Social Network Analysis and Mining (2023) 13:20
https://doi.org/10.1007/s13278-022-01020-5

ORIGINAL ARTICLE

Machine learning-based social media bot detection: a comprehensive literature review

Malak Aljabri[1] · Rachid Zagrouba[3] · Afrah Shaahid[2] · Fatima Alnasser[2] · Asalah Saleh[2] · Dorieh M. Alomari[4]

Received: 24 October 2022 / Revised: 30 November 2022 / Accepted: 20 December 2022 / Published online: 5 January 2023

Abstract
In today's digitalized era, Online Social Networking platforms are growing to be a vital aspect of each individual's daily life. The availability of the vast amount of information and their open nature attracts the interest of cybercriminals to create malicious bots. Malicious bots in these platforms are automated or semi-automated entities used in nefarious ways while simulating human behavior. Moreover, such bots pose serious cyber threats and security concerns to society and public opinion. They are used to exploit vulnerabilities for illicit benefits such as spamming, fake profiles, spreading inappropriate/false content, click farming, hashtag hijacking, and much more. Cybercriminals and researchers are always engaged in an arms race as new and updated bots are created to thwart ever-evolving detection technologies. This literature review attempts to compile and compare the most recent advancements in Machine Learning-based techniques for the detection and classification of bots on five primary social media platforms namely Facebook, Instagram, LinkedIn, Twitter, and Weibo. We bring forth a concise overview of all the supervised, semi-supervised, and unsupervised methods, along with the details of the datasets provided by the researchers. Additionally, we provide a thorough breakdown of the extracted feature categories. Furthermore, this study also showcases a brief rundown of the challenges and opportunities encountered in this field, along with prospective research directions and promising angles to explore.

Keywords Social media security · Bot detection · Machine learning · Social bots · Feature engineering · Cybersecurity

1 Introduction

In this modern world, OSNs such as Twitter, Facebook, Instagram, LinkedIn have become a crucial part of each one's life (Albayati and Altamimi 2019). It radically impacts daily human social interactions where users and their communities are the base for online growth, commerce, and information sharing. Different social networks offer a unique value chain and target different user segments. For instance, Twitter is known for being the most famous microblogging

✉ Malak Aljabri
mssjabri@uqu.edu.sa

Rachid Zagrouba
rmzagrouba@iau.edu.sa

Afrah Shaahid
2190009057@iau.edu.sa

Fatima Alnasser
2190003750@iau.edu.sa

Asalah Saleh
2160007924@iau.edu.sa

Dorieh M. Alomari
2180007089@iau.edu.sa

[1] Department of Computer Science, College of Computers and Information Systems, Umm Al-Qura University, Makkah 21955, Saudi Arabia

[2] SAUDI ARAMCO Cybersecurity Chair, Department of Computer Science, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

[3] SAUDI ARAMCO Cybersecurity Chair, Department of Computer Information Systems, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

[4] SAUDI ARAMCO Cybersecurity Chair, Department of Computer Engineering, College of Computer Science and Information Technology, Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam 31441, Saudi Arabia

social network for receiving rapid updates and breaking news. While Instagram usage is mainly by celebrities and businesses for marketing (Meshram et al. 2021). Whereas professional communities use LinkedIn. As social networks' popularity grows combined with the availability of vast personal information that users share makes the same valuable features of social platforms for ordinary people a tempting target for malicious entities (Adikari and Dutta 2020). The most prevalent form of malware on social media networks is thought to be bots (Aldayel and Magdy 2022; Cai, Li, and Zengi 2017b). Some bots are benign. However, the majority of bots are utilized to perform malicious activities such as fabricating accounts, faking engagements, social spamming, phishing, and spreading rumors to manipulate public opinion, such activities not only disturb the genuine users' experience but also lead to a negative effect on the public's and individual's security. As a result, in recent years, researchers have dedicated a significant amount of attention to social media bot detection (Ali and Syed 2022; Ferrara 2018; Rangel and Rosso 2019; Yang et al. 2012) and prevention (Thakur and Breslin 2021).

1.1 Social media platforms

OSNs have revolutionized communication technologies and are now an essential component of the modern web. The most popular social networks globally as of January 2022 are shown in Fig. 1, ordered by the number of monthly

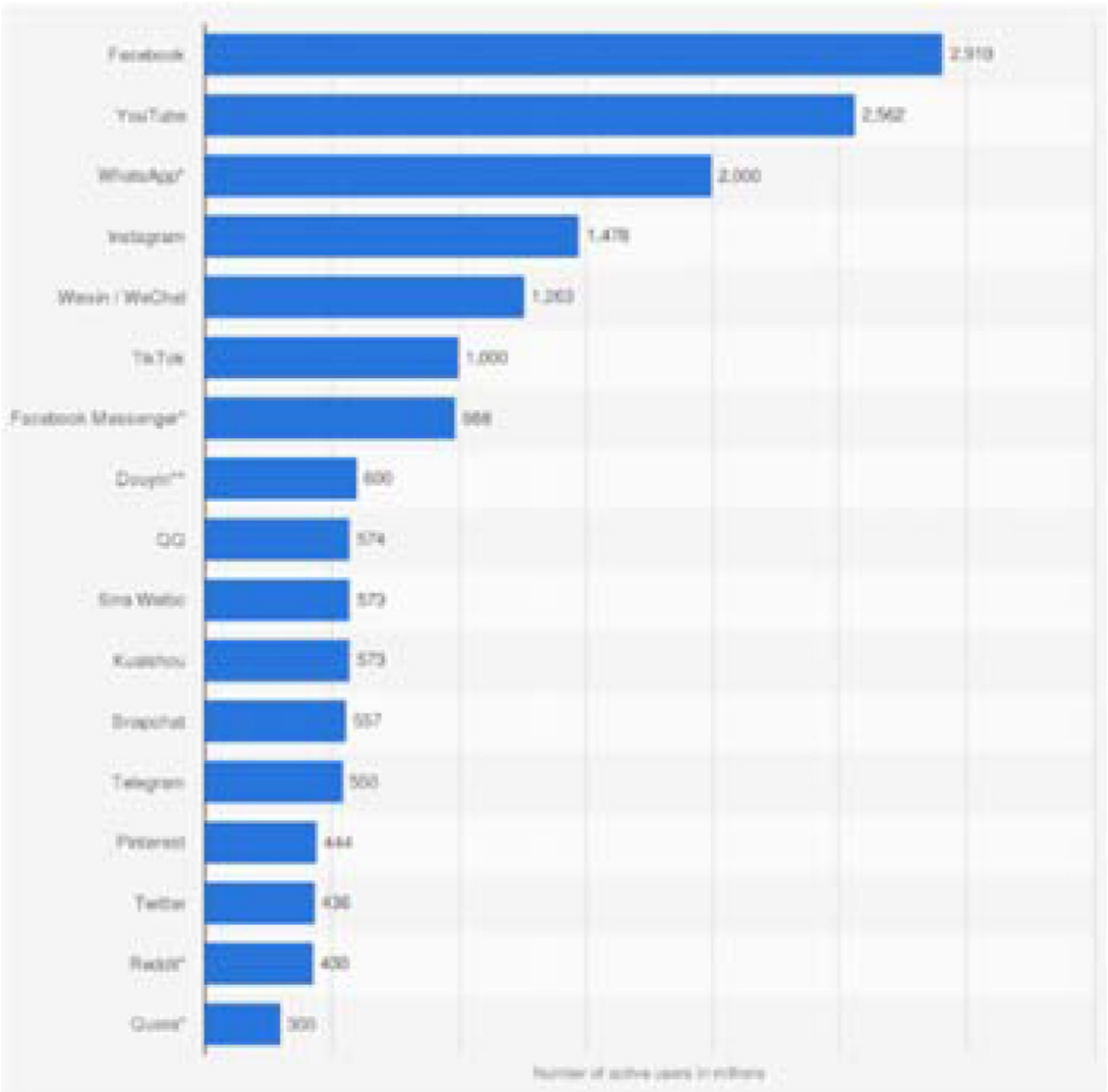

Fig. 1 Most popular social networks globally as of January 2022, ordered by no. of monthly active users.https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

active users in millions. The social media platforms which are included in the scope of our study are namely Twitter, Facebook, Instagram, LinkedIn, and Weibo. On these platforms, user growth and popularity have been increasing at an exponential rate. These platforms enable users to produce and exchange user-generated content (Kaplan and Haenlein 2010). For instance, only 2.375 billion people were using Facebook in the first quarter of 2019 (Siddiqui 2019), thereby representing one-third of the world population (Caers et al. 2013). One of the most widespread and extensively used OSN by people from all walks of life is Twitter. Twitter allows individuals to express their sentiments on different topics such as entertainment, the stock market, politics, and sports. (Wald et al. 2013). It is one of the fastest means of circulating information as a result extremely affects people's perspectives. Over the past few years, Twitter has become a replacement for mainstream media for obtaining news (Wald et al. 2013). On the other hand, Instagram is an OSN for sharing photos and videos and is accessible on both Android and iOS since 2012. Dated May 2019, there were more than a billion users registered on Instagram, according to collected data (Thejas et al. 2019). Moving on, Facebook is an online social networking site that makes it convenient for people to connect and share with family and friends. It was developed in 2004 initially for students by Mark Zuckerberg. With more than 1 billion users globally, Facebook is one of the biggest social networks in the current times (Santia et al. 2019). One of the most well-known professional social networks is LinkedIn, a platform that focuses on professional networking and career advancement (Dinath 2021). Sina-Weibo, also known as Chinese Twitter, was launched in 2009, and this microblogging website or application is one of China's biggest social media platforms. It offers a plethora of features which include posting images, instant messaging, Weibo stories, using location-based hashtags, trending topics, etc. Furthermore, it also gives businesses the privilege to set up accounts for the purpose of advertisements and services (Tenba Group 2022).

1.2 Social media security

Security and trustworthiness among users, service providers, platform owners, and third-party supervisors are critical factors for social media platforms' success and stable existence (Zhang and Gupta 2018). According to recent surveys (Shearer and Mitchell 2022), a considerable segment of the population prefers social networks to TV, newspapers, and other traditional media when looking for information. Trust in social networks as a source of information is predicted to rapidly grow (Kolomeets and Chechulin 2021). As a result, social bots can pose significant security risks by influencing public opinion and disseminating false information (Shao et al. 2017), spreading rumors and conspiracy theories (Ferrara 2020), creating fake reputations, and suppressing political competitors (Pierri et al. 2020; Benkler et al. 2017). Despite the fact that bots are extensively used, little research has been done to examine how they affect the social media environment. This indicates that nearly 48 million of the accounts on Twitter are bots (Sheehan 2018). It was also stated that Facebook acknowledges that 270 million of its accounts are fake (Sheehan 2018). Further, there is evidence that social media bots were utilized to attempt to influence political communication dates during the US midterm elections in 2010. There were also allegations that social bots on Twitter played a significant role in the 2016 US presidential election (Cresci et al. 2017; Mahesh 2020; Sedhai and Sun 2015). Bots can be employed to spread misinformation to promote a particular view of a public person, grow an account's following, and repost user-generated content. Bot detection on OSNs is therefore the most frequently requested security feature from businesses and law enforcement organizations (Kolomeets and Chechulin 2021). The dearth of publicly accessible datasets for OSNs such as Facebook, Instagram, and LinkedIn is one of the greatest obstacles in this research area. Unlike Twitter, this restriction results from some of these OSNs' limited data collection policies.

1.3 Types of bots on social media

The term "bot" refers to a robot, a computer program that works more quickly than humans at recurring, automated tasks. More precise terminology can be used to define bots in OSNs "a computer software that generates content automatically and engages with users of social media to replicate and possibly modify their behavior" (Benkler et al. 2017). Bots can be used for useful or harmful reasons and often replicate human behavior to some degree (Fonseca Abreu et al. 2020). Good bots can significantly reduce the need for human customer service representatives for some businesses, such as chatbots and news bots that automatically upload new articles or news for journalists or bloggers. Bots can be employed for negative as well as positive purposes. According to (Gorwa and Guilbeault 2020), bots are responsible for a sizable portion of online activity, are used to manipulate algorithms and recommender systems, stifle or promote political speech, and can be crucial in the spread of hyper-partisan "fake news." According to (Benkler et al. 2017), there are four different categories of social media bots: spambots, social bots, sybil bots, and cyborgs. Promoter bots, URL spambots, and false followers are only a few examples of the various types of spambots that spread harmful links, uninvited messages, and hijack popular subjects on social networks (Meshram et al. 2021). On the other hand, social bots are algorithmically controlled user accounts that mimic the activity of human users but carry out their tasks at a considerably faster rate while successfully

concealing their robotic identity (Ferrara 2018). While cyborgs bots are half-human, half-bot accounts that exist between people and bots, sybil bots are anonymous identities, i.e., user accounts, utilized for a significantly big effect (Gorwa and Guilbeault 2020). In this review, the collected papers were incorporating three categories of bots which are social bots spambots, and sybil bots.

1.4 The different machine learning-based techniques and algorithms

The development of algorithms that allow a computer to learn on its own from data and prior experiences is the core of ML, a subfield of artificial intelligence (AI). Arthur Samuel was the first to originate the term "Machine Learning" (Wiederhold and McCarthy 1992). ML system develops prediction models based on previous data and makes predictions up until new data are gathered. The amount of data used to create a model determines its accuracy (Saranya Shree et al. 2021). The various types of ML techniques include Supervised, Semi-supervised, Unsupervised, and Reinforcement. However, we have only included Supervised, Semi-supervised, and Unsupervised as a part of our study. Three types are present under supervised category: Classification, Regression, and Forecasting. Some of the most popular supervised algorithms include Random Forest (RF), Naïve Bayes (NB), Decision Trees (DT), Logistic Regression (LR), Support Vector Machine (SVM), Neural Networks (NN), and many more. Deep learning (DL) is a subset of supervised ML techniques that employs multiple layers to gradually extract higher-order features from the input data. In order to create patterns and process data, this AI technology mimics the actions and processes of the human brain (Gannarapu et al. 2020). Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs), Generative Adversarial Networks (GANs), etc., are some of the well-known DL algorithms. Whereas unsupervised learning algorithms are namely categorized into Clustering and Association which mainly deal with unlabeled data. Some of the primarily used ones include K-nearest Neighbor (KNN), K-means clustering, Principal Component Analysis (PCA), etc. However, a small amount of labeled data and a large amount of unlabeled data are utilized in semi-supervised learning, which results in a hybrid of supervised and unsupervised learning (Mahesh 2020).

1.5 Machine learning implementation on social media security

The way people use social media is evolving as a result of the proliferation of ML techniques in social media and the increased sophistication of cyberattacks on computer information systems (Aljabri et al. 2021a, b). On social networking sites such as Facebook and Twitter, numerous ML techniques were employed. For instance, existing ML algorithms can determine the user's location, carry out sentiment analysis, (Aljabri et al. 2021a, b) offer recommendations, and much more. Diverse ML methods have been successfully deployed to address wide-ranging problems in cybersecurity which include detecting malicious URLS, classification of firewall log data, phishing attacks detection, etc. (Aljabri et al. 2022a, b, c; AAljabri and Mirza 2022). However, malicious software can be used to target social media platforms and carry out cyber-attacks. In terms of social media, several efforts have been undertaken to investigate ML techniques to detect such types of malware. For instance, (Alom et al. 2020) detected Twitter spammers using DL techniques. Moreover, the study conducted by (Kantartopoulos et al. 2020) addressed the effects of hostile attacks and utilized KNN as a measure to tackle the problem. The authors presented a methodology that uses SVM and Ensemble algorithms to effectively detect cyberbullying (Gupta and Kaushal 2017). Additionally, models have been developed for social media systems' access control (Carminati et al. 2011). Yet, bot and fake account detection on social media platforms are still one of the primary challenges for cyber security researchers (Thuraisingham 2020).

1.6 Key contributions

This section firstly puts forth the existing literature reviews done on different social media platforms as shown in Table 1. It also briefly discusses the previously used taxonomies along with the prevailing gaps. Starting with the literature review on Twitter, (Alothali et al. 2019) included literature concerning from 2010 to 2018 based on various techniques which include Graph-based, Crowdsourcing, and ML. They analyzed the common aspects such as datasets, classifiers, and the selected features employed. The challenges present in the domain were also addressed. (Derhab et al. 2021) discussed existing techniques and put forth a taxonomy that addressed the state-of-the-art tweet-based bot detection techniques in the timeline from 2010 to 2020. Based on tweet-based bot detection techniques, they provided the main features utilized. For tweet-based bot detection, they also described big data analytics shallow and DL techniques, in addition to their performance results. Finally, the challenges and open issues in the area of tweet-based bot detection were presented and discussed (Derhab et al. 2021). Furthermore, (Orabi et al. 2020) discussed the studies from 2010 to 2019 on Graph-based, ML-based, Crowdsourcing, and Anomaly-based. Their research revealed some gaps in the literature, such as the fact that studies discussed mainly Twitter, and that unsupervised ML is rarely used, in addition to the majority of publicly available datasets being either inaccurate or

Table 1 Summary of existing literature reviews

References	Range of papers reviewed	Taxonomy	Open issue/ future discussion
Alothali et al. (2019)	2010–2018	Graph-based, crowdsourcing, and machine learning	✓
Derhab et al. (2021)	2010–2020	Shallow learning-based (supervised learning, semi-supervised learning, and unsupervised learning)	✓
		Deep learning-based-Deep learning-based	
Orabi et al. (2020)	2010–2019	Graph-based, machine-learning based, crowdsourcing, anomaly-based	✓
Gheewala and Patel (2018)	2010–2017	Clustering algorithms, classification algorithms, hybrid	✗
Ezarfelix et al. (2022)	2018–2021	Logistics regression, naive bayes, random forest, support vector machine	✓
Rao et al. (2021)	2015–2020	URL list-based spam filtering techniques, honeypot/honeynet-based techniques, machine learning and deep learning techniques	✓

insufficiently large. In (Gheewala and Patel 2018), contributed a review on ML twitter spam detection for the years 2010–2017 based on Clustering, Classification, and Hybrid algorithms. Some of the issues concluded were regarding the results being lowered as a result of concerns with feature fabrication, class imbalance, spam drift, etc., for spam detection. The study (Ezarfelix et al. 2022) performed was based only on the Instagram platform where a multitude of analyses, and evaluations have been performed on the studies from 2018 to 2021. It was concluded that in order to detect fake accounts, using NN is the most effective method. (Rao et al. 2021) presented a comprehensive review of the social spam detection techniques studied from 2015 to 2020 based on different social spam detection techniques which include Honeypot/Honeynet-based techniques, URL List-based spam filtering techniques, and ML and DL techniques. Numerous feature analysis and dimensionality reduction techniques used by different researchers were outlined. A thorough analysis was given, describing the datasets utilized, features used, ML/DL models used, performance measures used, and pros and cons of each model.

To the best of our knowledge, no study in the literature has carried out a comprehensive analysis of the existing studies in the time period (2015–2022) in the domain of applying ML-based techniques for social media bot detection (social bots, spambots, sybil bots). For this specific timeline, the existing reviews have studied either only ML or DL-based studies or only addressed a specific bot type. We perceived that there was a need for a recent literature review to be conducted so that researchers could identify the findings and gaps in this field and use that information as a roadmap for future research directions and further in-depth study. In response to this demand, in this study, we discuss what is currently known and being researched regarding the several concepts, theories, and techniques linked to bot detection on social media platforms.

This paper makes the following key contributions:

- Provide summaries and analysis of the used ML-based (supervised, semi-supervised, and unsupervised) classification techniques to detect various types of bots on some particular social media platforms.
- Provide a unique taxonomy based on the various ML-based techniques which has not been provided in the existing literature.
- Identify and analyze the most commonly extracted and used features on each social media platform.
- Study the most affected social media platform from malicious bots, the class of bots mostly found on these platforms. Additionally, highlight the most studied social platforms and analyze the gaps of research on other platforms.
- Examine and analyze the popular public datasets used for each platform and the methods used for the self-collected datasets.
- Highlight challenges and gaps in existing research thereby providing potential directions for further research.

The rest of this paper is structured as follows: Sect. 2 presents the methodology adopted for this paper. Section 3 puts forth a detailed analysis including tables and figures demonstrating the ML-based techniques used in the existing literature. In Sect. 4 based on all the reviewed studies, the datasets used, features extracted, and algorithms implemented are discussed thereby performing an extensive analysis. Section 5 sheds light on the insights gained and presents a discussion on the challenges and opportunities in existing research thereby providing future research directions. Section 6 provides a conclusion to summarize our literature review.

2 Methodology

The objective of this review is to study the existing literature from 2015 to 2022 in the domain of bot detection and classification using ML techniques on various social media platforms. We searched for social media bot detection-related papers on various well-known databases mainly Google Scholar, Mendeley, IEEE Xplore, ResearchGate, ScienceDirect, Elsevier, acm.org, arxiv.org, SpringerLink, MDPI, etc. The total number of papers reviewed were 105. All these 105 papers were summarized and elaborately discussed in this paper. Figure 2 demonstrates the range of the reviewed papers.

Figure 3 shows the created taxonomy for the paper. The first tier is based on ML-based techniques, followed by the second tier on the type of social media platform and lastly, the third tier is based on the type of social media bot which includes social bots, spambots, and sybil bots. The logic behind the taxonomy created in this literature review is mainly to identify the most effected social media platforms from bots, the class of bots mostly found on those platforms, and to highlight the most studied social platforms and analyze the gaps of research on other platforms. This is different from most existing literature reviews which focus on the ML techniques and algorithms used in research which can be inefficient to highlight findings and identify gaps since many studies use several techniques and algorithms applied on one platform.

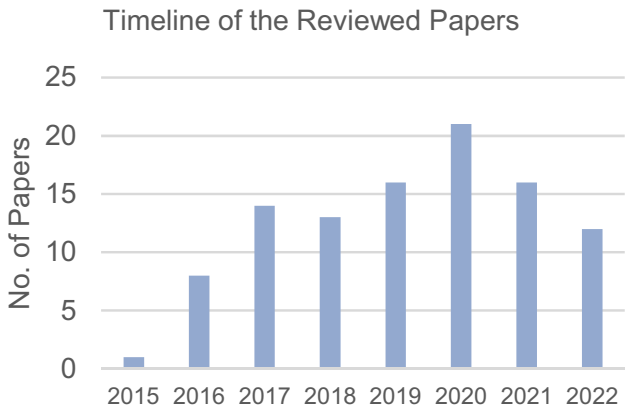

Fig. 2 Bar chart showcasing the range of the reviewed papers

3 Machine learning-based techniques for detecting bots on social media platforms

Numerous studies have been published addressing the use of ML-based techniques for bot detection. This section reviews the existing research studies on the subject by discussing previous studies and findings. The summaries are organized based on the three different ML types followed by different social media platforms and the affecting bot types.

3.1 Using supervised ML

Most of the studies we reviewed have implemented supervised ML and DL to detect social bots, spambots, and sybil bots which shall be discussed below.

3.1.1 Facebook—detecting social bots

Very few studies were found that used the supervised approach to detect social bots on Facebook. To improve classification accuracy, (Wanda et al. 2020) built a supervised learning architecture using a CNN model. To train and evaluate the model, the CNN used a Deep Neural Network (DNN) with a number of hidden layers. In order to minimize the objective function using the model's parameters, it also used a gradient descent. To optimize and accelerate training time in the NN, a pooling layer was used. The results with an optimizer Stochastic Gradient Descent (SGD) $m = 0.5$ showed a training loss of 0.5058 and a testing loss of 0.5060.

Secondly, 4.4 million publicly generated Facebook postings were collected and described in a dataset by (Dewan and Kumaraguru 2017). On their dataset of harmful posts, they used two different filtering techniques: one that used URL blacklists and another that used human annotations. They used NB, DT, RF, and SVM models, among other supervised learning methods. These models are based on a set of 44 publicly accessible attributes. After evaluation, RF was shown to have the highest accuracy of over 80%. Based on their findings, they proceed to develop Facebook Inspector (FBI), a browser plug-in that uses a Representational State Transfer (REST) API to identify harmful Facebook postings in real-time.

3.1.2 Facebook—detecting spambots

Some studies have identified spambots on Facebook using various data collection techniques. Due to the restrictive security policies on Facebook, accessing and acquiring relevant data is challenging.

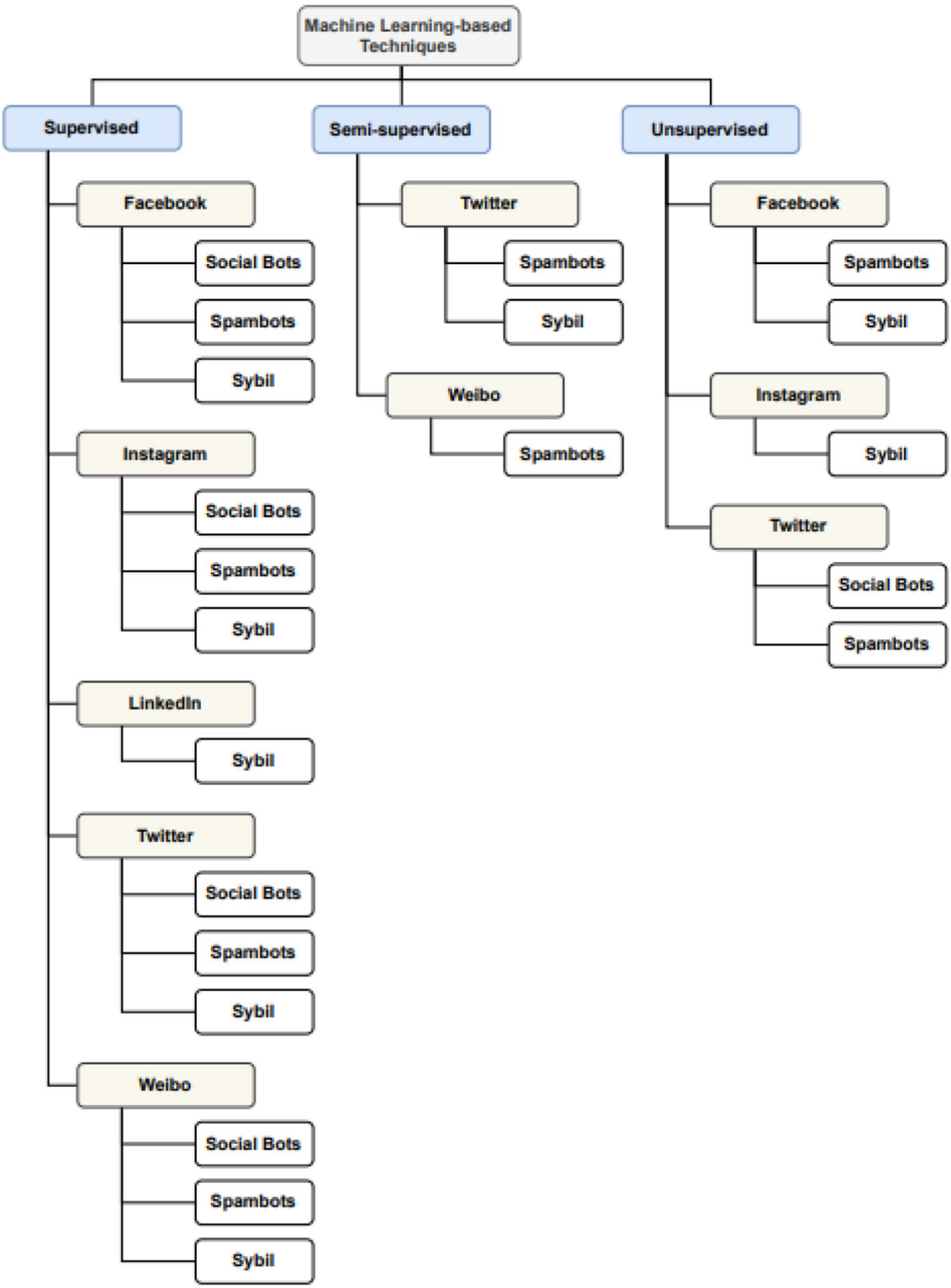

Fig. 3 Taxonomy of social media bot detection using ML-based techniques

Sahoo and Gupta (2020) implemented a spammer detection system on Facebook. The Particle Swarm Optimization (PSO) algorithm was used in this study to determine the popularity of the content and feature selection. The dataset included 1600 profile posts in total. Twelve profile- and content-based features were chosen after the generated content underwent data pre-processing. The PSO algorithm used these features as an input parameter to find fraudulent accounts. In this experiment, classifiers RF, RT, Bagging, JRip, J48, and AdaBoost were utilized. Using the classifier, the detection rate produced the best accuracy of 99.5%.

Followed by, (Rathore et al. 2018) who introduced an efficient spammer detection method called SpamSpotter that uses an Institute for Data, Systems, and Society (IDSS) to differentiate spammers from real Facebook users. A dataset made up of 1000 profiles was employed. The framework made use of features based on profiles and content. They used the Bayesian Network (BN), RF, Decorate (DE), J48, JRip, KNN, SVM, and LR as the eight supervised ML classifiers. The BN classifier outperformed all others with an accuracy of 0.984.

3.1.3 Facebook—detecting sybil bots

We found a reasonable number of studies that thrived in recognizing sybils (Fake profiles) on Facebook. This study proposed by (Albayati and Altamimi 2019) was about a smart system known as FBChecker that checks if a profile is fake. A set of behavioral and informational attributes were analyzed and classified by the system using the data mining approach. Four data mining algorithms which include KNN, DT, SVM, and NB were used. The RapidMiner data science platform was used to implement the selected classifiers. The dataset of 200 profiles was prepared by the authors. A Receiver Operating Characteristic Curve (ROC) graph comparison was created to check the accuracy and all classifiers showed a high accuracy rate, but SVM outperformed with an accuracy rate of 98%.

Subsequently, (Hakimi et al. 2019) proposed supervised ML techniques based on only five characteristics that play a key role in distinguishing fake and true users on Facebook. The important characteristics finalized were Average Post Likes Received, Average Post Comments, Average Post Comments Received, Average Post Liked, and Average Friends. A sample data of 800 users were generated by Mockaroo. The data were categorized into four clusters: Inactive User, Assume Fake account User, Fake account user, and Real User. Classifiers namely KNN, SVM, and NN were implemented. Results showed that KNN outperformed with an accuracy of 0.829. It was concluded that the features "likes", and "remarks" add a significant value to the job of detection.

Moreover, (Singh and Banerjee 2019) created a dataset on Facebook using their graph API to be utilized for sybil accounts detection. Also, a comparative analysis of various algorithms over the dataset was performed. The dataset contained 995 both real and fake accounts. Twenty-nine features were extracted including textual, categorical, and numerical features. AdaBoost, Bagging, XGBoost, Gradient Boost (GB), RF, LR, Support Vector Classifier (LinearSVC), and ExtraTree algorithms were applied for evaluation. AdaBoost was the best-performing algorithm with a 99% F1-score.

However, (Saranya Shree et al. 2021) suggested Natural Language Processing (NLP) pre-processing techniques and ML algorithms such as SVM and NB to classify fake and genuine profiles on Facebook. A dataset of 516 profiles was used and trained until 30 epochs. It predicted 91.5% fake accounts and 90.2% genuine accounts correctly.

Another strategy for identifying sybils on Facebook was presented by (Babu et al. 2021). By using the Facebook graph API, they gathered a dataset of 500 users from a survey of 500 Facebook users in order to better understand the nature and distinguishing characteristics of sybil. The tested dataset was used to identify fake profiles using the NB classifier. Seven profile-based features were used in the model. Their suggested solution had a 98% efficiency rate. Moving on, (Gupta and Kaushal 2017) has described an approach to detect fake accounts. The key contributions of the authors' work include a collection of a private dataset using the Facebook API through Python wrappers. After data collection, a set of 17 features was shortlisted which included likes, comments, shares, tag, apps usage, etc. A total of 12 supervised ML classification algorithms were used (from Weka), namely, k-Nearest Neighbor, Naive Bayes, Decision Tree classifiers (J48, C5.0, Reduced Error Pruning Trees Classification (REPT), Random Tree, Random Forest), etc. Two types of cross-validation were performed, namely, the holdout method, and tenfold cross-validation. A classification accuracy of 79% was achieved. The user activities contributed the maximum to the detection of fake accounts.

3.1.4 Instagram—detecting social bots

Only one study by (Sen et al. 2018) aimed to detect fake likes on Instagram thereby detecting social bots. A dataset of 151,117 likes of both fake and genuine likes was captured and labeled manually by the authors. A limitation of this study was the noisiness of the dataset. However, various types of features were extracted from the dataset, which were Network Effect, Internet Overlap, Liking Frequency, Influential Poster, Hashtag Features, and User-based features to be used with extensive analysis. LR, RF, SVM, AdaBoost,

XGBoost, NN, and Multilayer Perceptron (MLP) algorithms were applied. MLP showed the best results with 83% Precision and 81% Recall (AUC of 89%). According to the authors, the model's high efficacy in capturing the parameters that influence genuine liking behavior is the model's main strength.

3.1.5 Instagram—detecting spambots

Two studies were found that used the ML approach for fake and automated accounts detection on Instagram. Firstly, (Akyon and Esat Kalfaoglu 2019) contributed by generating two labeled public datasets. A dataset for fake accounts (1203 accounts) and another for bots (1400 accounts). However, both datasets had problems. The fake accounts dataset had an uneven number of real and fake accounts. As a result, the Synthetic Minority Over-sampling Technique-for-Nominal and Continuous (SMOTE-NC) algorithm was implemented. While cost sensitive genetic algorithm was implemented to correct the automated accounts dataset unnatural bias. Profile-centric features were fed into NB, LR, SVM, and NN algorithm. SVM and NN provided promising F1-scores for both datasets. 94% with oversampling for fake accounts and 86% for automated accounts dataset.

Similarly, a method to identify spam posts was also presented by (Zhang and Sun 2017). 1983 user profiles and 953,808 media posts made up a manually labeled dataset. Profile-based, Color Difference Histogram-based, and Media Post-based feature vectors were extracted from user profiles and media postings. The near duplicate posts were grouped into the same clusters using two-pass clustering techniques, Minhash clustering and K-medoids clustering. The best pair has an accuracy of 96.27%: RF, (maxDepth: 8, numTrees: 20, impurity: entropy).

3.1.6 Instagram—detecting sybil bots

Many studies were able to detect sybil bots starting with (Meshram et al. 2021) proposed an automated methodology for fake profiles detection. The authors collected 1203 accounts including real and fake accounts using Instagram API. In addition, a list of eight content- and behavior-based features were extracted. Authors needed to oversample the dataset using SMOTE-NC before applying any algorithm due to the unevenness of the real-fake accounts ratio. Afterward, NN, SVM, and RF algorithms were applied. RF depicted the best-performing results with an accuracy of 97%.

Whereas, using the same records and features, (Sheikhi 2020) presented a bagging classifier and performed a comparative analysis with five well-known ML algorithms, which were RT, J48, SVM, Radial Basis Function (RBF), MLP, Hoeffding Tree, and NB with 10-cross-validation. The bagging classifier showed better performance by successfully classifying 98% of the accounts. Moreover, the author presented the best feature types for different sizes of datasets.

Additionally, (Dey et al. 2019) also assessed fake and real different Instagram accounts. A publicly labeled dataset of sixteen accounts was obtained from Kaggle. Twelve profile-based features were extracted from the sample dataset. Missing Value Treatment, OuSybiltlier Detection, and Bivariate Analysis were carried out as a part of the Exploratory Data Analysis. Median imputation was done to deal with the outliers. For the extent of this paper, LR, and RF—two supervised classification algorithms were used. Lastly, out of the two mentioned classifiers, RF showed the best performance with 92.5% accuracy.

Subsequently, the research of (Purba et al. 2020) aimed to identify fake users' behavior. Furthermore, different approaches of classification have been proposed. 2-class (authentic, fake) and 4-classes (authentic, spammer, active fake user, inactive fake user) classifications. The total number of fake and authentic users in the dataset was 32,460 users. They used seventeen features based on metadata, media info, media tags, media similarity, and engagement. Using these features with RF, MLP, LR, NB, and J48 algorithms showed promising results. RF showed an accuracy of up to 91.76% for 4-classes classification. Moreover, analysis outcomes showed that metadata and statistics results are the foremost predictors for classification.

Nevertheless, (Kesharwani et al. 2021) utilized a six-layered DL model NN to classify fake and genuine Instagram accounts. The designed model used 12 profile-based features. An open dataset of 696 Instagram users available on Kaggle was used for this experiment and was collected using a crawler. The dataset had 10 profile-based features. The model's training was done using 20 epochs and therefore giving an accuracy of 93.63%.

Quite interestingly, (Bazm and Asadpour 2020) proposed a behavioral-based model. A labeled dataset was collected by the authors including 2000 accounts of both fake and genuine users. Seven behavioral features were extracted from the dataset. KNN, DT, SVM, RF, and AdaBoost algorithms were tested and analyzed. AdaBoost showed the best-performing results with an accuracy of 95%. Additionally, the Max feature was identified as the most effective for classification followed by standard deviation, following count, and entropy. Three of the above-mentioned most effective features were behavioral.

Lastly, the work of (Thejas et al. 2019) also focused on detecting valid and fake likes of Instagram posts by applying automated single and ensembled learning models. A labeled dataset of 10,346 observations and 37 features has been composed. The authors used numeric features and

text-based features to perform extensive analysis of fake likes related patterns. Various single classifiers have been used such as LR SVM, KNN, NB, and NN with different versions. Adjacent to ensembled-based classifiers as RF with multiple versions as well. Moreover, bot detection using an autoencoder has been experimented. RF showed the highest performance among all with 97% accuracy.

3.1.7 LinkedIn—detecting sybil bots

Only two studies were found on this platform to detect sybils, (Adikari and Dutta 2020) proposed a methodology for identifying bot-generated profiles based on limited publicly available data of profiles using data mining techniques. Many existing research assumes the availability of static and dynamic data of a profile, which is not the case with LinkedIn as it has more restrictive privacy policies that impede access to dynamic data. The profile features were extracted from a dataset of 74 profiles only. Thirty-four fake accounts were collected by searching blogs and websites for known LinkedIn fake accounts. The lack of verified fake accounts was a limitation of this research. NN, SVM, PCA, and Weighted Average algorithms were used in several combinations for detecting fake profiles. SVM showed the highest accuracy (87.34%) when employing PCA-selected features with a polynomial kernel.

Furthermore, (Xiao et al. 2015) proposed a scalable offline framework using the pipeline to identify clusters of fake accounts on LinkedIn. Cluster-level fake accounts are identified rather than account-level to detect fake accounts after registration rapidly. Statistical features generated by users at or after registration time, such as name, email address, company, were grouped into clusters. Cluster-level features were exclusively fed into the RF, LR, and SVM models. The authors have collected a set of labeled data for 260,644 LinkedIn accounts. RF algorithm's performance evidently provided the best results for all metrics; an AUC of 0.95 and a recall of 0.72 at 95% precision for out-sample test data.

3.1.8 Twitter—detecting social bots

Numerous studies were able to detect social bots on Twitter starting (Echeverrï£¡a et al. 2018) tested 20 unseen bot classes of varying sizes and characteristics using bot classifiers. Two datasets were collected using Twitter's API consisting of 2.5 million accounts. Twenty-nine Profile- and Content-based features were employed for classification. The classifiers used to test were GB Trees (XGBoost and LightGBoost Model (LGBM)), RF, DT, and AdaBoost. LGBM showed the highest accuracy rate of 97.84% on both the subsampling used—C30K and C500.

Moreover, (Fonseca Abreu et al. 2020) examined whether feature set reduction for Twitter bot detection yields comparable outcomes to large sets. Five Profile-based features were used for classification. The dataset used consisted of 4565 records of both social bots and genuine users. The ML algorithms tested namely were RF, SVM, NB, and one-class SVM. AUC's greater than 0.9 were obtained by all multi-class classifiers. However, RF exhibited the best results with an AUC of 0.9999.

Varol et al. (2017) used more than a thousand features which were based on metadata primarily based on friends, tweet content, sentiment, network patterns, and activity time series. A publicly accessible dataset of size 31 K that contains manually verified Twitter accounts as bots or real was used to train the model. The model's accuracy was evaluated using RF, AdaBoost, LR, and DT classifiers. The best performance was depicted by RF of 0.95 AUC. Furthermore, it was concluded that the most significant sources of data are user metadata and content features.

Twenty-eight features were extracted based on profile, tweets, and behavior (Knauth 2019). For easy future portability, language-agnostic features were mainly focused on. LR, SVM, RF, AdaBoost, and MLP classifiers were used for experiments. AdaBoost outperformed all competitors with an accuracy of 0.988. Smaller quantities of training data were analyzed, and it was shown that using a few, expressive characteristics provides good practical benefits for bot identification.

In this study, after a long process of feature extraction and data pre-processing, (Kantepe and Gañiz 2017) employed ML techniques. Thousand eight hundred accounts were used to get the data from Twitter API and Apache Spark, which was In this study, after a long process of feature extraction and data pre-processing, (Kantepe and Gañiz 2017) employed ML techniques. One thousand eight hundred accounts data was obtained with Twitter API and Apache Spark, which was then used to extract 62 different features. The features extracted were mainly profile-based features, Twitter features and periodic features. Four classifiers were used which include LR, Multinomial Naïve Bayes (MNB), SVM and GB. The highest accuracy result 86% was shown by the GB trees.

This research conducted by (Barhate et al. 2020) used two approaches for the detection of bots and analyzed their influence in trending a hashtag on Twitter. First, the bot probability of a user was calculated using a supervised ML technique and a new feature bot score. A total of 13 features were extracted for data pre-processing and Estimation of Distribution Algorithms (EDA). The data were trained using RF classifier, which produced an AUC result of 0.96. This study also came to the conclusion that bots had a high friend-to-follower ratio and a low follower growth rate.

The dataset that was acquired by (Pratama and Rakhmawati 2019) is from the supporters of the Indonesian presidential candidate on Twitter. The top five hashtags for each candidate were used to collect tweets, which were then manually labeled with the accounts' bot characteristics, resulting in a limit of about 4.000 tweets. SVM and RF, two ML models, are utilized for bot detection. These two models were trained with cross-validation ten-folds to improve the overall score. From these two models, RF has a higher overall score than SVM of 74% in F1-Score, Accuracy, and AUC. Comparing the 10 retrieved features from the dataset, they discovered that the account year creation had the biggest separation between humans and bots.

Davis et al. (2016) made use of RF classifier to evaluate and detect social bots by creating a system called BotOrNot. A public dataset of 31 K accounts was used to train the model. From six main groups of characteristics—network, user, friend, temporal, content, and sentiment features—the framework collected more than 1000 features. These various classifiers—one for each category of features and one for the overall score—were trained using extracted features. The system performance was assessed using ten-fold cross-validation, and an AUC value of 95% was obtained.

Likewise, a Twitter bot identification technique was also presented by (Shevtsov et al. 2022). 15.6 million tweets 'total, including 3.2 million accounts sent during the US Elections, were included in their dataset from Twitter. The XGBoost algorithm was used to pick 229 features from approximately 337 user-extracted features. Their suggested ML pipeline involves training and validating many three ML models which are SVM, RF, and XGBoost. Performance was best for XGBoost where their findings indicate that it performs well on the collected dataset compared to the training data section because of its great generalization capabilities. Only 2% of the F1 score is going from 0.916 to 0.896, and 0.03% of the ROC-AUC indicates a decline in performance from 0.98 to 0. 977.

Additionally, SPY-BOT, a post-filtering method based on ML for social network behavior analysis, was introduced by (Rahman et al. 2021). Six hundred training samples were used to extract eleven characteristics. They contrast the two ML algorithms LR and SVM throughout the training phase. After comparing outcomes, tuned SVM was the best performing. On the validation dataset, their method achieves up to 92.7% accuracy while up to 90.1% accuracy was obtained on the testing dataset. As result, they suggest that the proposed approach able to classify the users' behavior in Social Network-Integrated Industrial Internet of Things (SN-IIoT).

Also, a real-time streaming framework called Shot Boundary Determination (SBD) was also suggested by (Alothali, Alashwal, et al. 2021a) as a way to detect social bots before they launch an attack to protect users. To gather tweets and extract user profile features, the system uses the Twitter API. They used a publicly available Twitter dataset from Kaggle, which has a total of 37,438 records, as their offline dataset. Friends count, Followers count, Favorites count, Status count, Account age days, and Average tweets per day were the six features that were extracted and further used as input to their ML model. They use RF algorithm to differentiate between the bots and human accounts. The outcomes of their methodology demonstrated the effectiveness of retrieving, publishing the data, and monitoring the estimates.

Shukla et al. (2022) proposed a novel AI-driven multilayer condition-based social media bot detection framework called TweezBot. Moreover, the authors have performed a comparative analysis with several existing models and an extensive study of features, and exploratory data. The proposed method analyzed each Twitter-specific user profile features and activity-centric characteristics, such as profile name, location, description, verification status, and listed count. 2789 distinct user profiles were used to extract these features from a public labeled dataset from Kaggle. ML models used for comparative evaluation and analysis were RF, DT, Bernoulli Naïve Bayes (BNB), CNB, SVC, and MLP. TweezBot attained a maximum accuracy of 99.00049%.

Since bots are used to manipulate activities in politics as well (Fernquist et al. 2018) presented a study on political Twitter bots and their impact on the September 2018 Swedish general elections. To identify automatic behavior, an ML model that is independent of language was developed. The training data consist of both bots and genuine accounts. Three different datasets (Cresci et al. 2015; Gilani et al. 2017; Varol et al. 2017) were used to train the classification model. Furthermore, a list of 140 user metadata, Tweet and Time features were extracted. Various algorithms such as AdaBoost, LR, SVM, and NB were tested. RF outperformed with an accuracy of 0.957.

Similarly, (Beğenilmiş and Uskudarli 2018) made use of collective behavior features in hashtag-based tweet sets, which were compiled by searching for relevant hashtags. A dataset of 850 records was utilized to train the model using algorithms including RF, SVM, and LR. From tweets collected during the 2016 US presidential election, 299 features were retrieved. To capture the coordinated behavior, the features represent user and temporal synchronization characteristics. These models were developed to distinguish between organic and inorganic, political and non-political, and pro-Trump or pro-Hillary or neither tweet set behavior. The RF displayed the best outcomes, with an F-measure of 0.95. In conclusion, this study found that media utilization and tweets marked as favorites are the most dominant features and user-based features were the most valuable ones.

On the other hand, in this approach, (Rodríguez-Ruiz et al. 2020) one-class classification was suggested. One

benefit of one-class classifiers is that they do not need examples of abnormal behavior, such as bot accounts. The public dataset (Cresci et al. 2017) was used. Bagging-TPMiner (BTPM), Bagging-RandomMine (BRM), One-Class K-means with Randomly projected features Algorithm (OCKRA), one-class SVM, and NB were the classifiers that were taken into consideration. For categorization, only 13 numerical features were extracted. With an average AUC value of 0.921, Bagging-TPMiner outperformed all other classifiers over a number of experiments.

Moreover, (Attia et al. 2022) proposed a new multi-input DNN technique-based content-based bot detection model. They used the 6760 records from the public PAN 2019 Bots and Gender Profiling Task (Rangel and Rosso 2019) dataset. The proposed multi-input model includes three phases. Their proposed Multi-input model includes 3 phases. The first phase represents the first input as an N-gram model of a 3D matrix of 100*8*300 as model input to two-dimensional CNN. On the other hand, the second phase input is one-dimensional CNN model that has a vector with M length (100 tweets) as model input. The final phase has the previous models with fully connected neural networks to combine them. Each model was trained using suitable hyper-parameters values. Their model achieved a detection accuracy of 93.25% and outperforms other newly proposed models in bot detection.

In the work of (Sayyadiharikandeh et al. 2020) for each class of bots, they recommended training specialized classifiers and combining their conclusions using the maximum rule. In the most recent version of Botometer, they also produced Ensemble Specialized Classifier (ESC). Additionally, the authors used 18 different public labeled datasets from Bot Repository, and over 1200 features were extracted. Features were divided into 6 categories: metadata, retweet/mention networks, temporal features, content information, and sentiment features. Accordingly, a cross-domain performance comparison and analysis was performed using all the 18 different datasets. The authors recommend considering the three types of bot class as in (Cresci et al. 2017) dataset. Moreover, the authors provided a list of the most informative features per bot classes in the used public dataset.

A comprehensive comparative analysis was conducted by (Shukla et al. 2021) to determine the optimal feature encoding, feature selection, and ensembling method. From the Kaggle repository, a total of 37,438 records comprising the training and testing dataset were acquired. Scaling of numerical attributes and encoding of categorical attributes were two steps in the pre-processing of the dataset. A total of 19 attributes were extracted. The model used the classifiers: RF, Adaboost, NN, SVM, and KNN. It was determined that employing RF for blending produced the best results and the highest AUC score of 93%. Since the proposed approach uses Twitter profile metadata, it can detect bots more quickly than a system that analyzes an account's behavior. However, the system's reliance on static analysis reduces its efficiency.

Ramalingaiah et al. (2021) represented an effective text-based bag of words (BoW) model. BoW produces a numerical vector that can be utilized as inputs in different ML algorithms. Using resulted features from feature selection process, different ML algorithms were implemented like DT, KNN, LR, and NB to calculate their accuracies and compare it with their classifier which uses the BoW model to detect Twitter bots from a given training data. The utilized dataset from Kaggle with 2792 training entries and 576 testing entries for evaluation of their models. As a result, the performance of the decision tree gives the highest accuracy which further uses a bag of bots' algorithm to increase accuracy in detecting bots. Their classifier performs the best as it uses a bag of words model with test data yields an accuracy of over 99%.

A ML method based on benchmarking was proposed by (Pramitha et al. 2021) to choose the best model for bot account detection. Dataset obtained from Kaggle with 24,631 records then scraping was performed using the Twitter API to obtain profile features. Furthermore, over-sampling using SMOTE is applied to overcome imbalanced data and improve the models' accuracy. Both RF and XGBoost algorithms were evaluated. XGBoost algorithm outperforms RF, with an accuracy of 0.8908. Additionally, after ranking fifteen different features, they discovered that three significant features—verified, network, and geo-enable—can identify between human and bot accounts.

Many studies implemented effective DL algorithms instead of ML, such as a Behavior-enhanced Deep Model (BeDM) proposed by (Cai, Li, and Zengi 2017b) for bot detection using a real-world public labeled dataset of size 5658 accounts and 5,122,000 tweets from Twitter, which have been collected with honeypots. The model fused tweets content as temporal text data and the user posting behavior information using DL by applying a DNN to detect bots. The DL frameworks used in the BeDM are CNN and LSTM. Compared to Boosting (Gilani et al. 2016; Lee et al. 2006; Morstatter et al. 2016) baselines, the BeDM attained the highest F1 score of 87.32%, which proved the efficacy of the model.

Later in the same year, (Cai, Li, and Zeng 2017a) proposed analogous work. Yet, the novel Deep Bot Detection Model (DBDM) avoids the laborious feature engineering and automatically learns both behavioral and content representations based on the user representation. Additionally, DBDM took into consideration endogenous and exogenous factors that have an impact on user behavior. DBDM achieved a better results with an F1-score of 88.30%.

Additionally, (Hayawi et al. 2022) also proposed a DL framework, DeeProBot used eleven user profile metadata-based features. Five training and five testing datasets were

used from Bot Repository. Additionally, the text feature was embedded using GLoVe which aided in enhanced learning from the features. To detect bots, DeeProBot employed a hybrid Deep NN model. On the hold-out test set, DeeProBot gave an AUC of 0.97 for bot detection.

However, in a novel framework called GANBOT (Najari et al. 2022) modified the (Generative Adversarial Network) GAN concept. The generator and classifier were connected via an LSTM layer as a shared channel between them, reducing the convergence limitation on Twitter. By raising the likelihood of bot identification, the suggested framework outperformed the existing contextual LSTM technique. A total of 8386 from the Cresci2017 dataset were used. Results were assessed for four distinct vector dimensions: 25D, 50D, 100D, and 200D; the highest result was 949/0.951 for 200D.

A total of seventeen state-of-the-art methods for bot detection were described by (Kenyeres and Kovács 2022) together based on DL models. They classified Twitter feeds as bots or humans, based solely on the account's textual form of the tweets. PAN 2019 Bots and Gender Profiling task (Rangel and Rosso 2019) dataset was used which consisted of 11,560 labeled users. The core of seven models was based on LSTM networks, four based on Encoder Representations from Transformers (BERT) models, and one a combination of the two. For tweet classification, the best accuracy was obtained using fine-tuned BERT model of 0.828. While for account classification, the Adaboost model archived the best accuracy of 0.9. Their findings demonstrate that, even with a small dataset, DL models may compete with Classical Machine Learning (CML) methods.

Moreover, (Martin-Gutierrez et al. 2021) provide a multilingual method for detecting suspect Twitter accounts through DL. Dataset used in their work was collected using Twitter API of 37,438 Twitter accounts. Several experiments were conducted using different combinations of Word Embeddings to obtain a single vector regarding the text-based features of the user account. These features are later on concatenated with the rest of the metadata to build a potential input vector on top of a Dense Network denoted as Bot-DenseNet. The comparison of these experiments showed that the Bot-DenseNet when using the so-called RoBERTa Transformer as part of the input feature vector with an F1-score of 0.77, produces the best acceptable trade-off between performance and feasibility.

In this research, (Ping and Qin 2019) proposed a social bot detection model DeBD based on the DL algorithm CNN-LSTM for Twitter. CNN was used by DeBD to extract the joint features of the tweet content and their relationship. To carry out the experiments, a dataset of 5132 accounts was created. Secondly, the potential temporal features of the tweet metadata were extracted using LSTM. Finally, in order to achieve the purpose of detecting social bots, the temporal features were finally fused with the joint content features. The dataset used in this experiment was from (Cresci et al. 2017). All the experiments achieved a detection accuracy of more than 99%.

Daouadi et al. (2019) proved that a Deep Forest algorithm combined with thirteen metadata-based features is sufficient to accurately identify bot accounts on Twitter. Two datasets were used which were published by (Lee et al. 2006; Subrahmanian et al. 2016). The Twitter API was used to gather the dataset. The implementation was performed for more than 30 conventional algorithms, including Bagging, MLP, AdaBoost, RF, SL, etc. With an accuracy of 97.55%, the Deep Forest method surpassed the other conventional supervised learning techniques.

In this paper, (Cable and Hugh 2019) implemented the algorithms: NB, LR, Kernel SVM, RF, and LSTM-NN to identify political trolls across Twitter and compared their accuracies. A dataset of tweet ids related to the 2016 elections was used by scraping the Twitter API and obtaining a total of 142,560 unique tweets. The features were extracted using several methods: Word count, TF-IDF, and Word embeddings. The LSTM-NN obtained a test accuracy of 0.957.

Since it is important to determine the best features for enhancing the detection of social bots. To locate these ideal features, (Alothali, Hayawi, et al. 2021b) offer a hybrid feature selection (FS) technique. This method evaluates profile metadata features using random forest, naive Bayes, support vector machines, and neural networks. Using a public dataset made accessible by Kaggle that had a total of 18 profile metadata features, they investigated four feature selection approaches. In order to find the best feature subset, they employed filter and wrapper approaches. They discovered that, when compared to other FS methods, the cross-validation attribute evaluation performed the best. According to their findings, the random forest classifier has the best score using six optimal features: favorites count, verified, statuses count, average tweets per day, lang, and ID.

Lastly, (Sengar et al. 2020) proposed both ML and DL to distinguish bots from genuine users on Twitter. This was done by gathering user activity and profile-based features, then applying supervised ML and NLP to accomplish the goal. A labeled Twitter dataset which contains more than 5000 users and 200,000 tweets was used to train the classifiers. After analysis and feature engineering, eight features were extracted. Different learning models were compared and analyzed to determine the best-performing bot detection system namely KNN, DT, RF, AdaBoost, GB, Gaussian Naive Bayes (GNB), MNB, and MLP. Results showed that NN-based MLP algorithm gave the most accurate prediction with an accuracy of 95.08%. A CNN architecture was proposed for tweet level analysis by combining user and tweet metadata. The MIB Dataset (Cresci et al. 2017) was used.

The novel approach gave a staggering improvement. RF and GB gave the highest accuracy of 99.54%.

3.1.9 Twitter—detecting spambots

Some studies demonstrate the detection of spammers, starting with a hybrid method for identifying automated spammers based on their interactions with their followers was presented (Fazil and Abulaish 2018). Nineteen distinct features were retrieved, integrating community-based features with those from other categories like metadata-, content-, and interaction-based features. A real public dataset of 11,000 labeled users was used. The performance was analyzed using three supervised ML techniques namely RF, DT, and BN which were implemented in Weka. All three metrics—DR-0.976, FPR-0.017, and F-score 0.979, were found to be the best for RF. Lastly, it was determined that interaction- and community-based features are the most successful for spam identification in comparison after executing a feature ablation test and examining the discrimination capability of various features.

Oentaryo et al. (2016) categorized bots based on their behavior as broadcast, consumption, and spambots. A systematic profiling framework was developed which included a set of features and a classifier bank. Numeric, categorical, and series features were taken into consideration. The private manually labeled dataset used consisted of bots and non-bot 159 K accounts. Four supervised ML algorithms were employed which include: NB, RF, SVM, and LR. It was seen that LR outperforms the other classifiers by depicting an F1 score of 0.8228.

The research conducted by (Heidari et al. 2020) firstly, they created a new public data set containing profile-based features for more than 6900 Twitter accounts from the (Cresci et al. 2017) dataset where the input feature set consisted of age, gender, personality, and education from users' online posts. To build their system, they compare the following classifiers: RF, LR, AdaBoost, Feed-forward NN (FFNN), SGD. The results showed that the FFNN model with 97% accuracy provides the best results as compared with the other classifiers. Lastly, a new bot detection model was introduced which uses a contextualized representation of each tweet by using Embeddings from Language Model (ELMO) and Global Vectors for Word Representation (GloVe) in the word embedding phase to have a complete representation of each tweet's text. The model created multiple FFNN's models on top of multilayer bidirectional LSTM models to extract different aspects of a tweet's text. The model detected bots from human accounts, regardless of having the same user profile and achieved 94% prediction accuracy in two different testing datasets.

A spam detection AI approach for Twitter social networks was proposed by (Prabhu Kavin et al. 2022). The dataset (7973 accounts) was collected using Twitter Rest API and combined with the public dataset "The Fake Project" (Cresci et al. 2015). For pre-processing, dataset tokenization, stop word removal, and stemming were applied. User-based and content-based features were extracted from the dataset. To develop the model, a variety of ML methods, including SVM, ANN, and RF, were applied. With user-based features, the findings showed that SVM had the highest precision (97.45%), recall (98.19%), and F measure (97.32%).

In this research, (Eshraqi et al. 2016) determined a clustering algorithm that identified spam tweets (anomaly problem) on the basis of the data stream. The dataset consisted of 50,000 Twitter user accounts and 14 million tweets. The pre-processing was done by RapidMiner and then, transferred into Massive Online Analysis (MOA) for implementation. The features extracted were based on Graphs, Content, Time, and Keywords. When using the DenStream algorithm (Cao et al. 2006), regulating needed to be done properly. The model successfully identified 89% of available spam tweets. Furthermore, the results achieved by the model showed an accuracy of 99%.

Mateen et al. (2017) used 13 user-, content—as well as graph-based features to classify between human and spam profiles. The real public dataset used for this study was provided by (Gu 2022) which consisted of 11 K user accounts and 400 K tweets approximately. Three classifiers namely J48, DE, and NB were used for evaluation. J48 and DE outperformed the other classifiers using the hybrid technique of combined features by showing a 97.6% precision. Results showed that for the dataset employed, the hybrid technique significantly improved precision and recall. Additionally, compared to content- and graph-based features, which demonstrated 92% accuracy, user- and graph-based features correctly classified only 90% of cases.

Moreover, (Chen et al. 2017a, b) found that over time, the statistical characteristics of spam tweets in their labeled dataset changed, which impacted the effectiveness of the existing ML classifiers and is known as Twitter spam drift. Using Twitter's Streaming API, a public dataset of 2 million tweets was gathered. The Web Reputation Technology from Trend Micro was used to identify the tweets that were considered spam. The Lfun system, which was learned from unlabeled tweets, was proposed. Day 1 training and Day 2 to Day 9 testing results showed that RF only obtained DR ranging from 45 to 80%, whereas RF-Lfun increased to 90%. The Detection Rate of RF was roughly 85% from Day 2 training to Day 10 testing, but that of RF-Lfun was over 95%.

Kumar and Rishiwal (2020) explored and provided a framework for identifying spammers, content polluters, and bots using a ML approach based on NN usage. A data set consisting of 5572 tweets containing the text messages and their categorization labeling was used. Various algorithms were trained mainly MNB, Bernoulli, NB, SVM, and

Complementary NB. The most effective and best classification of spam account detection was shown by MNB with an accuracy of 99%.

In this study, (Güngör et al. 2020) used a dataset of 714 tweets that had been manually labeled and retrieved through the Twitter API. Eight profile-based features and five tweet-based features were extracted and analyzed. Additionally, a set of guidelines had been discovered via adding followers and friend FF rate, and spam accounts had been detected. For this experiment, the algorithms NB, J48, and LR were used. J48 performed the best, achieving an accuracy of 97.2%. In conclusion, the accuracy rate increased as a result of the usage of both tweet- and profile-based features.

By utilizing a dataset of 82 accounts of tweeters who use both Arabic and English, (Al-Zoubi et al. 2017) improved spam identification. J48, MLP, KNN, and NB were the algorithms used and compared in tenfold cross-validation with stratified sampling as a training/testing methodology. With an accuracy of 94.9, J48 demonstrated the best spam detection ability using the top seven features discovered by ReliefF.

For bot detection, (Heidari et al. 2021) analyzed the sentiment features of tweets' content for each account to measure their impact on the accuracy of ML algorithms. The authors have used (Cresci et al. 2017) dataset of the size of 12,736 accounts and 6,637,615 tweets. The bot detection methodology proposed by the authors is centered on the number of tweets that show a concentration on extreme opinions for an individual account. Whether the opinions are overly negative, positive, or neutral, it indicates the user is a bot. ML models such as RF, NN, SVM, and LR were examined using the proposed sentiment features. The highest result was achieved using Support Vector Regression (SVR) with an F1-score of 0.930.

The research work (Rodrigues et al. 2022) focused on identifying live tweets as spam or ham and performed sentiment analysis on both live and stored tweets to classify them as either positive, negative, or neutral. The proposed methodology used two different datasets from Kaggle. Vectorizers like TF-IDF and BoW models were used to extract sentiment features, which were then fed into a variety of ML and DL classifiers. The classifiers achieved the highest accuracy rate using LSTM in both spam detection with 98.74% and sentiment analysis with 73.81% accuracy.

The work (Andriotis and Takasu 2019) proposed a content-based approach to identify spambots. Technically, four public datasets were used in this study, which was (Cresci et al. 2017; Varol et al. 2017; Yang et al. 2012, 2013). Collectively, the datasets contain tweets of nearly up to 20 K accounts of both bots and genuine users. The methodology proposed employed metadata, content, and sentiment features. Furthermore, the performance of the KNN, DT, NB, SVM, RF, and AdaBoost algorithms was tested. AdaBoost showed the best result with a 0.95 F1-score. Additionally, the study depicted that sentiment features add value when combined with known features to bot detection algorithms.

Also, (Sadineni 2020) detect spam using a dataset from Kaggle that included 950 users and ten content-based attributes, demonstrating that SVM and RF outperform NB in terms of performance.

On the other hand, (Kudugunta and Ferrara 20182018) presented a contextual LSTM architecture based on a DNN that uses account metadata and tweet text to identify bots at the tweet level. The tweet text served as the primary input for the model. It was tokenized and converted into a series of GloVe vectors before being fed into the LSTM, which then fed the data into a 2-layer NN with ReLU activations. High classification accuracy can be attained using the suggested model. Additionally, the compared techniques for account-level bot identification that used synthetic minority oversampling reached over 99% AUC.

In this study, Arabic spam accounts were detected using text-based data with CNN models and metadata with NN models by (Alhassun and Rassam 2022) utilizing Twitter's premium API, and a dataset of 1.25 million tweets was collected. By flagging terminated accounts, data labeling was carried out. 13 features based on tweets, accounts, and graphs were retrieved. The findings demonstrated that the suggested combination framework used premium features to reach an accuracy of 94.27%. The performance of spam detection improved when premium features were compared to standard features when used with Twitter.

An efficient technique for spam identification was introduced by (Inuwa-Dutse et al. 2018). They suggested an SPD Optimized set of features that are apart from historical tweets. They focused on user-related attributes, user accounts, and paired user engagement. MaxEnt, Random Forest, ExtraTrees, SVM, GB, MLP, MLP+, and SVM were among the classification models that were utilized and evaluated based on three datasets, Honeypot (Lee et al. 2006), SPDautomated, and SPDmanual. The performance reached a peak of 99.93% when using GB on the SPD Optimized set. This technique can be used in real-time as the first step in a social media data gathering pipeline to increase the validity of research data.

Instead of employing the LCS method, (Sheeba et al. 2019) discovered spams using the RF classifier technique. The study used a dataset of 100,000 tweets. Latent Semantic Analysis was used to further identify the account after the RF classifier had identified it as a spambot using Latent Semantic Analysis (LSA). The proposed approach delivered benefits in terms of time consumption, high accuracy, and cost effectiveness.

An approach to spam identification based on DL methods was developed by (Alom et al. 2020). CNN architecture was utilized for the text-based classifier, while CNN and NN

were merged for the combined classifier to classify tweet text and metadata, respectively. On two distinct real-world public datasets, Honeypot (Lee et al. 2006) and 1KS-10 K (Yang et al. 2013), the suggested approach's performance was compared to those of five ML-based and two DL-based state-of-the-art approaches. For the datasets Honeypot and 1KS-10KN, the accuracy of 99.68% and 93.12%, respectively, was attained.

In this research, (Reddy et al. 2021) implemented some supervised classification algorithms to detect spammers on Twitter. Information was obtained from tweepyAPI which comprised 2798 accounts in the training set and 578 accounts in the test set. Eighteen profile-base features were extracted. In terms of accuracy, Extreme Machine Learning (EML) obtained a better accuracy of 87.5.

3.1.10 Twitter—detecting sybil bots

Firstly, (Narayan 2021) used ML algorithms for the detection and successful identification of bogus Twitter accounts/bots. The algorithms used were DT, RF, and MNB. The dataset used included 447 Twitter accounts. Twitter API was used for the excavation of the data. DT has been found to be more accurate as compared to RF and MNB.

In their work, (Bindu et al. 2022) proposed three efficient methods to successfully detect fake accounts. The classification algorithms used were as follows: Linear and radial SVM, RF, and KNN. The data set used contained a total of 3964 records. RF gave more accurate prediction results accordingly overcoming the overfitting problem. The K-Fold Cross-Validation Scores for RF include a mean of 0.979812 and a standard deviation of 0.019682. On the other hand, in comparison Radial SVM did not perform well, and gave more False Negatives. However, using the Ensemble approach, higher accuracy was achieved.

Likewise, (Alarifi et al. 2016) studied the features used for detecting sybil accounts. Twitter4j was used to gather a manually labeled sample dataset of 2000 Twitter accounts (humans, bots, and hybrid-both human and bot tweets). Eight content-based features were selected. Four supervised ML algorithms which include J48 (C4.5), Logistic Model Tree, RF, Logitboost, BN, SMO-P, SMO-R, and multilayer NN were used. RF performed the best with a DR of 91.39 for two-class and 88.00 for three-class classification. Lastly, in order to maximize the use of the classifier, the authors developed an efficient browser plug-in.

David et al. (2017) leveraged a public labeled dataset from the project BoteDeTwitter to build half of their data set related to Spain politics. Using the Twitter API, a sample of 853 bot profiles and the most recent 1000 tweets from each user's timeline was collected. To create an initial feature set, 71 features based on profiles, metadata, and content were extracted. The following supervised ML methods were compared: RF, SVM, NB, DT, and NNET. Even though the increases were not significant after the first six features, RF managed to get the highest average accuracy of 94% by using 19 features.

In (van der Walt and Eloff 2018) paper, Twitter data were mined using the twitter4J API and a non-relational database yielding a total of 169,517 accounts. Engineered traits that had previously been used to successfully identify fraudulent accounts made by bots were added to a sample of human accounts. Without relying on behavioral data, these features were applied to several supervised ML models, enabling training on very little data. The results show that engineered traits, which were previously employed to identify fake accounts created by bots, could only reasonably predict fake accounts created by humans with an F1 score of 49.75%.

Kondeti et al. (2021) implemented ML to detect fake accounts on the Twitter platform. Different ML algorithms were used such as SVM, LR, RF, and KNN along with six account metadata features likes, Lang-code, sex-code, status-count, friends-count, followers-count, and favorites-count. Further to improve these algorithms' accuracy, they used two different normalization techniques such as Z-Score and Min–Max. Their approach achieved high accuracy of 98% for both RF and KNN models.

Khaled et al. (2019) suggested a new algorithm—SVM-NN to efficiently detect sybil bots. Four public labeled datasets were used by the authors. A total of 4456 accounts of both fake and human classes, result from combining them. Sixteen user-based numerical features were extracted from the datasets after applying features reduction, and they were then fed into the SVM, NN, and SVM-NN algorithms. The authors of the researchers assert that their novel SVM-NN uses fewer features than existing models. SVM-NN was the best-performing algorithm as it showed an accuracy of around 98%.

In the study, (Ersahin et al. 2017) collected their own dataset of fake and real accounts using Twitter API. The dataset consisted of 1000 accounts' data later pre-processed using Entropy Minimization Discretization (EMD) on sixteen user-based numerical features. NB with EMD showed the best result with 90.41% accuracy.

However, in order to predict sybil bots on Twitter using deep-regression learning, (Al-Qurishi et al. 2018) introduced a new model. The authors used two publicly available labeled datasets that had been generated during the 2016 US election and collected using Twitter API. The first dataset consisted of 39,467 profiles and 42,856,800 tweets. Whereas the second dataset consisted of 3140 profiles and 4,152,799

tweets. The authors extracted 80 online and offline features based on Profile-, Content- (Temporal, Topic, Quality, and Emotion-based), and Graph. Accordingly, the features were fed into the Deep Learning Component (DLC) FFNN. When fed with noisy and unclear data, the results depicted an accuracy of 86%. Categorical features showed clear segregation that all sybil bots disable their geographical location and have an unverified account. While numerical features showed that sybil bots have a noticeably young account age (recently created). Additionally, the number of re-post and mentions are significantly higher in the sybil's accounts.

Gao et al. (2020) proposed a content-based method to detect sybils. The proposed method included three main phases: CNN, bi-SN-LSTM, and the dense layer and softmax classifier stacked to output the classification results. The proposed bi-SN-LSTM network, in contrast to the bi-LSTM, employs SELU as the activation function of its recurrent step, enabling limitless modifications to the state value. The proposed model achieved a high F1-score of 99.31% on the "My Information Bubble" (Cresci et al. 2015) dataset.

3.1.11 Weibo—detecting social bots

Data collection, feature extraction, and detection modules were all included in the DL technique known as TPBot proposed by (Yang et al. 2022). To begin with, the data collection module used a web crawler to obtain user data from Sina Weibo using dataset collected by (Wu et al. 2021). Then, depending on each user's profile, the feature extraction module extracted temporal-semantic and temporal-metadata features. Finally, in the detection module, a detection model based on BiGRU was developed. TPBot outperformed baselines, by achieving an F1-score of 98.37%. Additionally, experiments were carried out on two Twitter datasets (Cresci et al. 2015, 2017) to assess the generalization capabilities of TPBot, and on both datasets, it outperformed the baselines.

Behavioral analysis and feature study were performed by (Dan and Jieqi 2017) to extract the effective features of Weibo accounts and build a supervised model to detect bots. A dataset of 5840 accounts from the Sina-Weibo data warehouse was used to discriminate between real and bot users. Eleven users' behavioral-based features were extracted and fed into DT, C4.5, and RF algorithms. The RF algorithm performed measurably better with a 0.944 F-measure.

Moreover, (Huang et al. 2016) built a classifier that combined NB and Genetic Algorithm on Weibo. The genetic algorithm was used to create an optimal threshold matrix which efficiently increased the precision of the model by improving the conditional probability matrix. Two models were built using two different datasets. One dataset (1000) was crawled by R and the other consisted of spammers purchased from the sales platform (600) and legitimate users crawled from friends and relatives (400). 9 profile-based features were set as attributes. In the comparison of the performance with LR, DT, and NB showed a higher precision of 0.92.

3.1.12 Weibo—detecting spambots

In this paper, for effective spammer detection, an EML-based supervised ML approach was proposed by (Zheng, Zhang, et al. 2016b). The study started by crawling Weibo data to create a labeled dataset. 1000 messages, both spam and normal, were chosen from the collected dataset. Message content and user behavior-based features were then extracted for a total of 18 features, which were then fed into the classification algorithm. With a TPR of spammers and non-spammers reaching 99 and 99.95%, respectively, the experiment and evaluation demonstrated that the suggested approach offers good performance.

Zheng, Wang, et al. (2016a) proposed a two-phase-based spambot detection approach. In the first phase, authors took existing work about user features. In the second phase, the authors introduced content mining for spambot detection. Using web crawlers, a dataset of 517 accounts and 381,139 tweets was collected. Eighteen behavioral and content-based features were extracted. The experiment results were compared with SVM, DT, NB, and BN algorithms. The proposed two-phased method performed better than the mentioned algorithms with an accuracy of 90.67%.

However, (Wu et al. 2021) used DNN and active learning (DABot) as a technique to detect bots. They classified bots into three types: spammers, bots that engage with accounts to increase impressions, and bots involve with politics. Thirty features were extracted and classified as metadata, interactions, contents, and time. A data collection of 20 K users and 214,506 posts from all users was produced as a consequence of the authors manually labeling the user accounts. Different stages made up the modeled architecture: data input for each user, ResNet block, BiGRU block, Attention layer, and Interference layer.

Another spam detection technique was put forth by (Xu et al. 2021) and relied on the self-attention Bi-LSTM NN model in conjunction with ALBERT. Two datasets were employed in the experiment: one self-collected (582

accounts) and the other microblogPCU (2000 accounts). They converted the text from social network sites into word vectors using ALBERT and then, input those word vectors into the Bi-LSTM layer. The final feature vector was created after feature extraction and combination with the information focus of the self-attention layer. To get the result, the SoftMax classifier performed classification.

3.1.13 Weibo—detecting sybil bots

In this research, (Bhattacharya et al. 2021) suggested a detection model that performed improved prediction of fake Weibo accounts using a variety of Ensemble ML algorithms. The 918 HTML pages that made up the public Weibo dataset were obtained from Kaggle. Data scraping was used to construct the fake accounts dataset. Content-based attributes were extracted. Five supervised models—RF, SVC, NB, LR, and GB—were taken into consideration. For determining the final result, the RF classifier's highest F1 score of 0.93, precision, and recall were taken into account. Finally, a plot confusion matrix revealed an inaccurate prediction for 44 accounts, providing the opportunity for additional research.

3.2 Using semi-supervised ML

Few studies on only two platforms have implemented semi-supervised ML to detect spambots and sybil bots which are discussed below.

3.2.1 Twitter—detecting spambots

Sedhai and Sun (2018) were the earliest that utilized a semi-supervised approach for spam detection. Their proposed S3D approach contains two main components which are spam detection components in real-time mode, and model update components in batch mode to periodically update the detection models. For spam detection, they apply four detectors which are a blacklisted domain detector using blacklisted URLs, a near-duplicate detector to label near-duplicate tweets using clustering, a reliable ham detector to label tweets that are posted by trusted users and that do not contain spammy words, and a multi-classifier using NB, LR, and RF models to labels the remaining tweets. Their approach achieved good accuracy results for spam detection on the public HSpam14 dataset along with four types of features to represent tweet and cluster Hashtag, in addition to being effective in detecting new spamming forms.

In this research work, (Alharthi et al. 2019) proposed a semi-supervised ML technique that classified Twitter accounts as spam or genuine accounts based on their behavior and profile information. A dataset consisting of (500) active Arab users was collected through a Twitter API and manually labeled. Label spreading and label propagation algorithms were implemented using 16 extracted features. The features (TweetsAverage), (Number of the accounts' followers to the number of his/her friends), (Tweet Source), and (is all the tweets have the same source?) were proven to be the most efficient features. The proposed model achieved the following results an F-measure of 0.89, an accuracy of 0.91, and an AUC of 0.90.

3.2.2 Twitter—detecting sybil bots

In this study, (Zeng et al. 2021) used semi-supervised self-training learning by utilizing a Kaggle data set of real and fake Twitter accounts. In this suggested technique, a self-training method was applied to automatically classify Twitter accounts. Further, to effectively reduce the impact of class imbalance on the identification effect, the resampling technique was incorporated into the self-training process. The proposed framework displayed good identification results on six different base classifiers, particularly for the initial batch of small-scaled labeled Twitter accounts.

3.2.3 Weibo—detecting spambots

Only a single study based on a semi-supervised approach by (Ren et al. 2018) detected spambots on Weibo. The authors have collected the dataset (31,147 users and 754,112 tweets) using a crawler. Behavioral and Content-based features were utilized to feed the model. Compared to NB, LR, SVM, and J48 algorithms, the proposed approach showed better results in all the evaluation metrics applied.

3.3 Using unsupervised ML

Few studies on only three platforms have implemented unsupervised ML to detect social bots, spambots and sybil bots which are discussed below.

3.3.1 Facebook—detecting spambots

Sohrabi and Karimi (2018) carried out the Facebook platform's spam filtering mechanism for posts and comments.

Different exploration techniques and optimization techniques, including PSO, simulated annealing, ant colony optimization, and Differential Evolution (DE) could be used with the suggested filtering strategy. Seven metadata features were recovered from the dataset, which was made up of 200,000 wall posts and comments on them. They examined the DB index and DE clustering method, SVM, and DT, three algorithms with PSO-based feature selection. The hybrid algorithm created by integrating SVM and clustering techniques produced the best outcomes.

3.3.2 Facebook—detecting sybil bots

Fake Facebook profiles Detection using a group of supervised and unsupervised mining algorithms was performed by (Albayati and Altamimi 2020). The main components were the Crawler and the analyzer modules. A dataset of 982 profiles and a set of 12 behavioral and profile-based features. In the analyzer module, using the mining tool RapidMiner Studio, they implemented two unsupervised algorithms, K-Means and K-Medoids, along with three supervised algorithms: ID3, KNN, and SVM. The findings of the performance evaluation method revealed that supervised algorithms outperformed unsupervised algorithms in terms of accuracy rates. With a 97.7% accuracy rate, ID3 surpasses other classifiers.

3.3.3 Instagram—detecting sybil bots

In this paper, (Munoz and Paul Guillen Pinto 2020) detected fake profiles on Instagram. Web scrapping techniques were used for data extraction on the third-party site to Instagram. A dataset of 1086 true and false profiles was designed. 17 features were extracted based on metadata and multimedia information. Various ML algorithms such as DT, LR, RF, MLP, AdaBoost, GNB, Quadratic Discriminant Analysis, Gaussian process classification, SVM, and NN were deployed. RF obtained the best accuracy of 0.96 as well as the best true and false prediction precision.

3.3.4 Twitter—detecting social bots

A bot detection technique was put forth by (Chen et al. 2017a, b) that used shortened URLs and tweeting almost duplicate content over an extended period of time to look for a particular class of malicious bots. This method automatically gathered bot groups from real-time Twitter streams as opposed to earlier work. The following nine URL shortening services were investigated: bit.ly, ift.tt, ow.ly, goo.gl, tinyurl.com, dlvr.it, dld.bz, viid.me, and ln.is. The model is made up of four sequentially operating parts: a crawler, a duplicate filter, a collector, and a bot detector. In order to conduct the experiment, 500,000 tweets were collected. According to the experiments, bot networks and accounts made up a mean of 10.5% of all accounts that employed shortened URLs.

Interestingly, (Mazza et al. 2019) presented a visualization technique named Retweet Tweet (RTT) for gaining insights into the retweeting behavior of Twitter accounts. For the purpose of identifying retweeting social bots, Retweet-Buster (RTBUST), an unsupervised group-analysis method, was employed. Using the Twitter Premium Search API, a dataset of 10 M Italian retweets shared by 1446, 250 unique users was compiled. RTBust was built around an LSTM variational autoencoder. Based on the results of the Hierarchical Density-Based Spatial Clustering (HDBSCAN) algorithm, it was decided whether the account was a bot or legitimate. In comparison with using it with PCA and TICA, the proposed RTBUST technique using the VAE produced the best detection performance, i.e., $F1 = 0.87$.

Anwar and Yaqub (2020) proposed a quick way to isolate bots from the Twitter discussion space. The dataset used was unlabeled data collected through Twitter Search API during the 2019 Canadian elections. It consisted of 103,791 accounts and 546,728 tweets. 13 metadata features were extracted using PCA implemented in K-means clustering. Results showed that bots have a higher rate of retweet percentage, daily tweets, and daily favorite count, which are incorporated with the known characteristics of bots.

In this paper, to enhance the detection accuracy of social bots, (Wu et al. 2020) proposed an improved conditional GAN to extend imbalanced data sets prior to applying training classifiers. The Gaussian kernel density peak clustering algorithm (GKDPCA), an unsupervised modified clustering algorithm, was put into practice. 2433 users' data was compiled into a dataset. On the basis of six different feature types—user meta-data, sentiment, friends, content, network, and timing, eleven different features were retrieved. With an F1 score of 97.56%, the enhanced CGAN performed better than the three popular oversampling methods.

Khalil et al. (2020) used two unsupervised clustering algorithms DBSCAN and K-Mean. Six publicly available datasets (2232, 3465, and 1969) were used mentioned in (Kantartopoulos et al. 2020). Eight profile-based features were extracted. It was concluded that DBSCAN performed better by achieving an accuracy of 97.7%.

The second contribution of (Barhate et al. 2020) is aimed at using an unsupervised ML approach. Hashtag data from the Twitter API was mined and a dataset of 140 K users was

created. Using the PCA and K-means clustering algorithms, users were divided into four groups based on activity-related features. This enabled the analysis of each cluster's bot percentage. The age distribution of users in a trending hashtag was also plotted by the authors.

3.3.5 Twitter—detecting spambots

Some analyses were able to detect spammers successfully using unsupervised learning methods for instance, (Cresci et al. 2016) put forth a novel behavioral-based unsupervised approach for spambots accounts detection, inspired by biological DNA. The proposed methodology extracts and analyzes digital DNA sequences from users' actions. The authors manually created a dataset (4929 accounts) of verified spambot and genuine accounts. Each account got associated with a string that encodes its behavioral information. Compared to other benchmark work done, DNA fingerprinting model achieved the highest result with an MCC of 0.952.

Furthermore, (Koggalahewa et al. 2022) proposed an unsupervised spammer detection approach. In Stage 1, the clustering based on user interest distribution was performed. In Stage 2, spam detection was performed based on peer acceptance. Lastly, by assessing the user's peer acceptability against a threshold, a user was categorized as spam or genuine. Three datasets were used namely Social Honey Pot (Lee et al. 2006), HSpam14 million Tweets, and The Fake Project (Cresci et al. 2017). Detection accuracies pointed out that three features Local Outlier StandardScore (LOSS), Global Outlier Standard Score (GOSS), and Entropy when combined gave the best results. SMD performed the best with an accuracy of approximately 0.98 on the three datasets.

4 Discussion

To begin with, from all the reviewed studies we noticed that Twitter is the most researched platform with a total of 71 studies carried out, followed by 12 studies on Facebook, 11 studies on Instagram, 9 studies on Weibo, and lastly only 2 studies were conducted on LinkedIn. Appendix Table 2 summarizes all the reviewed ML-based studies focusing on the dataset used, feature's type, best-performing algorithm, and the highest result obtained, respectively. With respect to the most detected type of bot on each platform, Twitter had 36 studies on social bots, Facebook had 7 studies on sybil bots, Instagram had 8 studies on sybils, Weibo had 5 studies on spambots, and lastly, LinkedIn had only 2 studies which were on sybils.

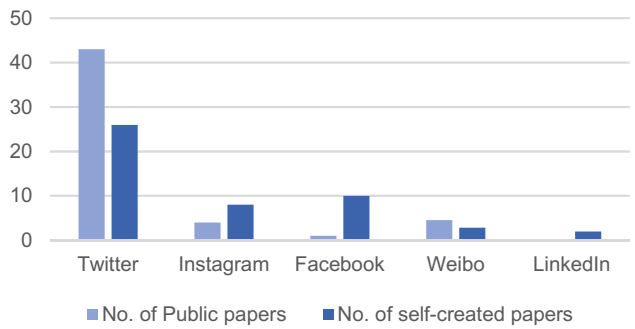

Fig. 4 Datasets distribution on each platform for the reviewed papers

Researchers in the reviewed papers used different datasets both publicly available and self-created to evaluate their models to classify bots from humans on the five addressed social media platforms. A summary of the 38 publicly available datasets has been provided in Appendix Table 3. From the Appendix Table 3, the most widely used datasets are MIB datasets which are the Cresci2017 and Cresci2015. However, the Cresci2017 dataset was the most used dataset by researchers because it includes five distinguished types of social media bots, namely genuine accounts, social spambots, traditional spambots, fake followers or Sybil, and a test set consisting of a mix between genuine and social spambots. Besides the variety of dataset's bot types, it is relatively a recent and large labeled dataset consisting of 12,736 accounts and 6,637,615 tweets in total, which may have attracted researchers to conduct their studies using the Cresci2017 dataset to detect spam and social bots in the Twitter platform. While Cresci2015 includes three fake follower's datasets, and two human accounts datasets making it more efficient in the detection of sybil bots on Twitter. The Fake Project dataset is one of the Cresci2015 which is much more used together with Honeypot dataset to detect spambots on Twitter. Different Kaggle's public datasets were used to detect different types of bots on Twitter. Due to the majority of papers related to Twitter compared to other platforms, more provided datasets are publicly available than the self-collected (private datasets) one. While other platforms such as Facebook and Instagram have more datasets that were self-created (private datasets). Weibo has almost equal types of datasets while LinkedIn has only self-created. Figure 4 illustrates public and collected datasets on each platform.

Despite the fact that there are numerous datasets available, some of them only contain human or bot IDs and labels. As a result, scraping is done using the appropriate collection API or method to obtain profile features or other information from an ID or account. For instance, the Twitter API is used to gather real-time datasets from publicly accessible Twitter data (Rodrigues et al. 2022). Many researchers have created their own datasets using these collection methods on different platforms as shown in Appendix Table 4. In

Twitter, Twitter API was the most used collection method while methods like Twitter4j, Tweepy, ML, Twitter Premium Search API, and REST API were less used. Instagram datasets were collected using Instagram API, Selenium Web Driver tool, 3rd-party Instagram websites, and some manually. For Facebook, mostly used Facebook Graph API to collect data while web crawler was mostly used on the Weibo platform. Lastly, for LinkedIn in only two studies, the dataset was collected manually.

To distinguish between human and automated users on social media platforms, it's critical to identify an ideal collection of attributes (Alothali, Hayawi, et al. 2021b). A general observation was made that bots have a high friend-to-follower ratio and a low follower growth rate. This can be done by using a variety of features that have been reported in various studies. On the basis of the extracted features in all the reviewed papers, the features were classified into the following categories: Content/Language, User (Profile), Metadata, Behavioral, Network (Community/Interaction), Sentiment, Timing/Temporal, Graph, Numeric/Categorical/Textual/Series, Statistical¸ User Friends, Media and Engagement, Entity and Link¸ Keywords, Internet Overlap, hashtag features, and Periodic features. Content features are based on linguistic cues computed through NLP, mainly part-of-speech tagging. User features are based on properties of the users' accounts and users' relationships. User Meta Data features are information regarding the profile's characteristics. Locating an information source via metadata is known to be effective. Behavior features are calculated by statistical properties from the data. Different aspects of information diffusion patterns are captured by network features. General-purpose and Twitter-specific sentiment analysis algorithms are used to build sentiment features. Time features include statistics of time. Graph features are extracted by modelling the social media platform as a social graph model. Descriptive statistics relative to an account's social contacts are included in user friend features. Interest Overlap features include overlap between two users such as Topical affinity. Appendix Table 5 of the literature review provides examples of features from the reviewed studies as well as a summary of the features used in various social media platforms. According to the table, the most popular feature types from all the reviewed papers are content-based, profile-based, metadata-based, and behavioral-based features on essentially all types of platforms. Content-based features were utilized in 44 studies, followed by user/profile-based features in 42 studies, metadata-based in 27 studies, and behavioral-based in 16 studies. User friend, media, engagement, and keywords on various types of platforms are among the less popular feature types.

In regard to the Twitter platform, 34 studies used profile-based features followed by 32 studies that used content-based features and achieved high results. Meta-data-based features were used in 17 studies. Features based on Timing, Statistical, Keywords, Interaction, Periodic, Latent, Numeric, Categorical, and Series were used only once by single studies and achieved reasonable results. Four studies (Alhassun and Rassam 2022; Al-Qurishi et al. 2018; Mateen et al. 2017; Eshraqi et al. 2016) utilized graph-based features. It came to the notice that when (Eshraqi et al. 2016) combined graph-based features along with Content, Time, and Keywords, a very high accuracy of 0.99 was achieved. Only 5 studies (Wu et al. 2020; Davis et al. 2016; Inuwa-Dutse et al. 2018; Sayyadiharikandeh et al. 2020; Varol et al. 2017) made use of the network-based features. (Inuwa-Dutse et al. 2018) combined such network- and profile-based features and achieved the highest result AUC 99.93%. The number of features utilized in all 71 studies ranged from as less as 5 features to as high as 1000 features. (Varol et al. 2017) and (Davis et al. 2016) used approximately 1000 features and achieved an AUC of 0.95 whereas (Fonseca Abreu et al. 2020) used only 5 profile-based features and still obtained an AUC of 0.999. Regarding crucial features, interaction- and community-based features hold high value in spambot detection (Fazil and Abulaish 2018).

In regard to the Facebook platform, 7 out of 12 studies utilized profile-based features followed by content-based being used by 5 studies. Moreover, by examining the results of this platform's studies, it can be concluded that the highest results were achieved when profile-based and content-based features were combined hence showing a high accuracy of 0.984 in the research conducted by (Rathore et al. 2018). Noteworthy, textual, Categorial, and Numerical-based features were used only in 1 study.

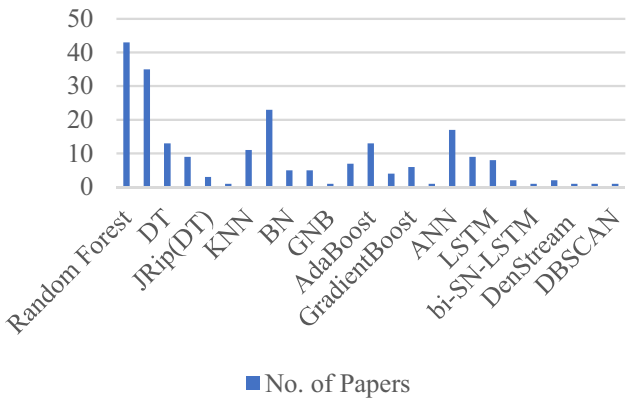

Fig. 5 Bar chart for ML algorithms

Singh and Banerjee (2019) but gave promising results (F1-score 0.99). Features such as "likes", "remarks", "user activities" contributed the maximum for the detection of sybils. Moving on to the third-most researched platform, Instagram, behavior-, content, and profile-based were used in 4 out of 11 studies. The combination of behavior- and content-based features showed the highest performance with an accuracy of 0.9845. In Weibo-based studies, content-based were the most widely used in addition to behavior-based features. Timing as well as semantic features were the least used but, on the contrary, gave the highest results. Since only 2 studies were found on LinkedIn, they made use of profile and statistical features. This platform needs to be extensively explored using other feature types which include content-, metadata-, behavioral-based, etc.

In terms of the different ML-based (supervised, semi-supervised, and unsupervised) techniques utilized in the reviewed papers which were built to compare and detect different types of social bots, Appendix Table 6 presents a list of all the respective papers that utilized the different algorithms. This Table 6 highlights the classifier bot type with the highest performance achieved for each algorithm. Figure 5 shows the number of papers that utilized each ML algorithm. As shown, RF is the best-performing and most applied algorithm among all algorithms in research and SVM is the second most applied algorithm in research followed by NB, DT, and AdaBoost. The least applied algorithms were GNB, ELM, bi-SN-LSTM, and clustering algorithms were the least applied algorithm. For supervised ML, the best performing algorithms of classical ML algorithms, the best performing were RF, JRip, AdaBoost, with their accuracy reaching up to 99.5%, and the least utilized algorithms were ID3 and GNB. As for DL algorithms, the best performing algorithm was CNN with the highest accuracy of 99.68% achieved though most perform well. The least utilized and least popular algorithm was ELM, even though it is considered simple with less training time. Moreover, it was noticed that ML classifiers work well with small-size datasets and DL algorithms with large-size datasets. However, no algorithm can be considered good or bad as it depends on a number of factors such as the dataset size, data preprocessing, and the number and type of features.

Further in terms of semi-supervised learning, despite them being powerful techniques in terms of discovering patterns in big data only four studies were found: three on Twitter and one on Weibo (Sedhai and Sun 2018; Alharthi et al. 2019; Zeng et al. 2021; Zeng et al. 2021). Since large datasets are derived from the Twitter platform which makes labeling an expensive and time-consuming process, semi-supervised techniques such as label propagation and label spreading show the ability to be applied more often. Moreover, integrating resampling along with self-training is a helpful way to reduce the impact of class imbalance when using semi-supervised learning. For the unsupervised learning approach, the DenStream unsupervised clustering algorithm achieved the highest result as compared with other used clustering algorithms like K-Mean and DBSCAN. Though this approach has less popularity and performance compared with the supervised approach (Albayati and Altamimi 2020), this method offers the benefit of not requiring a labeled dataset. Since, there was only one paper that applied this approach, hence additional investigation into this particular algorithm is not possible.

To conclude, not much evidence could be drawn from this analysis that the most researched bot type or the most researched social platforms are necessarily the ones most affected by social bots.

5 Challenges and opportunities

In this section, we shall put forth an elaborate discussion on challenges and future research directions based on our study and analysis. The findings showcased that social bot detection is challenging and this challenge is aggravated as the social network volume increases. To begin with, the most detected and researched bot types are social bots (42 studies), followed by spambots (34 studies), and lastly sybil bots (29 studies). Evidentially, Twitter is the most studied social network with a large number of bots of all types, especially social bots, mainly because of how easy it is to collect data through their API and the vast collection of accessible public datasets. However, social networks such as Instagram, LinkedIn, and Weibo need further in-depth study. Specifically, there is a dearth of studies on Facebook and LinkedIn due to the immense difficulty in obtaining publicly available datasets, which is caused by certain strict privacy policies on those networks. LinkedIn, in particular, does not have much of recent studies conducted on it. Furthermore, only sybil bots were found in the publicly available LinkedIn datasets. Moreover, with slight modifications, the ML techniques used for Instagram will have the potential to be applied to LinkedIn. From our studies, we conclude that Cresci2017 is the most used dataset in social media bot research due to its classification of bots based on their types. Whereas, Instagram has a greater number of sybil bots and two studies based on fake engagements. User and content-based features are the most frequently used for Instagram thereby showing high-accuracy results. Nevertheless, there is a scope for more research on this platform. In terms of features, on twitter (Fonseca Abreu et al. 2020) showed that even with 5 significant features, high results can be achieved. Therefore, new studies can be carried out by using as less features as possible. On Facebook, since profile-, content-, textual-,

categorial-, and numerical-based features contributed high value in various studies, a new research direction can be explored by combining all the above-mentioned five feature types. LinkedIn needs to be extensively explored by using feature types which include content-, metadata-, behavioral-based, etc.

In terms of the reviewed algorithms, it is seen that RF is the best performing in terms of accuracy and the most applied algorithm among all algorithms on all social media platforms in the conducted study. SVM is the second most applied algorithm in research followed by NB, DT, and AdaBoost. DenStream unsupervised clustering algorithm achieved the highest result compared with other used clustering algorithms like K-mean and DBSCAN. Though this approach has less popularity, it has added the advantage of not requiring a labeled dataset. Different algorithms based on Bayes Theorem were used to classify Spam and social bots like NB, MNB, GNB, and NB. However, MNB overperforms the others. Different types of algorithms were used to build Decision Trees like basic DT, J48, JRip and ID3. DT, and J48 were the most applied forms. Yet, the JRip algorithm achieved the best performance among them on spam detection. Different types of boosting algorithms were applied such as AdaBoost, XGBoost, and GradientBoost. AdaBoost was the most applied, whereas GradientBoost performed the best amongst them on social bots detection. The DL approach was mostly applied to detect social bots type on the Twitter platform. Among the different DL algorithms, CNN and LSTM were the highest-performing and most promising algorithms in terms of accuracy.

Comparing algorithms on different platforms, RF achieved the best accuracy result on Weibo and Instagram platform. While AdaBoost achieved highest Detection Rate on Facebook platform. On the other hand, CNN and ANN achieved highest accuracy on twitter platform.

Moving on, future enthusiastic researchers are encouraged to investigate and conduct studies on unstudied social media platforms such as TikTok, Telegram which are known to have bots. As seen above, only four studies employed semi-supervised learning techniques, and a few used unsupervised technique; therefore, these fields need more exploration and contribution. The semi-supervised approach gives unlabeled instances the same weight as labeled ones while also minimizing the cost of labeling the data. More importantly, it is advised that researchers make their datasets available to the scientific community. This will support the training of new models, their testing, and the evaluation of the existing models. Additionally, new public datasets that contain the most recent type of bots are needed. The main gap was observed in the collected research. Only a few papers proposed a ML-technique to detect bots at registration or creation in real-time. As none of the existing research are designed to catch bots and act before they make connections with real users. Whereas in practice, it is desired to detect bots as soon as possible after registration in order to prevent them from interacting with real users. However, this has its own challenges as the bot's detection needs to be done from the basic information provided during the registration time.

Lastly, as the great novelist, Patricia Briggs quotes "Knowledge is a better weapon than a sword". The users on various social media platforms need to gain cybersecurity awareness in order to not get deceived and be able to distinguish between bots and benign accounts and be responsible in situations if a malicious bot was recognized to immediately report it to the platform.

6 Conclusion

This paper made an effort to provide a comprehensive review of the existing studies in the area of utilizing ML for bots detection on social media platforms which are affected by three types of bots—social, spam, and sybils to provide a starting point for researchers to identify the knowledge gaps in this field and conduct future in-depth research.

Furthermore, the usage of supervised, semi-supervised, and unsupervised ML-based approaches was also summarized. Numerous ML and DL methods were analyzed for bots detection, including KNN, RF, DT, NB, SVM, KNN, LSTM, ANN, etc. Visual aids were created to analyze the reviewed papers based on the nature of their datasets, the various categories of features, as well as the performance of employed algorithms. From the analysis for bots detection, we discovered that RF exhibited the highest performance in terms of accuracy and is the most frequently used ML algorithm. Whereas CNN and LSTM were the highest-performing and promising DL algorithms in terms of accuracy. Last but not the least, we addressed and listed some of the challenges, limitations as well as recommended suggestions that can be utilized by enthusiastic future researchers for adding more value and thereby contributing to the field of cybersecurity.

Appendix

See Tables 2, 3, 4, 5 and 6.

Table 2 Summary of Social Media Bot Detection using ML Techniques

References	Technique	Platform	Bot type	Dataset	Feature's type	Best-performing algorithm	Highest result
Adikari and Dutta (2020)	Supervised ML	LinkedIn	Sybil	Shalinda2020	Profile	SVM	Accuracy (87.34%)
Akyon and Esat Kalfaoglu (2019)	Supervised ML	Instagram	Spam, sybil	Fatih2019	Behavioral, content	SVM, ANN	F1-score with oversampling (94%) without (86%)
Alarifi et al. (2016)	Supervised ML	Twitter	Sybil	Twitter4j	Content	RF	DR-91.39, 88.00
Alharthi et al. (2019)	Semi-supervised	Twitter	Spam	Alharthi2019	Behavioral, profile	Label spreading, Label propagation	F-measure (0.89) Accuracy (0.91) AUC (0.90)
Alhassun and Rassam (2022)	Supervised ML	Twitter	Spam	Atheer2022	Graph, profile, content	CNN	Accuracy (94.27%)
Alom et al. (2020)	DL	Twitter	Spam	Honeypot 1KS-10KN	Tweet text, users' metadata (content + profile)	CNN + ANN	Accuracy (99.68%)
Alothali, Alashwal, et al. (2021a)	Supervised ML	Twitter	Social	Kaggle	Profile	RF	–
Alothali, Hayawi, et al. (2021b)	Supervised ML	Twitter	Social	Twitter bots accounts (Kaggle)	Metadata	RF + CVAE method	AUC (94.3%)
Al-Qurishi et al. (2018)	DL	Twitter	Sybil	2016 US elections	Profile-, content-, graph-based	FFNN	Accuracy (86%)
Al-Zoubi et al. (2017)	Supervised ML	Twitter	Spam	Ala2017	Binary	J48 DT	Accuracy (94.9)
Andriotis and Takasu (2019)	Supervised ML	Twitter	Spam	Onur2017 Chao2011 Chao2012 Cersi2017	Metadata, content, sentiment	AdaBoost	F1-score (0.95)
Anwar and Yaqub (2020)	Unsupervised ML	Twitter	Social	2019 Canadian elections	Metadata	K-means clustering	–
Attia et al. (2022)	DL	Twitter	Social	CLEF 2019	Content	2D CNN	Accuracy (0.933)
Wu et al. (2020)	Unsupervised ML	Twitter	Social	Varol2017, Clark2016	Metadata, sentiment, friends, content, network, timing	GKDPCA	F1 score (97.56%)
Babu et al. (2021)	Supervised ML	Facebook	Social	Prateek2017	Profile	Bayesian	Efficiency (98%)
Barhate et al. (2020)	Supervised ML	Twitter	Social	Bot repository	Profile based (bot score)	RF	AUC (0.96)
Bazm and Asadpour (2020)	Supervised ML	Instagram	Sybil	Muhammad2020	Behavioral	AdaBoost	Accuracy (95%)
Beğenilmiş and Uskudarli (2018)	Supervised ML	Twitter	Social	Bgenilmis2017	Profile	RF	AUC (0.95)
Bhattacharya et al. (2021)	Supervised ML	Weibo	Sybil	Kaggle	Content	RF	F1-score (0.93)
Bindu et al. (2022)	Supervised ML	Twitter	Sybil	MIB projects	Profile	RF	Mean (0.979812)
Chen et al. (2017a, b)	Supervised ML	Twitter	Spam	Chao2017	Statistical	RF-Lfun	Detection rate (95%)
Cable and Hugh (2019)	DL	Twitter	Social	Littman r2018	Profile	LSTM	Accuracy (0.981)
Cai, Li, and Zeng (2017a)	DL	Twitter	Social	Fred2016	Behavior, content	DBDM	F1-score (88.30%)
Cai, Li, and Zengi (2017b)	DL	Twitter	Social	Fred2016	Behavioral, content	BeDM	F1-score (87.32%,)
Cresci et al. (2016)	Unsupervised ML	Twitter	Spam	StefanoCresci2016	Behavioral	DNA fingerprint	MCC (0.952)
Dan and Jieqi (2017)	Supervised ML	Weibo	Social	Sina Weibo data warehouse	Behavioral	RF	F-measure (0.944)

Table 2 (continued)

References	Technique	Platform	Bot type	Dataset	Feature's type	Best-performing algorithm	Highest result
Daouadi et al. (2019)	Supervised ML	Twitter	Social	Subrahmanian2016, Lee2011	Metadata, profile	Deep forest	Accuracy (97.55%)
David et al. (2017)	Supervised ML	Twitter	Sybil	BotsDeTwitter2016	Content	RF	Accuracy (94%)
Davis et al. (2016)	Supervised ML	Twitter	Social	Clayton2016	Network, user friends, temporal, content, sentiment	RF	AUC (95%)
Dewan and Kumaraguru (2017)	Supervised ML	Facebook	Social	Prateek2017	Entity, content, metadata, link	RF	Accuracy (80%)
Dey et al. (2019)	Supervised ML	Instagram	Sybil	Free4ever1	Profile	RF	Accuracy (92.5%)
Echeverrï£¡a et al. (2018)	Supervised ML	Twitter	Social, Spam	Echeverria2017 Besel2018 Chavoshi2016 Cresci2017, 2015 Yang2012 Subrahmanian2016 Z. Gilani2017	Content, profile	LGBM	Accuracy (97.84%)
Ersahin et al. (2017)	Supervised ML	Twitter	Sybil	Buket2017	User-based	NB w/EMD	Accuracy (90.41%)
Eshraqi et al. (2016)	Supervised ML	Twitter	Spam	Chao2011	Graph, content, time, keywords	DenSream	Accuracy (99%)
Fazil and Abulaish (2018)	Supervised ML	Twitter	Spam	GuofeiGu2017	Community, metadata, content, interaction	RF	F-score (0.979)
Fernquist et al. (2018)	Supervised ML	Twitter	Social	Cresci2015 Z. Gilani2017 O. Varol2017	Metadata, content, time	RF	Accuracy (0.957)
Fonseca Abreu et al. (2020)	Supervised ML	Twitter	Social, Spam	Cresci2017	Profile	RF	AUC (0.9999)
Gao et al. (2020)	DL	Twitter	Sybil	Cersi2015	Content based	(CNN + SELU) + bi-SN-LSTM	F1 score (99.31%)
Güngör et al. (2020)	Supervised ML	Twitter	Spam	Kubra2020	Content, profile	J48	Accuracy (97.2%)
Gupta and Kaushal (2017)	Supervised ML	Facebook	Sybil	Aditi2017	Profile, timeline	RF	Accuracy (94%)
Shukla et al. (2021)	Supervised ML	Twitter	Social	Kaggle	Metadata	RF	AUC (93%)
Hakimi et al. (2019)	Supervised ML	Facebook	Sybil	Ahmad2019	Content-based	KNN	Accuracy (0.829)
Hayawi et al. (2022)	DL	Twitter	Social	Varol 2017 Cresci2017a, b Yang2019–Yang2020b Mazza2019 Gilani2017	Metadata	GLoVe + LSTM	AUC (0.97)
Heidari et al. (2020)	Supervised ML & DL	Twitter	Spam	Cresci2017	Profile	-FFNN -Glove + ELMO(LSTM) + FFNN	Accuracy (0.971) Accuracy (0.981)
Heidari et al. (2021)	Supervised ML	Twitter	Spam	Cresci2017	Sentiment	SVM	F1-score (0.930)
Huang et al. (2016)	Supervised ML	Weibo	Social	Yingying2016	Profile-based	DT Bayesian algorithm	Precision (0.92)

Table 2 (continued)

References	Technique	Platform	Bot type	Dataset	Feature's type	Best-performing algorithm	Highest result
Inuwa-Dutse et al. (2018)	Supervised ML	Twitter	Spam	Honeypot, SPDautomated, SPDmanual	Account, network, user, optimised	GB	AUC (99.93%)
Kantepe and Gañiz (2017)	Supervised ML	Twitter	Social	Mücahit2017	Profile, content, periodic	GBT	Accuracy (86%)
Kenyeres and Kovács (2022)	Supervised ML & DL	Twitter	Social	PAN 2019 bots and gender profiling task	Tweet text, content, non-content	GLoVe + LSTM + Adaboost	Accuracy (0.9)
Kesharwani et al. (2021)	DL	Instagram	Sybil	Kaggle	Profile-based	ANN	Accuracy (93.63%)
Khaled et al. (2019)	Supervised ML	Twitter	Sybil	Cersi2015	User-based	SVM-NN	Accuracy (98%)
Khalil et al. (2020)	Unsupervised ML	Twitter	Social	Mohammad2020	Profile	DBSCAN	Accuracy (97.7%)
Knauth (2019)	Supervised ML	Twitter	Social	Cresci2017	Profile, content, behavior	AdaBoost	Accuracy (0.988)
Koggalahewa et al. (2022)	Unsupervised ML	Twitter	Spam	The fake project Social honey pot HSpam14	Content	SMD	Accuracy (0.98)
Kondeti et al. (2021)	Supervised ML	Twitter	Sybil	–	Account metadata	RF, KNN	Accuracy (98%)
Kudugunta and Ferrara (2018)	DL	Twitter	Spam	Cresci2017	Tweet text, account metadata	Contextual LSTM	AUC (> 96%)
Kumar and Rishiwal (2020)	Supervised ML	Twitter	Spam	Vinay2020	Content	MNB	Accuracy (99%)
Albayati and Altamimi (2020)	Supervised & unsupervised ML	Facebook	Sybil	Mohammed2020	Profile	ID3 DT	Accuracy (97.7%)
Albayati and Altamimi (2019)	Supervised ML	Facebook	Sybil	Mohammed2019	Profile	SVM	Accuracy (98%)
Martin-Gutierrez et al. (2021)	DL	Twitter	Social	David2021	Profile information (metadata), text	Bot-DenseNet + RoBERTa transformer	F1-score (0.77)
Mateen et al. (2017)	Supervised ML	Twitter	Spam	GuofeiGu2017	Profile-based, content-based, graph-based	J48, Decorate	Precision (97.6%)
Mazza et al. (2019)	Unsupervised ML	Twitter	Social	Michele2019	Latent	LSTM	F1 (0.87)
Meshram et al. (2021)	Supervised ML	Instagram	Sybil	Pranay2021	Behavioral. content	RF	Accuracy (98.45%)
Munoz and Paul Guillen Pinto (2020)	Unsupervised ML	Instagram	Sybil	Samuel2020	Metadata	RF	Accuracy (0.96)
Najari et al. (2022)	Supervised ML	Twitter	Social, Spam	Cresci2017	–	GLoVe + LSTM	949/.951—200D
Narayan (2021)	Supervised ML	Twitter	Sybil	Nirdhum2021	Profile	DT	Accuracy (93%)
Oentaryo et al. (2016)	Supervised ML	Twitter	Spam	Richard2016	Numeric, categorical, series	LR	F1-score (0.8228)
Ping and Qin (2019)	DL	Twitter	Social	Cersi2017	Temporal, content	CNN, LSTM	Accuracy (99%)
Prabhu Kavin et al. (2022)	Supervised ML	Twitter	Spam	Collected "the fake project"	User, content	SVM	F-score (0.973)
Pramitha et al. (2021)	Supervised ML	Twitter	Social	Kaggle	Profile	XGBoost	Accuracy (0.891)
Pratama and Rakhmawati (2019)	Supervised ML	Twitter	Social	Pandu2019	Twitter accounts	RF	F-score (0.74)

Table 2 (continued)

References	Technique	Platform	Bot type	Dataset	Feature's type	Best-performing algorithm	Highest result
Purba et al. (2020)	Supervised ML	Instagram	Sybil	Kristo2020	Metadata, media info, engagement, media tags, media similarity	RF	4-classes accuracy (91.76%)
Shukla et al. (2022)	Supervised ML	Twitter	Social	Kaggle	Profile	TweezBot	Accuracy (99.00049%)
Rahman et al. (2021)	Supervised ML	Twitter	Social	Md2021	Behavioral	SPY-BOT	Accuracy (0.901)
Ramalingaiah et al. (2021)	Supervised ML	Twitter	Social	Kaggle	Account	DT + BoW	Accuracy (> 0.99)
Rathore et al. (2018)	Supervised ML	Facebook	Spam	Shailendra2017	Profile-based, content-based	BN	Accuracy (0.984)
Reddy et al. (2021)	Supervised ML	Twitter	Spam	P. Muthi2021	Message content, behavior	EML	Accuracy (87.5)
Ren et al. (2018)	Semi-supervised	Weibo	Spam	Honglin2018	Behavior, content	–	Accuracy (0.936)
Rodrigues et al. (2022)	Supervised ML & DL	Twitter	Spam	Kaggle SMS	Sentiment	LSTM	Spam detection-accuracy (98.74%) Sentiment analysis-accuracy (73.81%)
Rodríguez-Ruiz et al. (2020)	Supervised ML	Twitter	Social Spam	Cresci2017	Content	Bagging-TPMiner	AUC (0.921)
Sadineni (2020)	Supervised ML	Twitter	Spam	Kaggle	Content based	RF	F1 score (0.83%) of spammers
Sahoo and Gupta (2020)	Supervised ML	Facebook	Spam	Somya2020	Profile, content	JRip	Detection rate (99.5%)
Saranya Shree et al. (2021)	Supervised ML	Facebook	Sybil	Instagram influencer dataset	Text, behavioral, graph	SVM + Naïve Bay	Accuracy (91.5%)
Sayyadiharikandeh et al. (2020)	Supervised ML	Twitter	Social	Botometer repository	Metadata, retweet/mention networks, temporal, content information, sentiment	ESC	Accuracy (0.99)
Sedhai and Sun (2018)	Semi-supervised	Twitter	spam	HSpam14	Hashtag, content user, domain	Clustring, NB, LR, RF	Best F1 scores
Sen et al. (2018)	Supervised ML	Instagram	Social	Indira2018	Network effect, internet overlap, liking frequency, influential poster, hashtag features, user-based features	MLP	Precision (83%)
Sengar et al. (2020)	Supervised ML & DL	Twitter	Social	Sandeep2017	User, Tweet metadata	RF, GB	Accuracy (99.54%)
Sheeba et al. (2019)	Supervised ML	Twitter	Spam	Sheeba2019	Content	RF	F1-score (66%)
Sheikhi (2020)	Supervised ML	Instagram	Sybil	Saeid2021	Behavioral, content	Bagging	98.45% accurate classification
Shevtsov et al. (2022)	Supervised ML	Twitter	Social	US 2020 Elections	User	XGBoost	F-score (0.896)
Singh and Banerjee (2019)	Supervised ML	Facebook	Sybil	Yeshwant2019	Textual, categorical, numerical	AdaBoost	F1-score (99%)

Table 2 (continued)

References	Technique	Platform	Bot type	Dataset	Feature's type	Best-performing algorithm	Highest result
Sohrabi and Karimi (2018)	Supervised & Unsupervised ML	Facebook	Spam	Mohammad2017	Metadata	(DB index, DE) + SVM	Accuracy (96.3%)
Thejas et al. (2019)	Supervised ML	Instagram	Sybil	Thejas2019	Text, numeric	RF	Accuracy (97%)
van der Walt and Eloff (2018)	Supervised ML	Twitter	Sybil	Estée2018	Account metadata	RF	F1 score (49.75%)
Varol et al. (2017)	Supervised ML	Twitter	Social	Onur2017	Metadata, content, network	RF	AUC (0.95)
Zhang and Sun (2017)	Supervised ML	Instagram	Spam	Wuxain2017	Content, user profile	RF	Accuracy (96.27%)
Wanda et al. (2020)	DL	Facebook	Social	VirusTotal, PhishTank	Profile	NN	Training loss (0.5058)
Xiao et al. (2015)	Supervised ML	LinkedIn	Sybil	Cao2015	Statistical	RF	AUC (0.95), recall (0.75) at 95% precision
Xu et al. (2021)	DL	Weibo	Spam	MicroblogPCU and weibo-Data (self-collected)	Semantic features from text	ALBERT + Bi-LSTM + self-attention	High accuracy
Wu et al. (2021)	DL	Weibo	Spam	SWLD-20 K	Metadata, interactions, contents, timing	ResNet + BiGRU	Accuracy (0.989)
Chen et al. (2017a, b)	Unsupervised ML	Twitter	Social	Zhouhan2017	–		
Yang et al. (2022)	DL	Weibo	Social	SWLD-20 K	Semantic, metadata, profile	TPBot	F-score (0.984)
Zheng, Wang, et al. (2016a)	Supervised ML	Weibo	Spam	Zinhu2015	Behavior, content	Two-phased Method	Accuracy (90.67%)
Zheng, Zhang, et al. (2016b)	Supervised ML	Weibo	Spam	Xianghan2015	Message content, user behavior	ELM	TRP (99%)

Table 3 Summary of Public Datasets Information

Platform	Datasets	Papers that used it	No. of papers	Bot type
Twitter	Cresci2017	Andriotis and Takasu (2019), Echeverrï£¡a et al. (2018), Fonseca Abreu et al. (2020), Hayawi et al. (2022), Heidari et al. (2020), Heidari et al. (2021), Knauth (2019), Sengar et al. (2020), Kudugunta and Ferrara (2018), Najari et al. (2022), Ping and Qin (2019), Rodríguez-Ruiz et al. (2020)	13	Spam, social
	Cresci2015	Bindu et al. (2022), Echeverrï£¡a et al. (2018), Fernquist et al. (2018), Gao et al. (2020), Khaled et al. (2019), Prabhu Kavin et al. (2022)	5	Sybil, social
	Kaggle	Alothali, Alashwal, et al. (2021a), Alothali, Hayawi et al. (2021b), Shukla et al. (2021), Knauth (2019), Pramitha et al. (2021), Shukla et al. (2022), Ramalingaiah et al. (2021), Rodrigues et al. (2022), Rodríguez-Ruiz et al. (2020), Sadineni (2020)	10	Spam, sybil, social
	Honeypot	Alom et al. (2020), Cai, Li, and Zeng (2017a), Inuwa-Dutse et al. (2018), Koggalahewa et al. (2022)	4	Spam
	The fake project	Koggalahewa et al. (2022), Prabhu Kavin et al. (2022)	2	Spam
	HSpam14	Koggalahewa et al. (2022), Sedhai and Sun (2018)	2	Spam
	Chao2011, Chao2012	Andriotis and Takasu (2019), Eshraqi et al. (2016), Fazil and Abulaish (2018)	3	Spam
	NSCLab	Chen et al. (2017a, b)	1	Spam
	O. Varol2017	Wu et al. (2020), Fernquist et al. (2018), Hayawi et al. (2022)	3	Social
	Z. Gilani2017	Echeverrï£¡a et al. (2018), Fernquist et al. (2018), Hayawi et al. (2022)	3	Social
	Onur2017	Andriotis and Takasu (2019), Varol et al. (2017)	2	Spam, social
	1KS-10KN	Alom et al. (2020)	1	Spam
	Sheeba2019	Sheeba et al. (2019)	1	Spam
	SPDautomate, SPDmanual	Inuwa-Dutse et al. (2018)	1	Spam
	PAN 2019 bots and gender profiling task	Attia et al. (2022), Kenyeres and Kovács (2022)	2	Social
	Yang2019/2020b/2012	Echeverrï£¡a et al. (2018), Hayawi et al. (2022)	2	Social
	GuofeiGu2017	Fazil and Abulaish (2018) Mateen et al. (2017)	2	Spam
	US (2020 elections	Shevtsov et al. (2022)	1	Social
	Littman2018	Cable and Hugh (2019)	1	Social
	Roeder2018	Cable and Hugh (2019)	1	Social
	Botometer repository	Barhate et al. (2020), Khalil et al. (2020), Sayyadiharikandeh et al. (2020)	3	Social
	Subrahmanian2016	Daouadi et al. (2019), Hayawi et al. (2022)	2	Social
	Echeverria2017, Besel2018, Chavoshi2016	Hayawi et al. (2022)	1	Social
	Lee2011	Daouadi et al. (2019)	1	Social
	Mohammad2019	Khalil et al. (2020)	1	Social
	Clark2016	Wu et al. (2020)	1	Social
	Mazza2019	Echeverrï£¡a et al. (2018), Hayawi et al. (2022)	2	Spam, social
	Clayton2016	Davis et al. (2016)	1	Social
	BotsDeTwitter	David et al. (2017)	1	Spam
	Begenilmi2017	Beğenilmiş and Uskudarli (2018)	1	Social
	Abdulrahman2016	Alarifi et al. (2016)	1	Sybil
	Morstatter2016	Cai, Li, and Zengi (2017b)		Social

Table 3 (continued)

Platform	Datasets	Papers that used it	No. of papers	Bot type
Instagram	Kaggle	Kesharwani et al. 2021)	1	Sybil
	Free4ever1	Dey et al. (2019)	1	Sybil
	Fatih2019	Akyon and Esat Kalfaoglu (2019)	1	Spam, sybil
	Instagram influencer dataset	Saranya Shree et al. (2021)	1	Sybil
Facebook	VirusTotal and PhishTank	Wanda et al. (2020)	1	Social
Weibo	MicroblogPCU	Xu et al. (2021)	1	Spam
	Sina Weibo data warehouse	Dan and Jieqi (2017)	1	Social
	SWLD-20 K	Wu et al. (2021), Yang et al. (2022)	2	Spam, social
	Kaggle	Bhattacharya et al. (2021)	1	Sybil

Table 4 Self-collected (private) Datasets and their Collection Methods

Platform	Papers	Data collection method	No. of papers
Twitter	Alarifi et al. (2016), van der Walt and Eloff (2018)	Twitter4j	2
	Reddy et al. (2021)	Tweepy	1
	Alhassun and Rassam (2022), Alothali, Alashwal, et al. (2021a), Al-Qurishi et al. (2018), Al-Zoubi et al. (2017), Anwar and Yaqub (2020), Wu et al. (2020), Barhate et al. (2020), Beğenilmiş and Uskudarli (2018), Bindu et al. (2022), Chen et al. (2017a, b), Cable and Hugh (2019), Daouadi et al. (2019), David et al. (2017), Echeverrï£¡a et al. (2018), Fazil and Abulaish (2018), Fonseca Abreu et al. (2020), Güngör et al. (2020), Kantepe and Gañiz (2017), Martin-Gutierrez et al. (2021), Narayan (2021), Oentaryo et al. (2016), Shukla et al. (2022), Rodrigues et al. (2022), Shevtsov et al. (2022), Pramitha et al. (2021), Varol et al. (2017), Chen et al. (2017a, b)	Twitter API	27
	Cresci et al. (2016), Shukla et al. (2021), Khaled et al. (2019), Pratama and Rakhmawati (2019)	Manually	4
	Ersahin et al. (2017)	Machine learning	1
	Cai, Li, and Zengi (2017b)	Honeypots	2
	Mazza et al. (2019)	Twitter premium search API	1
	Alharthali et al. (2019)	Twitter API	1
	Prabhu Kavin et al. (2022)	REST API	1
Instagram	Meshram et al. (2021), Sheikhi (2020)	Instagram API	2
	Akyon and Esat Kalfaoglu (2019), Bazm and Asadpour (2020), Sen et al. (2018)	Manually	3
	Munoz and Paul Guillen Pinto (2020)	Python	1
	Munoz and Paul Guillen Pinto (2020), Thejas et al. (2019)	Selenium web driver tool	2
	Purba et al. (2020)	3rd-party Instagram websites	1
Facebook	Wanda et al. (2020)	VirusTotal, PhishTank	1
	Babu et al. 2021), Dewan and Kumaraguru (2017), Gupta and Kaushal (2017), Rathore et al. (2018), Singh and Banerjee (2019), Sohrabi and Karimi (2018)	Facebook graph API	6
	Hakimi et al. (2019)	Mockaroo	1
	Albayati and Altamimi (2020), Rathore et al. (2018)	CRAWLER	2
Weibo	Bhattacharya et al. (2021), Ren et al. (2018), Wu et al. (2021), Yang et al. (2022), Zheng, Wang, et al. (2016a)	Web crawler and data scraping	5
	Huang et al. (2016)	Manually	1
LinkedIn	Adikari and Dutta (2020), Xiao et al. (2015)	Manually	2

Table 5 Summary of Features in Social Media Bot Detection using ML

Feature's type	Papers that used it	No. of papers	Platform	Description
Content/language	Akyon and Esat Kalfaoglu (2019), Alarifi et al. (2016), Alhassun and Rassam (2022), Al-Qurishi et al. (2018), Andriotis and Takasu (2019), Attia et al. (2022), Wu et al. (2020), Bhattacharya et al. (2021), Cai, Li, and Zeng (2017a), Cai, Li, and Zengi (2017b), David et al. (2017), Davis et al. (2016), Dewan and Kumaraguru (2017), Echeverrï£¡a et al. (2018), Eshraqi et al. (2016), Fazil and Abulaish (2018), Gao et al. (2020), Güngör et al. (2020), Hakimi et al. (2019), Kenyeres and Kovács (2022), Knauth (2019), Koggalahewa et al. (2022), Kumar and Rishiwal (2020), Mateen et al. (2017), Meshram et al. (2021), Ping and Qin (2019), Prabhu Kavin et al. (2022), Rathore et al. (2018), Reddy et al. (2021), Ren et al. (2018), Rodríguez-Ruiz et al. (2020), Sadineni (2020), Sahoo and Gupta (2020), Sayyadiharikandeh et al. (2020), Sheeba et al. (2019), Sheikhi (2020), Fernquist et al. (2018), Sedhai and Sun (2018), Shevtsov et al. (2022), Varol et al. (2017), Zhang and Sun (2017), Wu et al. (2021), Zheng, Wang, et al. (2016a), Zheng, Zhang, et al. (2016 b)	44	Facebook, Instagram, Twitter, Weibo	Content features are based on linguistic cues computed through NLP, mainly part-of-speech tagging
User (profile)	Adikari and Dutta (2020), Alharthi et al. (2019), Alhassun and Rassam (2022), Al-Qurishi et al. (2018), Al-Zoubi et al. (2017), Babu et al. 2021), Barhate et al. (2020), Beğenilmiş and Uskudarli (2018), Bindu et al. (2022), Cable and Hugh (2019), Daouadi et al. (2019), David et al. (2017), Dey et al. (2019), Echeverrï£¡a et al. (2018), Ersahin et al. (2017), Fonseca Abreu et al. (2020), Güngör et al. (2020), Gupta and Kaushal (2017), Heidari et al. (2020), Huang et al. (2016), Inuwa-Dutse et al. (2018), Kantepe and Gañiz (2017), Kesharwani et al. 2021), Khaled et al. (2019), Khalil et al. (2020), Knauth (2019), Albayati and Altamimi (2020), Albayati and Altamimi (2019), Mateen et al. (2017), Narayan (2021), Prabhu Kavin et al. (2022), Shukla et al. (2022), Rahman et al. (2021), Ramalingaiah et al. (2021), Rathore et al. (2018), Sahoo and Gupta (2020), Sedhai and Sun (2018), Sen et al. (2018), Sengar et al. (2020), Zhang and Sun (2017), Wanda et al. (2020), Yang et al. (2022)	42	Facebook, Instagram, Twitter, LinkedIn, Weibo	User features are based on properties of user account and users relationships

Table 5 (continued)

Feature's type	Papers that used it	No. of papers	Platform	Description
Metadata	Alom et al. (2020), Alothali, Alashwal, et al. (2021a), Alothali, Hayawi, et al. (2021b), Andriotis and Takasu (2019), Anwar and Yaqub (2020), Wu et al. (2020), Daouadi et al. (2019), David et al. (2017), Dewan and Kumaraguru (2017), Fazil and Abulaish (2018), Fernquist et al. (2018), Shukla et al. (2021), Hayawi et al. (2022), Kondeti et al. (2021), Kudugunta and Ferrara (2018), Martin-Gutierrez et al. (2021), Meshram et al. (2021), Munoz and Paul Guillen Pinto (2020), Pramitha et al. (2021), Pratama and Rakhmawati (2019), Purba et al. (2020), Sayyadiharikandeh et al. (2020), Sohrabi and Karimi (2018), van der Walt and Eloff (2018), Varol et al. (2017), Wu et al. (2021), Yang et al. (2022)	27	Instagram, Twitter, Weibo	User meta data features is information regarding the profile's characteristics. Locating an information source via metadata is known to be effective
Behavioural	Alharthi et al. (2019), Cai, Li, and Zengi (2017b), Meshram et al. (2021), Rahman et al. (2021), Akyon and Esat Kalfaoglu (2019), Sheikhi (2020), Knauth (2019), Zheng, Zhang, et al. (2016b), Saranya Shree et al. (2021), Cai, Li, and Zeng (2017a), Bazm and Asadpour (2020), Cresci et al. (2016), Dan and Jieqi (2017), Ren et al. (2018), Zheng, Wang, et al. (2016a), Reddy et al. (2021)	16	Facebook, Instagram, Twitter, Weibo	Behavior features are calculated by statistical properties from the data
Network (community/interaction)	Fazil and Abulaish (2018), Varol et al. (2017), Davis et al. (2016), Inuwa-Dutse et al. (2018), Sen et al. (2018), Sayyadiharikandeh et al. (2020), Wu et al. (2021), Wu et al. (2020)	8	Instagram, Twitter, Weibo	Different aspects of information diffusion patterns are captured by network features. They are built on retweets, mentions, hashtag occurrences etc On the basis of the frequency, all networks are weighed
Sentiment	Davis et al. (2016), Yang et al. (2022), Heidari et al. (2021), Rodrigues et al. (2022), Andriotis and Takasu (2019), Sayyadiharikandeh et al. (2020), Xu et al. (2021), Wu et al. (2020)	8	Twitter, Weibo	General-purpose and Twitter specific sentiment analysis algorithms are used to build sentiment features
Timing/temporal	Davis et al. (2016), Gupta and Kaushal (2017), Fernquist et al. (2018), Eshraqi et al. (2016), Wu et al. (2021), Sayyadiharikandeh et al. (2020), Ping and Qin (2019)	7	Facebook, Twitter, Weibo	Time features includes statistics of time such as retweets, consecutive tweets, statistics for the time between post etc Temporal capture timing patterns of content generation and consumption, such as tweet rate and inter-tweet time distribution
Graph	Eshraqi et al. (2016), Mateen et al. (2017), Saranya Shree et al. (2021), Al-Qurishi et al. (2018), Alhassun and Rassam (2022)	5	Instagram, Twitter	Graph features are extracted by modelling the social media platform as a social graph model
Numeric/categorical/textual/series	Oentaryo et al. (2016), Thejas et al. (2019), Saranya Shree et al. (2021), Singh and Banerjee (2019)	4	Facebook, Twitter	Examples: account ID, post ID

Table 5 (continued)

Feature's type	Papers that used it	No. of papers	Platform	Description
Statistical	Davis et al. (2016), Xiao et al. (2015), Chen et al. (2017a, b)	3	LinkedIn, Twitter	Examples: min, max, quartiles, Mean and variance, median, moments, and entropy etc
User friends	Davis et al. (2016), Wu et al. (2020)	2	Twitter	Descriptive statistics relative to an account's social contacts are included in user friend features
Media and engagement	Purba et al. (2020)	1	Instagram	Examples: average caption length, non-image percentage, engagement rate (like/comm.), location tag percentage, promotional keywords, followers' keywords, post interval
Entity and link	Dewan and Kumaraguru (2017)	1	Facebook	Examples: gender, parameter length, page category, has username, hyphen count username length, no. of subdomains, name length, no. of words in name, locale, likes on page, has HTTP/HTTPS, parameters count, path length etc
Keywords	Eshraqi et al. (2016)	1	Twitter	Examples: presence of the words "chat" and "naughty" in biography have a direct relationship with spam
Internet overlap, hashtag features	Sedhai and Sun (2018), Sen et al. (2018)	2	Instagram, Twitter	Interest overlap features include overlap between two users such as topical affinity
Periodic	Kantepe and Gañiz (2017)	1	Twitter	Examples: screen name, location, is protected, profile image

Table 6 ML Algorithms in Reviewed Papers

Algorithm	Papers that applied it	No. of papers	Highest classifier bot type detected	Performance measure (accuracy) (%)
RF	Alarifi et al. (2016), Alothali, Alashwal, et al. (2021a), Alothali, Hayawi, et al. (2021b), Andriotis and Takasu (2019), Barhate et al. (2020), Bazm and Asadpour (2020), Beğenilmiş and Uskudarli (2018), Bhattacharya et al. (2021), Bindu et al. (2022), Chen et al. (2017a, b), Dan and Jieqi (2017), David et al. (2017), Davis et al. (2016), Dewan and Kumaraguru (2017), Dey et al. (2019), Echeverrï£¡a et al. (2018), Fazil and Abulaish (2018), Fonseca Abreu et al. (2020), Gupta and Kaushal (2017), Shukla et al. (2021), Heidari et al. (2020), Heidari et al. (2021), Knauth (2019), Kondeti et al. (2021), Meshram et al. (2021), Munoz and Paul Guillen Pinto (2020), Narayan (2021), Oentaryo et al. (2016), Pratama and Rakhmawati (2019), Purba et al. (2020), Shukla et al. (2022), Rathore et al. (2018), Rodrigues et al. (2022), Sadineni (2020), Sedhai and Sun (2018), Sen et al. (2018), Sengar et al. (2020), Sheeba et al. (2019), Singh and Banerjee (2019), Thejas et al. (2019), Varol et al. (2017), van der Walt and Eloff (2018), Zhang and Sun (2017), Xiao et al. (2015)	44	Social	99.545
SVM	Adikari and Dutta (2020), Akyon and Esat Kalfaoglu (2019), Andriotis and Takasu (2019), Bazm and Asadpour (2020), Beğenilmiş and Uskudarli (2018), Bindu et al. (2022), David et al. (2017), Fernquist et al. (2018), Fonseca Abreu et al. (2020), Gupta and Kaushal (2017), Hakimi et al. (2019), Heidari et al. (2021), Khaled et al. (2019), Kantepe and Gañiz (2017), Knauth (2019), Kumar and Rishiwal (2020), Albayati and Altamimi (2019), Meshram et al. (2021), Munoz and Paul Guillen Pinto (2020), Oentaryo et al. (2016), Prabhu Kavin et al. (2022), Shukla et al. (2022), Rahman et al. (2021), Rathore et al. (2018), Ren et al. (2018), Rodrigues et al. (2022), Rodríguez-Ruiz et al. (2020), Sadineni (2020), Saranya Shree et al. (2021), Sen et al. (2018), Sheikhi (2020), Sohrabi and Karimi (2018), Thejas et al. (2019), Xiao et al. (2015), Zheng, Wang, et al. (2016a)	35	Sybil	98
NB	Akyon and Esat Kalfaoglu (2019), Andriotis and Takasu (2019), Babu et al. 2021), Bhattacharya et al. (2021), David et al. (2017), Ersahin et al. (2017), Fernquist et al. (2018), Fonseca Abreu et al. (2020), Güngör et al. (2020), Gupta and Kaushal (2017), Huang et al. (2016), Albayati and Altamimi (2019), Mateen et al. (2017), Munoz and Paul Guillen Pinto (2020), Oentaryo et al. (2016), Ren et al. (2018), Rodrigues et al. (2022), Rodríguez-Ruiz et al. (2020), Saranya Shree et al. (2021), Sedhai and Sun (2018), Sheikhi (2020), Purba et al. (2020), Thejas et al. (2019), Zheng, Wang, et al. (2016a)	24	Sybil	90.4
ANN	Adikari and Dutta (2020), Akyon and Esat Kalfaoglu (2019), Alarifi et al. (2016), Alom et al. (2020), Al-Qurishi et al. (2018), Hakimi et al. (2019), Heidari et al. (2020), Heidari et al. (2021), Kesharwani et al. 2021), Khaled et al. (2019), Meshram et al. (2021), Munoz and Paul Guillen Pinto (2020), Shukla et al. (2022), Sen et al. (2018), Thejas et al. (2019), Yang et al. (2022)	17	Sybil, Spam	94
DT	Bazm and Asadpour (2020), David et al. (2017), Echeverrï£¡a et al. (2018), Fazil and Abulaish (2018), Albayati and Altamimi (2019), Munoz and Paul Guillen Pinto (2020), Narayan (2021), Shukla et al. (2022), Ramalingaiah et al. (2021), Rodrigues et al. (2022), Sengar et al. (2020), Varol et al. (2017), Zheng, Wang, et al. (2016a)	13	Social	99
AdaBoost	Andriotis and Takasu (2019), Bazm and Asadpour (2020), Echeverrï£¡a et al. (2018), Fernquist et al. (2018), Heidari et al. (2020), Kenyeres and Kovács (2022), Knauth (2019), Munoz and Paul Guillen Pinto (2020), Sahoo and Gupta (2020), Sen et al. (2018), Sengar et al. (2020), Singh and Banerjee (2019), Varol et al. (2017)	13	Social	98.8

Table 6 (continued)

Algorithm	Papers that applied it	No. of papers	Highest classifier bot type detected	Performance measure (accuracy) (%)
KNN	Al-Zoubi et al. (2017), Andriotis and Takasu (2019), Bazm and Asadpour (2020), Fonseca Abreu et al. (2020), Gupta and Kaushal (2017), Hakimi et al. (2019), Kondeti et al. (2021), Albayati and Altamimi (2019), Rathore et al. (2018), Sengar et al. (2020), Thejas et al. (2019)	11	Sybil	98
J48 (DT)	Al-Zoubi et al. (2017), Güngör et al. (2020), Gupta and Kaushal (2017), Mateen et al. (2017), Rathore et al. (2018), Ren et al. (2018), Sahoo and Gupta (2020), Sheikhi (2020), Purba et al. (2020)	9	Spam	97.6
MLP	Al-Zoubi et al. (2017), Knauth (2019), Munoz and Paul Guillen Pinto (2020), Purba et al. (2020), Sen et al. (2018), Sengar et al. (2020), Sheikhi (2020)	7	Social	83
CNN	Alhassun and Rassam (2022), Alom et al. (2020), Attia et al. (2022), Cai, Li, and Zeng (2017a), Cai, Li, and Zengi (2017b), Gao et al. (2020), Martin-Gutierrez et al. (2021), Ping and Qin (2019), Wu et al. (2021)	9	Spam	99.68
Gradient boost	Bhattacharya et al. (2021), Echeverrï£¡a et al. (2018), Inuwa-Dutse et al. (2018), Kantepe and Gañiz (2017), Sengar et al. (2020), Singh and Banerjee (2019)	6	Social	99.54
LSTM (RNN)	Cai, Li, and Zeng (2017a), Cai, Li, and Zengi (2017b), Hayawi et al. (2022), Kenyeres and Kovács (2022), Mazza et al. (2019), Kudugunta and Ferrara (2018), Ping and Qin (2019), Wanda et al. (2020)	8	Sybil	99.31
BN	Al-Zoubi et al. (2017), Fazil and Abulaish (2018), Gupta and Kaushal (2017), Rathore et al. (2018), Zheng, Wang, et al. (2016a)	5	Spam	98.4
MNB	Kantepe and Gañiz (2017), Rodrigues et al. (2022), Kumar and Rishiwal (2020), Narayan (2021), Sengar et al. (2020)	5	Spam	99
XGBoost	Pramitha et al. (2021), Sen et al. (2018), Shevtsov et al. (2022), Singh and Banerjee (2019)	4	Social	89.6
JRip (DT)	Gupta and Kaushal (2017), Rathore et al. (2018), Sahoo and Gupta (2020)	3	Spam	99.5
Bi-LSTM (RNN)	Heidari et al. (2020), Xu et al. (2021)	2	Spam	98.1
BiGRU (RNN)	Wu et al. (2021), Yang et al. (2022)	2	Social	98.87
ID3 (DT)	Albayati and Altamimi (2020)	1	Sybil	97.7
GNB	Kantepe and Gañiz (2017)	1	Social	86
ELM	Zheng, Zhang, et al. (2016b)	1	Spam	99
bi-SN-LSTM	Gao et al. (2020)	1	Spam	F1(99.31)
DenStream	Eshraqi et al. (2016)	1	Spam	99
K-means	Koggalahewa et al. (2022)	1	Spam	96.9
DBSCAN	Mazza et al. (2019)	1	Social	97.7

Author contributions Conceptualization, MA, RZ, AS, FA, AS, DA; Methodology, MA, RZ, AS, FA, AS, DA; Formal Analysis, MA, RZ, AS, FA, AS, DA; Writing-Reviewing, MA, RZ, AS, FA, AS, DA; Project Administration, MA. All authors have read and agreed to the published version of the manuscript.

Funding We would like to thank SAUDI ARAMCO Cybersecurity Chair at Imam Abdulrahman Bin Faisal University (IAU) for supporting and funding this research work.

Declarations

Conflict of interest The authors declare that there is no conflict of interest.

References

Adikari S, Dutta K (2020) Identifying fake profiles in LinkedIn

Akyon FC, Esat Kalfaoglu M (2019) Instagram fake and automated account detection. In: Proceedings—2019 innovations in intelligent systems and applications conference, ASYU 2019. https://doi.org/10.1109/ASYU48272.2019.8946437

Alarifi A, Alsaleh M, Al-Salman AM (2016) Twitter turing test: identifying social machines. Inf Sci. https://doi.org/10.1016/j.ins.2016.08.036

Albayati MB, Altamimi AM (2019) An empirical study for detecting fake Facebook profiles using supervised mining techniques. Inf Slovenia. https://doi.org/10.31449/inf.v43i1.2319

Albayati M, Altamimi A (2020) MDFP: a machine learning model for detecting fake Facebook profiles using supervised and unsupervised mining techniques. Int J Simul Syst Sci Technol. https://doi.org/10.5013/ijssst.a.20.01.11

Aldayel A, Magdy W (2022) Characterizing the role of bots' in polarized stance on social media. Soc Netw Anal Mining. https://doi.org/10.1007/s13278-022-00858-z

Alharthi R, Alhothali A, Moria K (2019) Detecting and characterizing Arab spammers campaigns in Twitter. Proc Comput Sci 163:248–256. https://doi.org/10.1016/j.procs.2019.12.106

Alhassun AS, Rassam MA (2022) A combined text-based and metadata-based deep-learning framework for the detection of spam accounts on the social media platform Twitter. Processes. https://doi.org/10.3390/pr10030439

Ali A, Syed A (2022) Cyberbullying detection using machine learning. Pak J Eng Technol 3(2):45–50. https://doi.org/10.51846/vol3iss2pp45-50

Aljabri M, Aljameel SS, Mohammad RMA, Almotiri SH, Mirza S, Anis FM, Aboulnour M, Alomari DM, Alhamed DH, Altamimi HS (2021a) Intelligent techniques for detecting network attacks: review and research directions. In Sens. https://doi.org/10.3390/s21217070

Aljabri M, Chrouf SM, Alzahrani NA, Alghamdi L, Alfehaid R, Alqarawi R, Alhuthayfi J, Alduhailan N (2021b) Sentiment analysis of Arabic tweets regarding distance learning in Saudi Arabia during the covid-19 pandemic. Sensors 21(16):5431. https://doi.org/10.3390/s21165431

Aljabri M, Altamimi HS, Albelali SA, Al-Harbi M, Alhuraib HT, Alotaibi NK, Alahmadi AA, Alhaidari F, Mohammad RM, Salah K (2022a) Detecting malicious URLs using machine learning techniques: review and research directions. IEEE Access 10:121395–121417. https://doi.org/10.1109/access.2022.3222307

Aljabri M, Alhaidari F, Mohammad RM, Mirza S, Alhamed DH, Altamimi HS, Chrouf SM (2022b) An assessment of lexical, network, and content-based features for detecting malicious urls using machine learning and deep learning models. Comput Intell Neurosci 2022:1–14. https://doi.org/10.1155/2022/3241216

Aljabri M, Alahmadi AA, Mohammad RM, Aboulnour M, Alomari DM, Almotiri SH (2022c) Classification of firewall log data using multiclass machine learning models. Electronics 11(12):1851. https://doi.org/10.3390/electronics11121851

Aljabri M, Mirza S (2022) Phishing attacks detection using machine learning and Deep Learning Models. In: 2022 7th international conference on data science and machine learning applications (CDMA). https://doi.org/10.1109/cdma54072.2022.00034

Alom Z, Carminati B, Ferrari E (2020) A deep learning model for Twitter spam detection. Online Soc Netw Media. https://doi.org/10.1016/j.osnem.2020.100079

Alothali E, Alashwal H, Salih M, Hayawi K (2021a) Real time detection of social bots on Twitter using machine learning and Apache Kafka. In: 2021a 5th cyber security in networking conference, CSNet 2021a. https://doi.org/10.1109/CSNet52717.2021.9614282

Alothali E, Hayawi K, Alashwal H (2021b) Hybrid feature selection approach to identify optimal features of profile metadata to detect social bots in Twitter. Soc Netw Anal Mining. https://doi.org/10.1007/s13278-021-00786-4

Alothali E, Zaki N, Mohamed EA, Alashwal H (2019) Detecting social bots on Twitter: a literature review. In: Proceedings of the 2018 13th international conference on innovations in information technology, IIT 2018. https://doi.org/10.1109/INNOVATIONS.2018.8605995

Al-Qurishi M, Alrubaian M, Rahman SMM, Alamri A, Hassan MM (2018) A prediction system of Sybil attack in social network using deep-regression model. Future Gener Comput Syst. https://doi.org/10.1016/j.future.2017.08.030

Al-Zoubi AM, Alqatawna J, Faris H (2017) Spam profile detection in social networks based on public features. In: 2017 8th international conference on information and communication systems, ICICS 2017. https://doi.org/10.1109/IACS.2017.7921959

Andriotis P, Takasu A (2019) Emotional bots: content-based spammer detection on social media. In: 10th IEEE international workshop on information forensics and security, WIFS 2018. https://doi.org/10.1109/WIFS.2018.8630760

Anwar A, Yaqub U (2020) Bot detection in twitter landscape using unsupervised learning. ACM Int Conf Proc Series. https://doi.org/10.1145/3396956.3401801

Attia SM, Mattar AM, Badran KM (2022) Bot detection using multi-input deep neural network model in social media. In: 2022 13th international conference on electrical engineering (ICEENG), p 71–75. https://doi.org/10.1109/ICEENG49683.2022.9781863

Barhate S, Mangla R, Panjwani D, Gatkal S, Kazi F (2020) Twitter bot detection and their influence in hashtag manipulation. In: 2020 IEEE 17th India council international conference, INDICON 2020. https://doi.org/10.1109/INDICON49873.2020.9342152

Bazm, M. and Asadpour, M. (2020) "Behavioral Modeling of Persian Instagram Users to detect Bots." Available at: https://doi.org/10.48550/arXiv.2008.03951

Beğenilmiş E, Uskudarli S (2018) Organized behavior classification of tweet sets using supervised learning methods. ACM Int Conf Proc Series. https://doi.org/10.1145/3227609.3227665

Benkler Y et al (2017) Partisanship, propaganda, and disinformation: online media and the 2016 U.S. presidential election, search issue lab. Issue lab. Available at: https://search.issuelab.org/resource/partisanship-propaganda-and-disinformation-online-media-and-the-2016-u-s-presidential-election.html. Accessed 9 Oct 2022

Bhattacharya A, Bathla R, Rana A, Arora G (2021) Application of machine learning techniques in detecting fake profiles on social media. In: 2021 9th international conference on reliability, Infocom technologies and optimization (trends and future directions), ICRITO 2021. https://doi.org/10.1109/ICRITO51393.2021.9596373

Bindu K et al (2022) Detection of fake accounts in Twitter using data science. Int Res J Mod Eng Technol Sci 4(5), pp. 3552-3556.

Cable, J. and Hugh, G. (2019) Bots in the Net: Applying Machine Learning to Identify Social Media Trolls. rep. Available at: http://cs229.stanford.edu/proj2019spr/report/74.pdf

Caers R, de Feyter T, de Couck M, Stough T, Vigna C, du Bois C (2013) Facebook: a literature review. New Media Soc. https://doi.org/10.1177/1461444813488061

Cai C, Li L, Zeng D (2017a) Detecting social bots by jointly modeling deep behavior and content information. Int Conf Inf Knowl Manag Proc Part F131841. https://doi.org/10.1145/3132847.3133050

Cai C, Li L, Zengi D (2017b) Behavior enhanced deep bot detection in social media. In: 2017b IEEE international conference on intelligence and security informatics: security and big data, ISI 2017b. https://doi.org/10.1109/ISI.2017.8004887

Cao F, Ester M, Qian W, Zhou A (2006) Density-based clustering over an evolving data stream with noise. In: Proceedings of the sixth SIAM international conference on data mining, 2006. https://doi.org/10.1137/1.9781611972764.29

Carminati B, Ferrari E, Heatherly R, Kantarcioglu M, Thuraisingham B (2011) Semantic web-based social network access control. Comput Secur 30(2–3):108–115. https://doi.org/10.1016/j.cose.2010.08.003

Chen C, Wang Y, Zhang J, Xiang Y, Zhou W, Min G (2017a) Statistical features-based real-time detection of drifted Twitter spam. IEEE Trans Inf Forensics Secur. https://doi.org/10.1109/TIFS.2016.2621888

Chen Z, Tanash RS, Stoll R, Subramanian D (2017b) Hunting malicious bots on twitter: an unsupervised approach. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 10540 LNCS. https://doi.org/10.1007/978-3-319-67256-4_40

Cresci S, di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2015) Fame for sale: efficient detection of fake Twitter followers. Decis Support Syst. https://doi.org/10.1016/j.dss.2015.09.003

Cresci S, di Pietro R, Petrocchi M, Spognardi A, Tesconi M (2016) DNA-inspired online behavioral modeling and its application to spambot detection. IEEE Intell Syst. https://doi.org/10.1109/MIS.2016.29

Cresci S, Spognardi A, Petrocchi M, Tesconi M, di Pietro R (2017) The paradigm-shift of social spambots: evidence, theories, and tools for the arms race. In: 26th international world wide web conference 2017, WWW 2017 companion. https://doi.org/10.1145/3041021.3055135

Dan J, Jieqi T (2017) Study of bot detection on Sina-Weibo based on machine learning. In: 14th international conference on services systems and services management, ICSSSM 2017—Proceedings. https://doi.org/10.1109/ICSSSM.2017.7996292

Daouadi KE, Rebaï RZ, Amous I (2019) Bot detection on online social networks using deep forest. Adv Intell Syst Comput. https://doi.org/10.1007/978-3-030-19810-7_30

David I, Siordia OS, Moctezuma D (2017) Features combination for the detection of malicious Twitter accounts. In: 2016 IEEE international autumn meeting on power, electronics and computing, ROPEC 2016. https://doi.org/10.1109/ROPEC.2016.7830626

Davis, C. A., Varol, O., Ferrara, E., Flammini, A., & Menczer, F. (2016). BotOrNot. Proceedings of the 25th International Conference Companion on World Wide Web - WWW . https://doi.org/10.1145/2872518.2889302

Derhab A, Alawwad R, Dehwah K, Tariq N, Khan FA, Al-Muhtadi J (2021) Tweet-based bot detection using big data analytics. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3074953

Dewan P, Kumaraguru P (2017) Facebook Inspector (FbI): towards automatic real-time detection of malicious content on Facebook. Soc Netw Anal Mining. https://doi.org/10.1007/s13278-017-0434-5

Dey A, Reddy H, Dey M, Sinha N (2019) Detection of fake accounts in Instagram using machine learning. Int J Comput Sci Inf Technol. https://doi.org/10.5121/ijcsit.2019.11507

Dinath W (2021) Linkedin: a link to the knowledge economy. In: Proceedings of the European conference on knowledge management, ECKM. https://doi.org/10.34190/EKM.21.178

Echeverrï£¡a J, de Cristofaro E, Kourtellis N, Leontiadis I, Stringhini G, Zhou S (2018) LOBO. In: Proceedings of the 34th annual computer security applications conference, p 137–146. https://doi.org/10.1145/3274694.3274738

Ersahin B, Aktas O, Kilinc D, Akyol C (2017) Twitter fake account detection. Int Conf Comput Sci Eng (UBMK) 2017:388–392. https://doi.org/10.1109/UBMK.2017.8093420

Eshraqi N, Jalali M, Moattar MH (2016) Detecting spam tweets in Twitter using a data stream clustering algorithm. In: 2nd international congress on technology, communication and knowledge, ICTCK 2015. https://doi.org/10.1109/ICTCK.2015.7582694

Ezarfelix J, Jeffrey N, Sari N (2022) Systematic literature review: Instagram fake account detection based on machine learning. Eng Math Comput Sci J. https://doi.org/10.21512/emacsjournal.v4i1.8076

Fazil M, Abulaish M (2018) A hybrid approach for detecting automated spammers in Twitter. IEEE Trans Inf Forensics Secur. https://doi.org/10.1109/TIFS.2018.2825958

Fernquist J, Kaati L, Schroeder R (2018) Political bots and the Swedish general election. In: 2018 IEEE international conference on intelligence and security informatics, ISI 2018. https://doi.org/10.1109/ISI.2018.8587347

Ferrara, E. (2018). Measuring Social Spam and the Effect of Bots on Information Diffusion in Social Media. Computational Social Sciences, 229-255. https://doi.org/10.1007/978-3-319-77332-2_13

Ferrara, E. (2020). What types of COVID-19 conspiracies are populated by Twitter bots?. First Monday, 25(6). https://doi.org/10.5210/fm.v25i6.10633

Fonseca Abreu JV, Ghedini Ralha C, Costa Gondim JJ (2020) Twitter bot detection with reduced feature set. In: Proceedings—2020 IEEE international conference on intelligence and security informatics, ISI 2020. https://doi.org/10.1109/ISI49825.2020.9280525

Gannarapu S, Dawoud A, Ali RS, Alwan A (2020) Bot detection using machine learning algorithms on social media platforms. In: CITISIA 2020—IEEE conference on innovative technologies in intelligent systems and industrial applications, proceedings. https://doi.org/10.1109/CITISIA50690.2020.9371778

Gao T, Yang J, Peng W, Jiang L, Sun Y, Li F (2020) A content-based method for Sybil detection in online social networks via deep learning. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2975877

Gheewala S, Patel R (2018) Machine learning based twitter spam account detection: a review. In: Proceedings of the 2nd international conference on computing methodologies and communication, ICCMC 2018. https://doi.org/10.1109/ICCMC.2018.8487992

Gilani Z, Wang L, Crowcroft J, Almeida M, Farahbakhsh R (2016) Stweeler: a framework for Twitter bot analysis. In: WWW 2016 companion—proceedings of the 25th international conference on World Wide Web. https://doi.org/10.1145/2872518.2889360

Gilani Z, Farahbakhsh R, Tyson G, Wang L, Crowcroft J (2017) Of bots and humans (on Twitter). In: Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017, p 349–354. https://doi.org/10.1145/3110025.3110090

Gorwa R, Guilbeault D (2020) Unpacking the social media bot: a typology to guide research and policy. Policy Internet 12(2):225–248. https://doi.org/10.1002/poi3.184

Güngör KN, Ayhan Erdem O, Doğru İA (2020) Tweet and account based spam detection on Twitter, p 898–905. https://doi.org/10.1007/978-3-030-36178-5_79

Guofei Gu (no date) Welcome to Guofei Gu's Homepage. Available at: https://people.engr.tamu.edu/guofei/index.html. Accessed 12 Oct 2022

Gupta A, Kaushal R (2017) Towards detecting fake user accounts in facebook. In: ISEA Asia security and privacy conference 2017, ISEASP 2017. https://doi.org/10.1109/ISEASP.2017.7976996

Hakimi AN, Ramli S, Wook M, Mohd Zainudin N, Hasbullah NA, Abdul Wahab N, Mat Razali NA (2019) Identifying fake account in facebook using machine learning. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 11870 LNCS. https://doi.org/10.1007/978-3-030-34032-2_39

Hayawi K, Mathew S, Venugopal N, Masud MM, Ho PH (2022) DeeProBot: a hybrid deep neural network model for social bot detection based on user profile data. Soc Netw Anal Mining. https://doi.org/10.1007/s13278-022-00869-w

Heidari M, Jones JH, Uzuner O (2020) Deep contextualized word embedding for text-based online user profiling to detect social bots on Twitter. In: IEEE international conference on data mining workshops, ICDMW, 2020-November. https://doi.org/10.1109/ICDMW51313.2020.00071

Heidari M, Jones JHJ, Uzuner O (2021) An empirical study of machine learning algorithms for social media bot detection. In: 2021 IEEE international IOT, electronics and mechatronics conference, IEMTRONICS 2021—Proceedings. https://doi.org/10.1109/IEMTRONICS52119.2021.9422605

Huang, Y., Zhang, M., Yang, Y., Gan, S., & Zhang, Y. (2016) The Weibo Spammers' Identification and Detection based on Bayesian-algorithm. Proceedings of the 2016 2nd Workshop on Advanced Research and Technology in Industry Applications. https://doi.org/10.2991/wartia-16.2016.271

Inuwa-Dutse I, Liptrott M, Korkontzelos I (2018) Detection of spam-posting accounts on Twitter. Neurocomputing. https://doi.org/10.1016/j.neucom.2018.07.044

Kantartopoulos P, Pitropakis N, Mylonas A, Kylilis N (2020) Exploring adversarial attacks and defences for fake Twitter account detection. Technologies. https://doi.org/10.3390/technologies8040064

Kantepe M, Gañiz MC (2017) Preprocessing framework for Twitter bot detection. In: 2nd international conference on computer science and engineering, UBMK 2017. https://doi.org/10.1109/UBMK.2017.8093483

Kaplan AM, Haenlein M (2010) Users of the world, unite! The challenges and opportunities of social media. Bus Horiz. https://doi.org/10.1016/j.bushor.2009.09.003

Kenyeres A, Kovács G (2022) "Conference: XVIII. Conference on hungarian computational linguistics." Available at: https://www.researchgate.net/publication/358801180_Twitter_bot_detection_using_deep_learning

Kesharwani M, Kumari S, Niranjan V (2021) "Detecting fake social media account using deep neural networking. Int Res J Eng Technol (IRJET), 8(7), pp. 1191-1197.

Khaled S, El-Tazi N, Mokhtar HMO (2019) Detecting fake accounts on social media. In: Proceedings—2018 IEEE international conference on big data, big data 2018. https://doi.org/10.1109/BigData.2018.8621913

Khalil H, Khan MUS, Ali M (2020) Feature selection for unsupervised bot detection. In: 2020 3rd international conference on computing, mathematics and engineering technologies: idea to innovation for building the knowledge economy, ICoMET 2020. https://doi.org/10.1109/iCoMET48670.2020.9074131

Knauth J (2019) Language-agnostic twitter bot detection. In: International conference recent advances in natural language processing, RANLP, 2019-September. https://doi.org/10.26615/978-954-452-056-4_065

Koggalahewa D, Xu Y, Foo E (2022) An unsupervised method for social network spammer detection based on user information interests. J Big Data. https://doi.org/10.1186/s40537-021-00552-5

Kolomeets M, Chechulin A (2021) Analysis of the malicious bots market. In: Conference of open innovation association, FRUCT, 2021-May. https://doi.org/10.23919/FRUCT52173.2021.9435421

Kondeti P, Yerramreddy LP, Pradhan A, Swain G (2021) Fake account detection using machine learning, p 791–802. https://doi.org/10.1007/978-981-15-5258-8_73

Kudugunta S, Ferrara E (2018) Deep neural networks for bot detection. Inf Sci. https://doi.org/10.1016/j.ins.2018.08.019

Kumar G, Rishiwal V (2020) Machine learning for prediction of malicious or SPAM users on social networks. Int J Sci Technol Res, 9(2), pp. 926-932

Lee K, Eoff BD, Caverlee J (2006) Seven months with the devils: a long-term study of content polluters on Twitter. Icwsm 2011

Mahesh, B. (2020) "Machine Learning Algorithms - A Review," International Journal of Science and Research (IJSR), 9(1), pp. 381–386. Available at: https://doi.org/10.21275/ART20203995.

Martin-Gutierrez D, Hernandez-Penaloza G, Hernandez AB, Lozano-Diez A, Alvarez F (2021) A deep learning approach for robust detection of bots in Twitter using transformers. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3068659

Mateen M, Iqbal MA, Aleem M, Islam MA (2017) A hybrid approach for spam detection for Twitter. In: Proceedings of 2017 14th international bhurban conference on applied sciences and technology, IBCAST 2017. https://doi.org/10.1109/IBCAST.2017.7868095

Mazza M, Cresci S, Avvenuti M, Quattrociocchi W, Tesconi M (2019) RTbust: exploiting temporal patterns for botnet detection on twitter. In: WebSci 2019—proceedings of the 11th ACM conference on web science. https://doi.org/10.1145/3292522.3326015

Meshram EP, Bhambulkar R, Pokale P, Kharbikar K, Awachat A (2021) Automatic detection of fake profile using machine learning on Instagram. Int J Sci Res Sci Technol. https://doi.org/10.32628/ijsrst218330

Morstatter F, Wu L, Nazer TH, Carley KM, Liu H (2016) A new approach to bot detection: striking the balance between precision and recall. In: 2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM), p 533–540. https://doi.org/10.1109/ASONAM.2016.7752287

Munoz SD, Paul Guillen Pinto E (2020) A dataset for the detection of fake profiles on social networking services. In: Proceedings—2020 international conference on computational science and computational intelligence, CSCI 2020. https://doi.org/10.1109/CSCI51800.2020.00046

Najari S, Salehi M, Farahbakhsh R (2022) GANBOT: a GAN-based framework for social bot detection. Soc Netw Anal Mining. https://doi.org/10.1007/s13278-021-00800-9

Narayan N (2021) Twitter bot detection using machine learning algorithms. In: 2021 4th international conference on electrical, computer and communication technologies, ICECCT 2021. https://doi.org/10.1109/ICECCT52121.2021.9616841

Naveen Babu M, Anusha G, Shivani A, Kalyani C, Meenakumari J (2021) Fake profile identification using machine learning. Int J Recent Adv Multidiscip Topics 2(6):273–275

Oentaryo RJ, Murdopo A, Prasetyo PK, Lim EP (2016) On profiling bots in social media. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), p 10046 LNCS. https://doi.org/10.1007/978-3-319-47880-7_6

Orabi M, Mouheb D, al Aghbari Z, Kamel I (2020) Detection of bots in social media: a systematic review. Inf Process Manag. https://doi.org/10.1016/j.ipm.2020.102250

Pierri F, Artoni A, Ceri S (2020) Investigating Italian disinformation spreading on Twitter in the context of 2019 European elections. PLoS ONE. https://doi.org/10.1371/journal.pone.0227821

Ping H, Qin S (2019) A social bots detection model based on deep learning algorithm. In: Int Conf Commun Technol Proc, ICCT, 2019-October. https://doi.org/10.1109/ICCT.2018.8600029

Prabhu Kavin B, Karki S, Hemalatha S, Singh D, Vijayalakshmi R, Thangamani M, Haleem SLA, Jose D, Tirth V, Kshirsagar PR, Adigo AG (2022) Machine learning-based secure data acquisition for fake accounts detection in future mobile communication networks. Wirel Commun Mob Comput. https://doi.org/10.1155/2022/6356152

Pramitha FN, Hadiprakoso RB, Qomariasih N, Girinoto (2021) Twitter bot account detection using supervised machine learning. In: 2021 4th international seminar on research of information technology and intelligent systems, ISRITI 2021. https://doi.org/10.1109/ISRITI54043.2021.9702789

Pratama PG, Rakhmawati NA (2019) Social bot detection on 2019 Indonesia president candidate's supporter's tweets. Proc Comput Sci. https://doi.org/10.1016/j.procs.2019.11.187

Purba KR, Asirvatham D, Murugesan RK (2020) Classification of instagram fake users using supervised machine learning algorithms. Int J Electr Comput Eng. https://doi.org/10.11591/ijece.v10i3.pp2763-2772

Rahman MA, Zaman N, Asyhari AT, Sadat SMN, Pillai P, Arshah RA (2021) SPY-BOT: machine learning-enabled post filtering for social network-integrated industrial internet of things. Ad Hoc Netw. https://doi.org/10.1016/j.adhoc.2021.102588

Ramalingaiah A, Hussaini S, Chaudhari S (2021) Twitter bot detection using supervised machine learning. J Phys Conf Series 1950(1):012006. https://doi.org/10.1088/1742-6596/1950/1/012006

Rangel F, Rosso P (2019) Overview of the 7th author profiling task at Pan 2019: Bots and gender profiling in twitter. In: CEUR workshop proceedings, p 2380

Rao S, Verma AK, Bhatia T (2021) A review on social spam detection: challenges, open issues, and future directions. Exp Syst Appl. https://doi.org/10.1016/j.eswa.2021.115742

Rathore S, Loia V, Park JH (2018) SpamSpotter: an efficient spammer detection framework based on intelligent decision support system on Facebook. Appl Soft Comput J. https://doi.org/10.1016/j.asoc.2017.09.032

Reddy PM, Venkatesh K, Bhargav D, Sandhya M (2021) Spam detection and fake user identification methodologies in social networks using extreme machine learning. SSRN Electron J. https://doi.org/10.2139/ssrn.3920091

Ren H, Zhang Z, Xia C (2018) Online social spammer detection based on semi-supervised learning. ACM Int Conf Proc Series. https://doi.org/10.1145/3302425.3302429

Rodrigues AP, Fernandes R, Shetty A, Lakshmanna K, Shafi RM (2022) Real-time Twitter spam detection and sentiment analysis using machine learning and deep learning techniques. Comput Intell Neurosci 2022:1–14. https://doi.org/10.1155/2022/5211949

Rodríguez-Ruiz J, Mata-Sánchez JI, Monroy R, Loyola-González O, López-Cuevas A (2020) A one-class classification approach for bot detection on Twitter. Comput Secur. https://doi.org/10.1016/j.cose.2020.101715

Sadineni PK (2020) Machine learning classifiers for efficient spammers detection in Twitter OSN. SSRN Electron J. https://doi.org/10.2139/ssrn.3734170

Sahoo SR, Gupta BB (2020) Popularity-based detection of malicious content in facebook using machine learning approach. Adv Intell Syst Comput. https://doi.org/10.1007/978-981-15-0029-9_13

Santia GC, Mujib MI, Williams JR (2019) Detecting social bots on facebook in an information veracity context. In: Proceedings of the 13th international conference on web and social media, ICWSM 2019

Saranya Shree S, Subhiksha C, Subhashini R (2021) Prediction of fake Instagram profiles using machine learning. SSRN Electron J. https://doi.org/10.2139/ssrn.3802584

Sayyadiharikandeh M, Varol O, Yang KC, Flammini A, Menczer F (2020) Detection of novel social bots by ensembles of specialized classifiers. Int Conf Inf Knowl Manag Proc. https://doi.org/10.1145/3340531.3412698

Sedhai S, Sun A (2015) Hspam14: a collection of 14 million tweets for hashtag-oriented spam research. In: SIGIR 2015—proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. https://doi.org/10.1145/2766462.2767701

Sedhai S, Sun A (2018) Semi-supervised spam detection in Twitter stream. IEEE Trans Comput Soc Syst 5(1):169–175. https://doi.org/10.1109/tcss.2017.2773581

Sen I, Singh S, Aggarwal A, Kumaraguru P, Mian S, Datta A (2018) Worth its weight in likes: towards detecting fake likes on instagram. In: WebSci 2018—proceedings of the 10th ACM conference on web science. https://doi.org/10.1145/3201064.3201105

Sengar SS, Kumar S, Raina P (2020) Bot detection in social networks based on multilayered deep learning approach. Sens Transducers 244(5):37–43

Shao C, Ciampaglia GL, Varol O, Yang K, Flammini A, Menczer F (2017) The spread of low-credibility content by social bots. Nat Commun. https://doi.org/10.1038/s41467-018-06930-7

Shearer E, Mitchell A (2022) News use across social media platforms in 2020, Pew Research Center's Journalism Project. Available at: https://www.journalism.org/2021/01/12/news-use-across-social-media-platforms-in-2020. Accessed 9 Oct 2022

Sheeba JI, Pradeep Devaneyan S (2019) Detection of spambot using random forest algorithm. SSRN Electron J. https://doi.org/10.2139/ssrn.3462968

Sheehan BT (2018) Customer service chatbots: anthropomorphism adoption and word of mouth. Griffith University, University of Queensland, Queensland

Sheikhi S (2020) An efficient method for detection of fake accounts on the instagram platform. Revue Intell Artif. https://doi.org/10.18280/ria.340407

Shevtsov A, Tzagkarakis C, Antonakaki D, Ioannidis S (2022) Explainable machine learning pipeline for Twitter bot detection during the 2020 US Presidential Elections. Softw Impacts 13:100333. https://doi.org/10.1016/j.simpa.2022.100333

Shukla R, Sinha A, Chaudhary A (2022) TweezBot: an AI-driven online media bot identification algorithm for Twitter social networks. Electron (switzerland). https://doi.org/10.3390/electronics11050743

Shukla H, Jagtap N, Patil B (2021) Enhanced Twitter bot detection using ensemble machine learning. In: Proceedings of the 6th international conference on inventive computation technologies, ICICT 2021. https://doi.org/10.1109/ICICT50816.2021.9358734

Siddiqui A (2019) Facebook 2019 Q1 earnings: The social media giant boasts 2.7 billion monthly active users on its all services, Digital Information World. Available at: https://www.digitalinformationworld.com/2019/04/facebook-q1-2019-report.html. Accessed 9 Oct 2022

Singh Y, Banerjee S (2019) Fake (sybil) account detection using machine learning. SSRN Electron J. https://doi.org/10.2139/ssrn.3462933

Sohrabi MK, Karimi F (2018) A feature selection approach to detect spam in the Facebook social network. Arab J Sci Eng. https://doi.org/10.1007/s13369-017-2855-x

Subrahmanian VS, Azaria A, Durst S, Kagan V, Galstyan A, Lerman K, Zhu L, Ferrara E, Flammini A, Menczer F (2016) The DARPA

Twitter bot challenge. Computer 49(6):38–46. https://doi.org/10.1109/MC.2016.183

Tenba Group (2022) What is Sina Weibo? Know your Chinese social media!, Tenba Group. Available at: https://tenbagroup.com/what-is-sina-weibo-know-your-chinese-social-media. Accessed 9 Oct 2022

Thakur S, Breslin JG (2021) Rumour prevention in social networks with layer 2 blockchains. Soc Netw Anal Mining. https://doi.org/10.1007/s13278-021-00819-y

Thejas GS, Soni J, Chandna K, Iyengar SS, Sunitha NR, Prabakar N (2019) Learning-based model to fight against fake like clicks on Instagram posts. In: Conference proceedings—IEEE SOUTHEASTCON, 2019-April. https://doi.org/10.1109/SoutheastCon42311.2019.9020533

Thuraisingham B (2020) The role of artificial intelligence and cyber security for social media. In: Proceedings—2020 IEEE 34th international parallel and distributed processing symposium workshops, IPDPSW 2020. https://doi.org/10.1109/IPDPSW50202.2020.00184

van der Walt E, Eloff J (2018) Using machine learning to detect fake identities: bots vs humans. IEEE Access. https://doi.org/10.1109/ACCESS.2018.2796018

Varol O, Ferrara E, Davis CA, Menczer F, Flammini A (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the 11th international conference on web and social media, ICWSM 2017

Wald R, Khoshgoftaar TM, Napolitano A, Sumner C (2013) Predicting susceptibility to social bots on Twitter. In: Proceedings of the 2013 IEEE 14th international conference on information reuse and integration, IEEE IRI 2013. https://doi.org/10.1109/IRI.2013.6642447

Wanda P, Hiswati ME, Jie HJ (2020) DeepOSN: bringing deep learning as malicious detection scheme in online social network. IAES Int J Artif Intell. https://doi.org/10.11591/ijai.v9.i1.pp146-154

Wiederhold G, McCarthy J (1992) Arthur Samuel: Pioneer in machine learning. IBM J Res Dev 36(3):329–331. https://doi.org/10.1147/rd.363.0329

Wu B, Liu L, Yang Y, Zheng K, Wang X (2020) Using improved conditional generative adversarial networks to detect social bots on Twitter. IEEE Access. https://doi.org/10.1109/ACCESS.2020.2975630

Wu Y, Fang Y, Shang S, Jin J, Wei L, Wang H (2021) A novel framework for detecting social bots with deep neural networks and active learning. Knowl Based Syst. https://doi.org/10.1016/j.knosys.2020.106525

Xiao C, Freeman DM, Hwa T (2015). Detecting clusters of fake accounts in online social networks. In: AISec 2015—proceedings of the 8th ACM workshop on artificial intelligence and security, co-located with CCS 2015. https://doi.org/10.1145/2808769.2808779

Xu G, Zhou D, Liu J (2021) Social network spam detection based on ALBERT and combination of Bi-LSTM with self-attention. Secur Commun Netw. https://doi.org/10.1155/2021/5567991

Yang C, Harkreader R, Gu G (2013) Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Trans Inf Forensics Secur. https://doi.org/10.1109/TIFS.2013.2267732

Yang Z, Chen X, Wang H, Wang W, Miao Z, Jiang T (2022) A new joint approach with temporal and profile information for social bot detection. Secur Commun Netw 2022:1–14. https://doi.org/10.1155/2022/9119388

Yang C, Harkreader R, Zhang J, Shin S, Gu G (2012) Analyzing spammers'social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter. In: WWW'12—proceedings of the 21st annual conference on World Wide Web. https://doi.org/10.1145/2187836.2187847

Zeng Z, Li T, Sun S, Sun J, Yin J (2021) A novel semi-supervised self-training method based on resampling for Twitter fake account identification. Data Technol Appl 56(3):409–428. https://doi.org/10.1108/dta-07-2021-0196

Zhang W, Sun HM (2017) Instagram spam detection. In: Proceedings of IEEE Pacific Rim international symposium on dependable computing, PRDC. https://doi.org/10.1109/PRDC.2017.43

Zhang Z, Gupta BB (2018) Social media security and trustworthiness: overview and new direction. Future Gener Comput Syst. https://doi.org/10.1016/j.future.2016.10.007

Zheng X, Zhang X, Yu Y, Kechadi T, Rong C (2016b) ELM-based spammer detection in social networks. J Supercomput 72(8):2991–3005. https://doi.org/10.1007/s11227-015-1437-5

Zheng X, Wang J, Jie F, Li L (2016a) Two phase based spammer detection in Weibo. In: Proceedings—15th IEEE international conference on data mining workshop, ICDMW 2015. https://doi.org/10.1109/ICDMW.2015.22