_count	K-means clustering	Recall = 90%	Semi-supervised learning
Chavoshi et al. [34]	Humans and bots	Tweets, URL_count, hashtag_count, mentions_count, and screen_name	Warped correlation	Precision = 94%	Unsupervised learning

as friend_follower ratio and account age, were applied to study the behavior of bots and humans. The known behaviors of bots, such as generating more content via tweets, retweets, replies, and mentions than humans, were observed. The significant contribution of the study of Gilani et al. [32, 33] was to highlight different behaviors of bots and humans at different levels of popularity groups. Specifically, well-known accounts were observed to have fewer bot accounts and tweets in their dataset, whereas ordinary accounts had more bot accounts and tweets. Rodríguez-Ruiz et al. [27] proposed using one-class SVM (OCSVM) anomaly detection (AD) to classify malicious bots and humans, experimented using the Twitter dataset of Cresci et al. [4] that encompasses various types of malicious bots, such as spam and retweet bots, and achieved an AUC of 89%. Chavoshi et al. [34] used unsupervised learning via a warped correlation method to classify Twitter users as humans or bots and detected bots with a 94% precision. They demonstrated that human activities, such as tweeting patterns, were not highly correlated, as opposed to bots.

Dorri et al. [24] presented a SocialBotHunter tool used to detect social botnets from human Twitter users. SocialBotHunter is based on OCSVM, which adopts the SSML approach. There are three main steps involved in SocialBotHunter: (i) feature extraction, (ii) anomaly score, and (iii) botnet detection. For each user, features including followers_following ratio, tweet repetition rate, account age (in days), inter-tweet time, mention_count, URL_count, and average number of spam words per tweet are used as inputs for the ML algorithm. The authors used the 1KS-10KN dataset containing 1000 spam bots and 10 000 human users to evaluate their proposed methodology. SocialBotHunter achieved a high detection rate with a recall of 99%. Shi et al. [35] proposed using a semi-supervised K-means clustering algorithm based on the transition probability of clickstream to classify humans and malicious Twitter bots. They used the CyVOD dataset in their experiment and achieved a high detection rate with a recall of 90%. In Table 2, we list the aforementioned existing well-known approaches for detecting social media bots as there exists a plethora of approaches. For each study, we state the dataset type (e.g. human or bot), significant features used, method used, performance of that method, and approach (e.g. supervised, semi-supervised, and unsupervised).

Table 2 suggests that SML approaches are most used to detect social media bots, in particular RF.

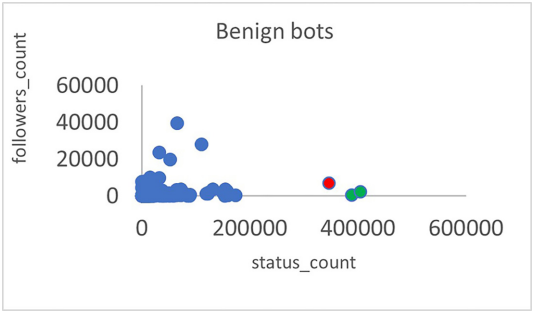

Figure 1: Anomalies in the benign bot dataset.

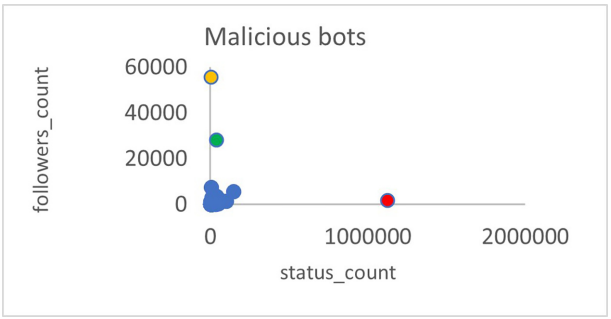

Figure 2: Anomalies in the malicious bot dataset.

FS

An FS problem can be formulated as a task of identifying optimal features in a dataset given some optimization criteria [36, 37]. Various FS methods have been proposed for OSN data [13, 14, 37]. Khaled et al. [31] proposed a hybrid SVM-NN model to detect malicious Twitter accounts. They collected real human data on Twitter through volunteers and bought fake (malicious) Twitter accounts through online black markets. In this hybrid model, 16 features were considered for the classification of human and malicious Twitter accounts. Four FS methods—the principal component analysis (PCA), correlation, regression, and wrapper-SVM—were applied to multiple data subsets of human and malicious accounts to identify significant features. In their FS experiments, it was discovered that followers_count, friends_count, favorites_count, and listed_count are significant features that successfully differentiate human and malicious accounts. It has been shown in the literature that most malicious bot accounts are usually fake [7, 30]; hence, we aim to investigate if the features found in [31] and other studies could be useful in classifying benign and malicious bots.

Benford's law

Benford's law (BL; see Appendices A and B), also known as the law of anomalous numbers, was discovered in 1881 by an astronomer—Simon Newcomb. Over the years, it has been applied in various domains, including network intrusions and OSNs [38]. Because various scholars [38–42] have demonstrated that BL applies to OSN data—especially Twitter—we do not buttress this point in this study. We only highlight the findings of each study. The actual distribution of each feature in Table 4 is compared with BL distribution using chi-squared and Kolmogorov–Smirnov tests. This goodness-of-fit statistics was used to examine whether there was a significant difference between the observed and expected distributions (see [43] for more details). The goodness-of-fit test is formulated as follows:

Null hypothesis (H_0) = a feature obeys the first significant leading digit (FSLD) distribution.

Alternative hypothesis (H_1) = a feature violates the FSLD distribution.

If P-value < 0.05, we reject H_0; otherwise, we cannot reject H_0.

Golbeck [39, 41] pioneered the application of BL to various OSN datasets, such as Facebook and Twitter datasets. These datasets contained features of humans and malicious bots. It was demonstrated by Golbeck [39, 41] that the friend_count, followers_count, and status_count obeyed BL (FSLD distribution) on the human dataset. The same features were shown to violate BL on the malicious bot dataset. Striga and Podobnik [38] used the Facebook dataset to demonstrate that BL (FSLD distribution) holds for likes, posts, and comments for human users. A similar finding was observed by Maurus and Plant [40], namely, that YouTube views, likes, dislikes, and comments conformed to BL for human users. In our previous study [42], BL was applied to Twitter datasets to identify significant features that differentiated human and malicious automated programs (bots). Further, it was demonstrated that features such as favorite_count and friends_count obeyed BL on human datasets, whereas they violated BL on malicious bot datasets. On the basis of the literature, not all automated programs are malicious [11, 12, 14, 23]. In this study, we perform further investigation if BL can identify features indicative of anomalous behavior between benign and malicious bots. To the best of our knowledge, BL has not been applied elsewhere to study behavioral features that can differentiate benign and malicious bots.

We hypothesized on the basis of the literature that benign and malicious bots differ in behavioral patterns, such as tweeting patterns [11, 22], and these patterns can be used to identify significant features. For instance, the raw data of Twitter numerical features may appear normal in absolute terms; however, analyzing the distribution of significant leading digits of these values may uncover anomalies [40, 42]. In general, numerical Twitter features, such as followers_count and friends_count, provide more insight into the behavior of an account, as opposed to nonnumerical features, such as *profile_has_image* [5, 40]. BL predicts that the distribution of features such as status_count is uneven [39]. Specifically, features with numbers beginning with "1" are expected to occur in ~30% of cases compared with features with numbers beginning with "9." The phenomenon of BL can be explained as follows: any positive real number $x \in \mathbb{R}^+$ can be written as a scientific notation $x = S(x) \times 10^k$, where $S(x) \in [0, 10)$ is called the significand and k is an integer called the exponent [43]. For example, a feature called status_count that has a value $x = 302$ can be written as 3.02×10^2 in scientific notation, where 3.02 is the significand, 2 is the exponent, and 3 is the FSLD. The main advantage of studying leading digits as opposed to absolute numbers is that every number has a unique FSLD (i.e. 1, 2, …, 9). For example, the FSLD distribution of a feature such as status_count can be analyzed on benign and malicious bot datasets to determine if this feature follows the FSLD distribution or not.

BL is a straightforward method to implement and does not require any parameter fitting, making it superior to other well-known nonuniform distributions, such as Power law and Zipf's law [40]. BL only cares about the FSLD distribution, and the lack of proportion in data sizes is insignificant [43]. Therefore, BL can be adopted as an FS method [42] to solve problems with imbalanced datasets, as in this study. Traditional FS methods such as the filter, wrapper, and embedded methods are generally ineffective for imbalanced datasets [33, 44]. Although advanced FS methods such as the ensemble RF (ERF) can handle imbalanced datasets, they are computationally expensive [45].

Table 3: Datasets: used in this study

Author	Dataset description	Notes
Oentaryo et al. [11]	Malicious bots—80, benign bots—373	Benign and malicious bots
Yang et al. [54]	Bots—698	Botwiki
Mazza et al. [55]	Bots—358	Botnets
Yang et al. [25]	Bots—17 882	Political bots

Table 4: Sample: of numerical features from the Twitter metadata used to detect malicious bots

Feature	Description
Account age (in days)	Number of days of accounts existence [28]
Screen_name length	Number of characters in the screen name [30]
Favorites_count	Number of tweets liked by a user [31]
URL_count	Number of URLs in the user profile [33]
Hashtag_count	Number of tweets and retweets for a specified # key phrase [11]
Lists_count	Number of pinned favorite lists [31]
Statuses_count	Number of tweets an account has [11]
Satus.retweet_count	Number of retweets an account has [11]
Status.reply_count	Number of replies per user's status [11]
Friends_count	Number of users an account is following [18]
Followers_count	Number of followers an account currently has [18]
Status.favorite_count	Number of likes per user's status [11]

Table 5: BL: test for benign and malicious bots, wherein bold features are significant

Feature	Benign	Malicious
Favourites_count	**Cannot reject H_0.**	**Reject H_0.**
Lists_count	**Cannot reject H_0.**	**Reject H_0.**
Statuses_count	**Cannot reject H_0.**	**Reject H_0.**
Status.retweet_count	**Cannot reject H_0.**	**Reject H_0.**
Friends_count	**Cannot reject H_0.**	**Reject H_0.**
Followers_count	**Cannot reject H_0.**	**Reject H_0.**
Status.favorite_count	Reject H_0.	Reject H_0.
Screen_name length	Reject H_0.	Reject H_0.
Account age (in days)	Reject H_0.	Reject H_0.
URL_count	Reject H_0.	Reject H_0.
Status.reply_count	Reject H_0.	Reject H_0.
Hashtag_count	Reject H_0.	Reject H_0.

AD

AD on big data platforms such as OSNs is an interesting subfield of ML that has attracted considerable attention recently [46]. Considering Twitter as an example, a sudden increase in the number of tweets is an anomaly event that can signify real trending topics that may be contaminated by fake news, as it was during the coronavirus outbreak [47]. AD is concerned with identifying anomalies in massive datasets [5, 48], such as detecting malicious bot accounts on Twitter. ML algorithms are also commonly used to detect malicious bots on OSNs [13, 14]. Well-known AD methods include density-, cluster-, distance-, graph-, and spectral-based techniques [48–50]. The underlying assumption in the cluster- and distance-based techniques is that "normal" datasets are expected to be clustered around the same point, whereas anomalies will be far from the normal points (see Figs 1 and 2 using a real-world Twitter dataset from [11]). The spectral-based techniques assume that high-dimensional data can be projected into a lower-dimensional subspace where normal and abnormal points can be identified. The PCA is a well-known spectral-based technique [50]. Statistical methods assume that "normal" data distribution occur in high probability regions, whereas anomalies occur in lower probability regions [48].

The main challenge with detecting anomalies (malicious bots) on OSNs is identifying features indicative of anomalous behavior [49, 50]. Usually, these features are identified on the basis of historical data using FS methods, such as ERF [45]. OSNs have many features, such as the number of friends and followers, that can be considered in an ML algorithm [28, 29]. Identifying meaningful features is crucial to designing effective ML algorithms, but this process can be computationally expensive [48, 49]. In this study, we demonstrate that BL is an effective rule for selecting features of benign and malicious bot data.

ML algorithms

AD algorithms can operate under supervised, unsupervised, and semi-supervised learning conditions depending on the availability of labeled datasets of normal and anomalous instances [46, 48]. In this study, labeled datasets should indicate benign or malicious bots as well as their features. Supervised AD techniques assume that there is a sufficient number of equally labeled normal and anomalous data in the training sets [48]. For example, for us to apply supervised AD, we would require large labeled datasets of benign and malicious bots. Unsupervised AD techniques do not assume any labels of the training data; they implicitly assume that normal datasets are grouped around the same point, whereas anomalies deviate from this point. For example, benign bot data are expected to be grouped around the same point, whereas malicious bot data deviate from this point [46]. We do not consider unsupervised learning in this study as it can often cluster data into multiple subgroups, whereas we aim to classify bots as either benign or malicious. Semi-supervised methods fall in between supervised and unsupervised methods [48, 51]. Semi-supervised algorithms assume that the training set comprises only normal labeled cases. If there are any unlabeled anomalous cases, they should be minimal (not >10% of the dataset) [46]. In this study, we adopt the

Table 6: Significant: features identified for differentiating benign and malicious bots

	BL	ERF	FPR	FDR	FWE
Significant features	Favorites_count, lists_count, statuses_count, status.retweet_count, friends_count, and followers_count	Statuses_count, followers_count, friends_count, lists_count, favorites_count, and status.retweet_count	Status_count, favorites_count, lists_count, friends_count, followers_count, and status.retweet_count	Status_count, favorites_count, lists_count, friends_count, followers_count, and status.retweet_count	Status_count, favorites_count, lists_count, friends_count, followers_count, and status.retweet_count

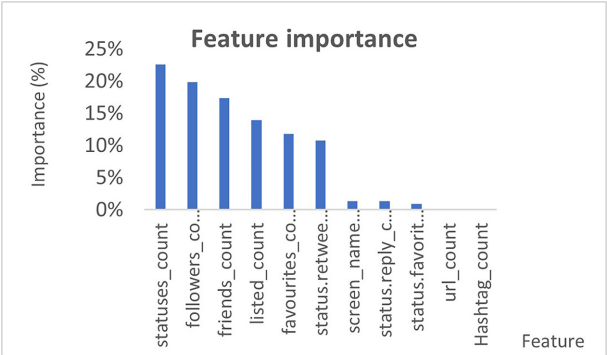

Figure 3: Feature importance using ERF. The size of a bar indicates the importance of each feature.

Table 7: Confusion: matrix for benign and malicious bots

	Actual malicious	Actual benign
Predicted malicious	True positive (TP)	False positive (FP)
Predicted benign	False negative (FN)	True negative (TN)

Table 8: GMM: benign versus malicious bots using significant features of BL

Model	Threshold	Precision	Recall	F1	MCC
GMM	$2.94e^{-05}$	69%	79%	77%	66%

semi-supervised learning approach, given that we have a sample of labeled datasets of benign and malicious bots in Table 3 to train the proposed models.

Semi-supervised algorithms are generally applicable to many real-life problems, such as OSNs, as in most cases datasets for normal cases can be obtained, but anomalous data are often unavailable [44, 46]. The detection of anomalies strongly depends on the modeling of normal or expected distribution [49, 50]. It is difficult to set boundaries on OSNs that define normal behavior, as users' behavior evolves [5, 7]. The rule of thumb is that data with most records grouped are assumed to be normal, whereas points that deviate from this normal group are anomalies [52, 53]. Anomalies are detected on the basis of position or distance from normal behavior. In Figs 1 and 2, we illustrate individual and group anomalies in a simple 2D plot for a benign and malicious dataset using the dataset in Table 3.

Datasets and Methodology

We used the same dataset from [11] to investigate features indicative of anomalous behavior between benign and malicious bot data. BL is used to identify significant features between benign and malicious bots. Our dataset comprised metadata of benign and malicious Twitter bots. From the original list of benign and malicious bot accounts [11] shared with us, we discovered that some accounts had been deactivated or inactive, hence were excluded from our data. Table 3 summarizes the datasets used in this study.

To the best of our knowledge, the dataset from [11] is the main dataset that contains clearly labeled benign and malicious bots; therefore, we refer to this dataset as "labeled." Further, the dataset is not sufficiently large (Table 3). There exist a couple of options for addressing this problem, such as falsifying data with labels or using unlabeled real data. We opted for the latter option. Now, the question is, can these datasets be combined in one experiment? The answer is yes, given that the Twitter datasets are based on the same data structure of features, such as status_count. To confirm this, we performed the Mann–Whitney U test to demonstrate that no bias was introduced by combining datasets. The results for this are found in Appendix G (Table A17). The datasets from [25, 54, 55] do not contain labels (i.e. benign or malicious bots); hence, we treat these datasets as "unlabeled."

Figures 1 and 2 depict datasets of status_count vs. followers_count for benign and malicious bots, respectively. In both figures, data points (blue) are relatively grouped, and hence can be assumed to be "normal" behavior. The colored data points are distanced from these normal data and can be deemed anomalies. We aim to demonstrate that BL can effectively identify features indicative of anomalous behavior between benign and malicious bots, which will help the proposed ML models to distinguish between benign and malicious bots.

FS using BL

We applied BL to identify features that can differentiate benign and malicious bots meaningfully. It has been demonstrated by [39, 40] that Twitter datasets satisfy minimum dataset requirements by BL; therefore, BL can be applied [39, 40]. In Table 4, we indicate features generally found in the metadata of benign and malicious bots [11]. Notably, these are the same features that were used in the past to differentiate human and malicious bot datasets.

Then, we compared the BL distribution with the actual distribution for each feature in Table 4 to identify significant features among benign and malicious bots, (A graphical representation of BL tests is provided in Appendix B).

A feature is only significant if it obeys BL on the benign bot dataset and simultaneously violates BL on the malicious bot dataset. If a feature violates or does not violate BL on both datasets, then that feature is not deemed significant [40]. On the basis of the results reported in Table 5, features in bold are significant for differentiating

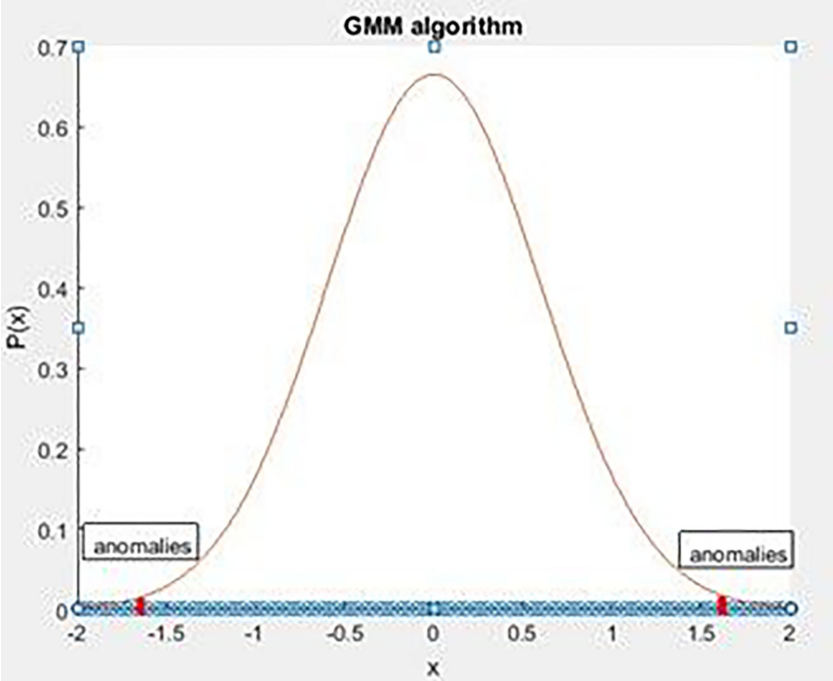

Figure 4: GMM for the first 300 samples, where malicious bots are indicated as anomalies. Malicious bot samples are found on the lower regions of the distribution as anomalies, whereas benign bots are found on the higher probability regions.

Table 9: GMM: benign versus malicious bots using all features in Table 4

Model	Threshold	Precision	Recall	F1	MCC
GMM	$2.04e^{-04}$	62%	74%	67%	58%

malicious and benign bots. These results are intuitive and align with the known behaviors of bots mainly discussed in [11, 31]. For example, consider the lists_count feature that allows users to pin tweets that will appear at the top of their profile, regardless of time. Malicious bots are observed to pin more tweets than benign bots so that when other users visit their profiles, they will immediately see those tweets conveying a particular message. Features such as screen_name length and account age violated BL on both datasets, and thus are not deemed significant. This is rational as the features do not provide much insight into terms of user behavior. We observed that the identified significant features for differentiating benign and malicious bots in Table 5 are the same as those in our previous study [42], which used BL to distinguish between humans and malicious bots. This suggests that benign bots and human accounts display similar behaviors. To further illustrate the effectiveness of BL in differentiating benign and malicious bots, we implemented four FS methods on the same dataset from [11]. Their results are consistent as indicated in Tables 5 and 6.

FS using ERF

ERF works for high-dimensional imbalanced data problems, such as OSNs, given the hierarchical structure that allows them to learn from majority and minority classes [56]. In this study,

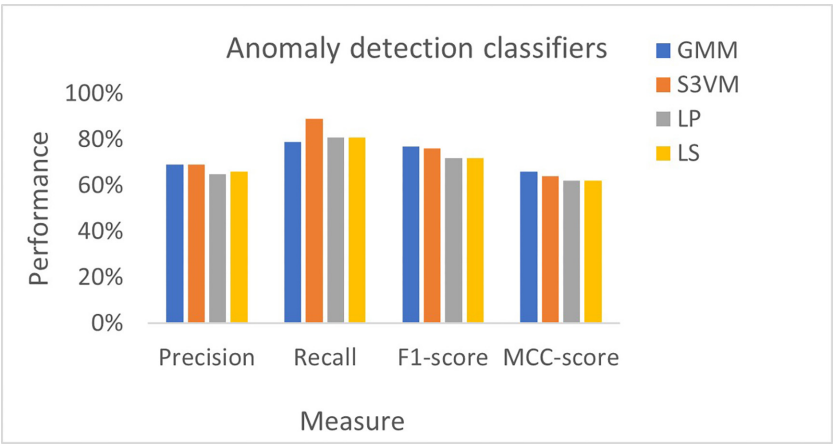

Figure 5: AD classifier performance using features identified by BL.

Table 10: S3VM: benign versus malicious bots using significant features of BL

Model	Precision	Recall	F1	MCC
S3VM	69%	89%	76%	64%

Table 11: S3VM: benign versus malicious bots using all features in Table 4

Model	Precision	Recall	F1	MCC
S3VM	61%	76%	67%	60%

Table 12: LP: benign versus malicious bots using significant features of BL

Model	Precision	Recall	F1	MCC
LP	65%	81%	72%	62%

Table 13: LP: benign versus malicious bots using all features in Table 4

Model	Precision	Recall	F1	MCC
LP	58%	74%	63%	59%

Table 14: LS: benign versus malicious bots using significant features of BL

Model	Precision	Recall	F1	MCC
LS	66%	81%	72%	62%

benign and malicious bots are considered majority and minority classes, respectively. The results in Fig. 3 indicate that status_count, followers_count, friends_count, listed_count, favorite_count, and retweet_count are important features for differentiating benign and malicious bots. The results agree with the BL results in Table 5. The length of a bar indicates the importance of each feature.

Notably, BL was found to produce similar results as ERF in Fig. 3. However, BL is highly computationally efficient compared with ERF. Thus, BL could be more beneficial to ML algorithms used to classify benign and malicious bots in real time. ERF can efficiently analyze big data; hence, it can be expected to reduce the computational cost of identifying meaningful features. Next, we implemented three well-known FS methods from ScikitLearn library. ScikitLearn library was chosen because it is one of the biggest open-source libraries, mostly used by academics and practitioners.

FS using false positive rate

The false positive rate (FPR) method selects significant features in a dataset based on univariate statistical tests. This method selects significant features based on a *P*-value computed using chi-squared or analysis of variance; see [37] for more details. Using a *P*-value of 0.05, status_count, favorites_count, lists_count, friends_count, followers_count, and status.retweet_count were identified as significant features.

FS using false discovery rate

The false discovery rate (FDR) method identifies significant features using the Benjamini–Hochberg algorithm and an upper bound alpha on the expected FDR; see [37] for more details. The FDR method produced the same results as the FPR method for an alpha of 0.05.

FS using familywise error

The familywise error (FWE) method identifies significant features using the Bonferroni algorithm and an upper bound alpha; see [37] for more details. The FWE method produced the same results as the FPR method for an alpha of 0.05.

In summary, significant features identified by BL to differentiate benign and malicious bots proved to be consistent with those identified by the ERF, FPR, FDR, and FWE.

Evaluation measures

In this study, the proposed methods are evaluated using standard measures for imbalanced datasets, including precision, recall, Matthews correlation coefficient, and F_1 score. The mathematical formulas for these measures are easily derived from the confusion matrix in Table 7 (see [57, 58]).

$$Precision = \frac{TP}{TP+FP}.$$

$$Recall = \frac{TP}{TP+FN}.$$

$$F_1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}.$$

$$MCC - score = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\,(TP+FN)\;(TN+FP)\;(TN+FN)}}.$$

Classification of Benign and Malicious Bots

In our experimental setup, we split the datasets in Table 3 into two groups: the training and test datasets. The training dataset was used to train the ML model and fit its parameters, whereas the test dataset was used to test the performance of an ML model. The training dataset comprised 103 labeled datasets (30 malicious bots and 73 benign bots) and 18 938 unlabeled datasets (i.e. bot samples from [25, 54, 55]). The test dataset comprised 350 samples (50 malicious bots and 300 benign bots). The significant features used in this classification are indicated in Table 5, which are favorites_count, lists_count, statuses_count, status.retweet_count, friends_count, and followers_count.

Semi-supervised Gaussian mixture model

To address research question (iii), we evaluated various semi-supervised AD algorithms for classifying malicious and benign bots. The semi-supervised Gaussian mixture model (GMM) AD is considered suitable for evaluating continuous data such as OSN data that have many features [44, 46, 48, 52, 53]. The GMM algorithm is summarized in Appendix C.

Semi-supervised SVM

Next, we applied semi-supervised SVM (S3VM) on the same dataset to classify benign and malicious bots, as part of our experimental study. S3VM applies the maximum margin principle, which aims to design a binary classifier using labeled and unlabeled datasets [59]. Hence, S3VM was considered suitable for the binary classification

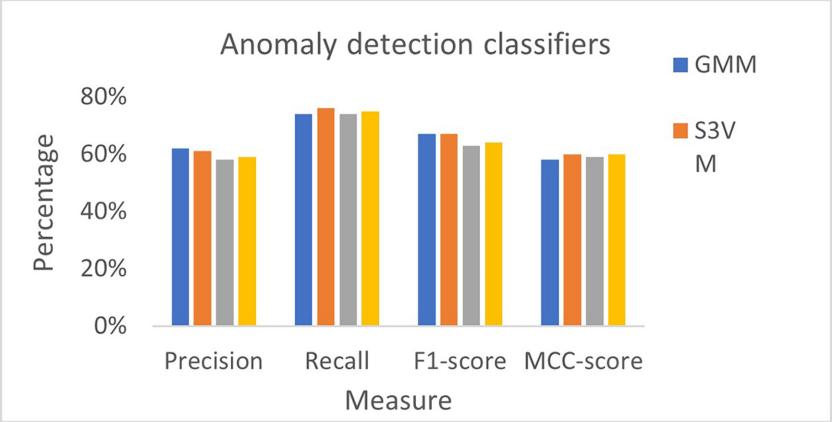

Figure 6: AD classifier performance using all features in Table 4.

Table 15: LS: benign versus malicious bots using all features in Table 4

Model	Precision	Recall	F1	MCC
LS	59%	75%	64%	60%

problem considered in this study. S3VM is summarized in Appendix D, and further details are found in [45, 59].

Semi-supervised label propagation

The label propagation (LP) method is a graph-based SSML model that uses a sample of labeled nodes on a graph to extend the labeling of all nodes on the graph until convergence is achieved [45]. This method is suitable for the considered problem, given that we have a dataset of labeled samples of benign and malicious bots. Specifically, we consider N labeled data points denoted by $+1$ (benign bots) and -1 (malicious bots) and M unlabeled data points denoted by $y = 0$. Let $G = \{V, E\}$ represent a graph with every vertex comprising labels $V = \{+1, -1, 0\}$, and the edge is based on affinity matrix W. Further, the features of the dataset are denoted by X. The intuition behind the LP method is that two nodes are connected if they are "similar;" therefore, unlabeled data points can be labeled by propagating labeled data points to their neighbors by iterating until convergence. The LP method is summarized in Appendix E; for further details, see [45].

Semi-supervised label spreading

The label spreading (LS) method is similar to the LP method, except that the labeled data points on the vertex may change during the iteration process. The clamping factor $\alpha \; \varepsilon \; (0, 1]$ determines whether labeled data points will update. If $\alpha = 0$, the LS method will not update the original labels as in the case of the LP method. The LS method is summarized in Appendix F; for further details, see [45].

Results and Discussions

In this study, we classified benign and malicious bots of OSNs through features indicative of anomalous behavior. In total, 6 of 12 features were identified by BL to be significant for differentiating between benign and malicious bots (Table 6). The six features were then used as inputs into semi-supervised AD classifiers. A total of four SSML algorithms were implemented to classify benign and malicious bots. GMM was implemented to address the binary classification problem, and it produced good results (Table 8 and Fig. 4) when using features identified by BL. Using all features found in Table 4 did not improve GMM, as indicated in Table 9. Further, S3VM was implemented to solve the same classification problem. S3VM outperformed other algorithms when using features identified by BL in Table 5 (Fig. 5). For all algorithms, the inclusion of all features in Table 4 did not improve the classification performance, as indicated in Fig. 6. One possible explanation for this is that adding less meaningful features, such as account age and screen_name length, could negatively impact the classification. Such features were observed to violate BL on both datasets; thus, they were not deemed to be significant. This is intuitive, as these features did not provide much insight into the behavior of benign and malicious bots. LP and LS also produced good results as demonstrated in Tables 12 and 14, respectively.

Figure 5 is based on the results in Tables 8, 10, 12, and 14, whereas Fig. 6 is based on the results in Tables 9, 11, 13, and 15.

Finally, we employed the findings of [11] to benchmark the proposed algorithms. For our purposes, we were more interested in the correct classification of malicious and benign bots. Oentaryo et al. [11] implemented four SML methods and found that LR and SVM produced the best results in terms of recall and F_1 score. Particularly, LR and SVM achieved recalls of 81% and 76% and F_1 scores of 74% and 73%, respectively. In our experiments, S3VM produced the best results with a recall of 89% and an F_1 score of 76%. In summary, our results demonstrated that using features indicative of anomalous behavior could yield better results, as opposed to using just "many" features. Features in Table 4 were previously used to successfully classify human and malicious bots. However, we have demonstrated that using all these features does not improve the classification. Moreover, including unlabeled bot dataset in the ML slightly improved the classification in terms of recall and F_1 score. Given that we used raw metadata from Twitter, our findings are scalable to other OSN domains, such as Facebook and LinkedIn.

The main limitation of this study is the lack of publicly available datasets of benign and malicious bots. Further study is required to design a real-time system such as a botomoter for classifying bots as either benign or malicious.

Conclusions

In this study, we investigated which OSN features are useful in differentiating malicious and benign bots. Twitter datasets encompass-

ing benign and malicious bots were used to discover suspicious behaviors, such as "like fraud" and "retweet" spam caused by malicious bots. Finally, once significant features were identified, we implemented four semi-supervised AD algorithms, including GMM, S3VM, LP, and LS, to classify malicious and benign bots. S3VM produced the best results with a recall of 89% and an F_1 score of 76%. Further, we demonstrated that features effective in classifying human users and malicious bots are not equally effective in classifying benign and malicious bots. We believe that our findings will help to minimize cyber threats caused by malicious bots and improve the end-user experience on OSNs. Further research will investigate behavioral activities (including status content features) of different types of benign bots (e.g. sport and comic bots) and malicious bots (e.g. phishing and fraudulent advertisement bots) on various OSNs.

Supplementary Data

Supplementary data is available at *Cybersecurity Journal* online.

Acknowledgment

Innocent Mbona would like to thank the University of Pretoria and Bank Seta for funding this study. We also thank the Singapore Management University (SMU) for having granted us access to their malicious and benign bot dataset.

Author Contributions

Innocent Mbona (Conceptualization, Investigation, Methodology, Writing - original draft), and Jan Eloff (Supervision, Validation, Writing - review & editing).

References

1. Appel G, Grewal L, Hadi R. *et al*. The future of social media in marketing. *J Acad Mark Sci* 2020;**48**:79–95.
2. Freeman DM, Hwa T. Detecting clusters of fake accounts in online social networks categories and subject descriptors. In: *Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security, AIsec 15*. p. 91–101. New York, NY: Association for Computing Machinery, 2015.
3. Xu X, Liu L, Li B. A survey of CAPTCHA technologies to distinguish between human and computer. *Neurocomputing* 2020;**408**:292–307.
4. Cresci S, Spognardi A, Petrocchi M. *et al*. The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. In: *Proceedings of the 26th International Conference on World Wide Web Companion 2017*. p. 963–72. New York, NY: Association for Computing Machinery, 2017. https://doi.org/10.1145/3041021.3055135.
5. Chauhan V, Pilaniya A, Middha V. *et al*. Anomalous behavior detection in social networking. In: *Proceedings of the 8th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2017*. Piscataway, NJ: Institute of Electrical and Electronics Engineers Inc, 2017.
6. Kayes I, Iamnitchi A. Privacy and security in online social networks: a survey. *Online Soc Networks Media* 2017;**3–4**:1–21.
7. Gurajala S, White JS, Hudson B. *et al*. Fake twitter accounts: profile characteristics obtained using an activity-based pattern detection approach. In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. NW Washington, DC: IEEE Computer Society, 2015. https://doi.org/10.1145/2789187.2789206.
8. Paavola J, Helo T, Jalonen H. *et al*. The automated detection of trolling bots and cyborgs and the analysis of their impact in the social media. In: *Proceedings of the European Conference on Cyber Warfare and Security ECCWS 2016*. p. 237–44. Berkshire: Academic Publishing International Ltd, 2016.
9. Al-Qurishi M, Al-Rakhami M, Alamri A. *et al*. Sybil defense techniques in online social networks: a survey. *IEEE Access* 2017;**5**:1200–19.
10. Roy P, Sood M. Implementation of ensemble-based prediction model for detecting sybil accounts in an osn. In: *Advances in Intelligent Systems and Computing*. Berlin: Springer, 2021, 709–23.
11. Oentaryo RJ, Murdopo A, Prasetyo PK. *et al*. On profiling bots in social media. In: *Lecture Notes Computer Science (Including Subseries Lecture Notes in Artificial Intelligence, Lecture Notes on Bioinformatics) 10046 LNCS*. Berlin: Springer, 2016, 92–109. https://doi.org/10.1007/978-3-319-47880-7_6.
12. Stieglitz S, Brachten F, Ross B, Jung AK. Do social bots dream of electric sheep? A categorisation of social media bot accounts. In: *Proceedings of the 28th Australasian Conference on Information Systems ACIS 2017*. p. 1–11. Auckland: Auckland University of Technology Library, 2017.
13. Latah M. Detection of malicious social bots: a survey and a refined taxonomy. *Expert Syst Appl* 2020;**151**:113383.
14. Orabi M, Mouheb D, Al Aghbari Z. *et al*. Detection of bots in social media: a systematic review. *Inf Process Manag* 2020;**57**:102250.
15. Hinds J, Williams EJ, Joinson AN. "It wouldn't happen to me": privacy concerns and perspectives following the Cambridge Analytica scandal. *Int J Hum Comput Stud* 2020;**143**:102498.
16. Hatfield JM. Social engineering in cybersecurity: the evolution of a concept. *Comput Secur* 2018;**73**:102–13.
17. Abreu JVF, Fernandes JHC, Gondim JJC, Ralha CG. Bot development for social engineering attacks on Twitter. arXiv preprint arXiv:2007.11778 2020.
18. Akyon FC, Esat Kalfaoglu M. Instagram fake and automated account detection. In: *Proceedings of the Conference on Innovations in Intelligent Systems and Applications, ASYU 2019*. Çankaya: IEEE Turkey, 2019. https://doi.org/10.1109/ASYU48272.2019.8946437.
19. Dutta HS, Aggarwal K, Chakraborty T. *DECIFE: Detecting Collusive Users Involved in Blackmarket following Services on Twitter*. New York NY: Association for Computing Machinery, 2021.
20. Forelle MC, Howard PN, Monroy-Hernandez A, Savage S. Political bots and the manipulation of public opinion in Venezuela. *SSRN Electron J* 2015;**1**:1–8. https://doi.org/10.2139/ssrn.2635800.
21. Post T 533M Facebook Accounts Leaked Online: Check if You Are Exposed | Threatpost. https://threatpost.com/facebook-accounts-leaked-check-exposed/165245/. (12 March 2022, date last accessed).
22. Brachten F, Mirbabaie M, Stieglitz S. *et al*. Threat or opportunity? Examining social bots in social media crisis communication. In: *Proceedings of the 29th Australian Conference on Information Systems, ACIS 2019*. p. 1–8. Broadway: UTS ePRESS, 2018. https://doi.org/10.5130/acis2018.bo.
23. Radware Radware Research "Inside Good Bots": Understanding Management of Benign Bot Traffic. https://www.radwarebotmanager.com/shieldsquare_research_inside_good_bots/. (23 January 2021, date last accessed).
24. Dorri A, Abadi M, Dadfarnia M. SocialBotHunter: Botnet detection in twitter-like social networking services using semi-supervised collective classification. In: *Proceedings of the 16th IEEE International Conference on Dependable, Autonomic and Secure Computing, IEEE International Conference on Pervasive Intelligence and Computing, 4th IEEE International Conference on Big Data Intelligence and Computing*. p. 496–503. Piscataway, NJ: IEEE, 2018. https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00097.
25. Yang KC, Varol O, Davis CA. *et al*. Arming the public with artificial intelligence to counter social bots. *Hum Behav Emerg Technol* 2019;**1**:48–61.
26. Zabihimayvan M, Sadeghi R, Rude HN. *et al*. A soft computing approach for benign and malicious web robot detection. *Expert Syst Appl* 2017;**87**:129–40.
27. Rodríguez-Ruiz J, Mata-Sánchez JI, Monroy R. *et al*. A one-class classification approach for bot detection on Twitter. *Comput Secur* 2020;**91**:101715.
28. Varol O, Ferrara E, Davis CA. *et al*. Online Human Bot Interaction. In: *Proceedings of the 11th International AAAI Conference on Web and Social Media, ICWSM 2017*. p. 280–9. Palo Alto, CA: AAAI Press, 2017.
29. Ferrara E, Varol O, Davis C. *et al*. The rise of social bots. *Commun ACM* 2016;**59**:96–104.

30. Van Der Walt E, Eloff J. Using machine learning to detect fake identities: bots vs humans. *IEEE Access* 2018;**6**:6540–9.
31. Khaled S, El-Tazi N, Mokhtar HMO. Detecting fake accounts on social media. In: *Proceedings of the 2018 IEEE International Conference on Big Data, Big Data 2018*. p. 3672–81. Seattle, WA: IEEE, 2019. doi: 10.1109/BigData.2018.8621913.
32. Gilani Z, Farahbakhsh R, Tyson G. *et al.* Of bots and humans (on Twitter). In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Network Analysis and Mining, ASONAM 2017*. p. 349–54. New York, NY: Association for Computing Machinery, 2017. https://doi.org/10.1145/3110025.3110090.
33. Gilani Z, Farahbakhsh R, Tyson G. *et al.* A large-scale behavioural analysis of bots and humans on twitter. *ACM Trans Web* 2019;**13**:1–24.
34. Chavoshi N, Hamooni H, Mueen A. DeBot: Twitter bot detection via warped correlation. In: *Proceedings of the 16th IEEE International Conference on Data Mining (ICDM)*. p. 817–22. Barcelona: IEEE, 2017. https://doi.org/10.1109/icdm.2016.0096.
35. Shi P, Zhang Z, Choo KKR. Detecting malicious social bots based on clickstream sequences. *IEEE Access* 2019;**7**:28855–62.
36. Vinet L, Zhedanov A. A "missing" family of classical orthogonal polynomials. *J Phys A Math Theor* 2011;**44**:085201.
37. Pilnenskiy N, Smetannikov I. Feature selection algorithms as one of the Python data analytical tools. *Futur Internet* 2020;**12**:54.
38. Striga D, Podobnik V. Benford's law and Dunbar's number: does Facebook have a power to change natural and anthropological laws?. *IEEE Access* 2018;**6**:14629–42.
39. Golbeck J. Benford's law applies to online social networks. *PLoS ONE* 2015;**10**:e0135169.
40. Maurus S, Plant C. Let's see your digits: anomalous-state detection using Benford's Law. In: *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Part F1296*. p. 977–86. Nova Scotia: IEEE, 2017. https://doi.org/10.1145/3097983.3098101.
41. Golbeck J. Benford's Law can detect malicious social bots. *First Monday* 2019;1–8:**24**. https://doi.org/10.5210/fm.v24i8.10163.
42. Mbona I, Eloff JHP. Feature selection using Benford's law to support detection of malicious social media bots. *Inf Sci* 2022;**582**:369–81.
43. Miller SJ. *Benford's Law: Theory and Applications*. Princeton, NJ: Princeton University Press, 2015.
44. Ding N, Ma HX, Gao H. *et al.* Real-time anomaly detection based on long short-term memory and Gaussian Mixture Model. *Comput Electr Eng* 2019;**79**:106458.
45. Bonaccorso G. *Mastering Machine Learning Algorithms: Expert Techniques for Implementing Popular Machine Learning Algorithms, Fine-Tuning your Models, and Understanding How They Work*. Birmingham: Packt Publishing, 2020.
46. Xu X, Liu H, Yao M. Recent progress of anomaly detection. *Complex* 2019;**2019**:1–11.
47. van der Linden S, Roozenbeek J, Compton J. Inoculating against fake news about COVID-19. *Front Psychol* 2020;**11**:1–7.
48. Islam MR, Liu S, Wang X. *et al.* Deep learning for misinformation detection on online social networks: a survey and new perspectives. *Soc Netw Anal Min* 2020;**10**:1–20.
49. Thudumu S, Branch P, Jin J, Singh J. A comprehensive survey of anomaly detection techniques for high dimensional big data. *J Big Data* 2020;**7**:42. https://doi.org/10.1186/s40537-020-00320-x.
50. Kurka DB, Godoy A, Von Zuben FJ. Online social network analysis: a survey of research applications in computer science. *Soc Inf Process Netw* 2015;**1**:1–55.
51. van Engelen JE, Hoos HH. A survey on semi-supervised learning. *Mach Learn* 2020;**109**:373–440.
52. Mittal R, Bhatia MPS. Anomaly detection in multiplex networks. *Proc Comput Sci* 2018;**125**:609–16.
53. Sun X, Zhang C, Li G. *et al.* Detecting users' anomalous emotion using social media for business intelligence. *J Comput Sci* 2018;**25**:193–200.
54. Yang KC, Varol O, Hui PM, Menczer F. Scalable and generalizable social bot detection through data selection. In: *Proceedings of the 34th AAAI Conference on Artificial Intelligence*. p. 1096–103. New York, NY: Association for the Advancement of Artificial Intelligence, 2020. https://doi.org/10.1609/aaai.v34i01.5460.
55. Mazza M, Cresci S, Avvenuti M. *et al.* RTbust: exploiting temporal patterns for botnet detection on twitter. In: *Proceedings of the 11th ACM Conference on Web Science, WebSci 2019*. p. 183–92. Boston, MA: Association for Computing Machinery, 2019. https://doi.org/10.1145/3292522.3326015.
56. Amr T. *Hands-on Machine Learning with Scikit-Learn and Scientific Python Toolkits*. Vol. **384**. Birmingham: Packt Publishing, 2019.
57. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 2020;**21**:1–13.
58. Thabtah F, Hammoud S, Kamalov F. *et al.* Data imbalance in classification: experimental evaluation. *Inf Sci* 2020;**513**:429–41.
59. Ding S, Zhu Z, Zhang X. An overview on semi-supervised support vector machine. *Neural Comput Appl* 2017;**28**:969–78.
60. Kumar S, Leskovec J, Cheng J. *et al.* An army of me: Sockpuppets in online discussion communities. In: *Proceedings of the 26th International Conference on World Wide Web*. p. 857–66. Geneva: International World Wide Web Conference Committee, 2017. https://doi.org/10.1145/3038912.3052677.