1-1-2023

Social media bot detection with deep learning methods: a systematic review

Kadhim Hayawi
Zayed University

Susmita Saha
Mass Dynamics

Mohammad Mehedy Masud
United Arab Emirates University

Sujith Samuel Mathew
Zayed University

Mohammed Kaosar
Murdoch University

Neural Computing and Applications
https://doi.org/10.1007/s00521-023-08352-z

REVIEW

Social media bot detection with deep learning methods: a systematic review

Kadhim Hayawi[1] · Susmita Saha[2] · Mohammad Mehedy Masud[3] · Sujith Samuel Mathew[1] · Mohammed Kaosar[4]

Received: 28 December 2021 / Accepted: 13 February 2023

Abstract

Social bots are automated social media accounts governed by software and controlled by humans at the backend. Some bots have good purposes, such as automatically posting information about news and even to provide help during emergencies. Nevertheless, bots have also been used for malicious purposes, such as for posting fake news or rumour spreading or manipulating political campaigns. There are existing mechanisms that allow for detection and removal of malicious bots automatically. However, the bot landscape changes as the bot creators use more sophisticated methods to avoid being detected. Therefore, new mechanisms for discerning between legitimate and bot accounts are much needed. Over the past few years, a few review studies contributed to the social media bot detection research by presenting a comprehensive survey on various detection methods including cutting-edge solutions like machine learning (ML)/deep learning (DL) techniques. This paper, to the best of our knowledge, is the first one to only highlight the DL techniques and compare the motivation/effectiveness of these techniques among themselves and over other methods, especially the traditional ML ones. We present here a refined taxonomy of the features used in DL studies and details about the associated pre-processing strategies required to make suitable training data for a DL model. We summarize the gaps addressed by the review papers that mentioned about DL/ML studies to provide future directions in this field. Overall, DL techniques turn out to be computation and time efficient techniques for social bot detection with better or compatible performance as traditional ML techniques.

Keywords Social media · Bots · Deep learning · Machine learning · Bot detection · Systematic review

Abbreviations

DL	Deep learning
ML	Machine learning
SVM	Support vector machine
SMOTE	Synthetic minority oversampling
ENN	Edited Nearest Neighbours
CGAN	Conditional generative adversarial network
ADASYN	Adaptive synthetic
GloVe	Global Vectors for Word Representation
USE	Universal Sentence Encoder
LIWC	Linguistic Inquiry and Word Count
MTF	Markov transition field
GAF	Gramian Angular Field
SST2	Stanford Sentiment Treebank
BERT	Bidirectional Encoder Representations from Transformers
GPT	Generative pre-trained transformer
LSTM	Long short-term memory
CNN	Convolutional neural network
RNN	Recurrent neural network
GRU	Gated recurrent unit
BiGRU	Bidirectional gated recurrent unit
BeDM	Behaviour enhanced deep model
Bi-SN-LSTM	Bidirectional self-normalizing LSTM network
ResNet	Residual network
MLP	Multilayer Perceptron

✉ Kadhim Hayawi
abdul.hayawi@zu.ac.ae

[1] College of Interdisciplinary Studies, Computational Systems, Zayed University, Abu Dhabi, UAE

[2] Mass Dynamics, Melbourne, Australia

[3] College of Information Technology, United Arab Emirates University, Al Ain, UAE

[4] Discipline of IT, Media and Communications, Murdoch University, Perth, Australia

Published online: 06 March 2023

GCNN	Graph convolutional neural network
PSO	Particle swarm optimization
FFNN	Feed-forward neural network
RDNN	Regularized deep neural network
C-DRL	Content-based deep reinforcement learning
SNA-DRL	Social network analysis-based deep reinforcement learning
HAN	Hierarchical attention networks
RF	Random forest
LR	Logistic regression
SL	Simple logistic
DNN	Deep neural networks

1 Introduction

Nowadays, social media platforms are expanding at a fast pace in terms of number of users, data size and applications. They are basically internet-based applications that facilitate exchange of user-generated contents. A high rate of usage of such platforms creates a revolution in human communication. For instance, Facebook had almost one-third of the world population[1] as its users in the first quarter of 2019, and by 2015, the estimated number of users had grown to 1.3 billion in Twitter [1]. In general, people use social media platforms for interacting with others via various type of posts, by following them and being followed. In some platforms like Twitter, trending topics are discussed on a daily basis [2]. Malicious individuals and organizations exploit the flexibility and power of social media to gain influence by creating fake automated accounts, often called social bots or sybil accounts and, in this study, social media bots. These accounts can exploit the regular services for malicious purposes by manipulating the discussion and public opinion, spreading rumours and fake news, promoting harmful products/services, defaming other people or being fake followers of a user to handcraft a fake popularity and spamming/social phishing/profile cloning/collusion attacks [3, 4]. These attacks can be catastrophic. Some of the vicious examples of bot infiltration are attacks during US presidential election, Russiagate hoax[2] attack and rumour spread during Boston Marathon blasts [5] in Twitter. A very high percentage of bot accounts in social media, such as between 9 and 15% accounts (equivalent to 48 million accounts) in Twitter according to a recent study [6], makes these platforms highly vulnerable.

Hence, social bot detection research got a huge interest as a defence mechanism against such threats. Moreover, the human-like characteristic of these bots and a continuous evolution [7, 8] in their strategy make it very difficult to distinguish them from legitimate users and thus invoke the need of more sophisticated and dynamic countermeasures.

Machine learning can be a smart approach for learning the pattern of bot behaviour and thus for efficient detection, especially by making an effective use of the mammoth streaming data generated from these online platforms. Several machine learning techniques, including supervised, unsupervised and reinforcement learning, have been proposed to detect bots in Twitter and other platforms [9–11]. In general, ML models show good performance measures for social bot detection with easy implementation; however, they exhibit time and computation expensive feature extraction, slower learning time and less performance in case of large features. Deep learning is a special branch of machine learning which is distinguishable from traditional ML approaches by its layered architecture and ability to process and extract features from complex data such as images, text and speech. Many DL models are shown to outperform the traditional, shallow and ML classifiers for the bot detection task. In addition, to overcome the challenge of cyborgs with human-like behavioural attributes, DL techniques especially generative adversarial networks [12] can be really effective.

To the best of our knowledge, six review articles have been published so far in this field highlighting social bot detection techniques and taxonomy (see details in supplementary Table S1). Hence, the existing surveys cover the broader field of social bot detection, highlighting all of the technical approaches such as structure-based, crowdsourced, hybrid, graph-based, machine learning and other techniques such as dynamic time warping, digital DNA-based or natural language processing (NLP) approaches [8, 13–17]. We have found at least one review article that surveyed machine learning algorithms in general for different categories of social media analysis including bot detection [14].

The motivation behind the current systematic review on only the deep learning articles for social media bot detection is mainly based on two reasons. First, the DL approaches showed high potential to detect the benign/malicious bots and to keep pace with their fast-evolving and highly variable characteristics, which is a pressing need right now. Second, none of the previous systematic reviews attempted to synthesize the efficacy, failure and challenges of exclusively these techniques for social media bot detection. It is important to figure out the status of the deep learning research in comparison with other techniques, including the classical machine learning

[1] https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

[2] https://www.nbcnews.com/tech/tech-news/after-mueller-report-twitter-bots-pushed-russiagate-hoax-narrative-n997441.

algorithms, to guide the future research in social media bot detection to the right direction and to maximize the usability of current data sources. Third, while a previous review article presented machine learning research in general for social media, narrowing it down to social media bot detection and going deeper from both technical and application points of view is crucial for ensuring future success in this field.

Here, we presented a comprehensive analysis on the model architecture, processing workflows, effectiveness and limitations of these cutting-edge approaches. Figure 1 shows a generalized overview of the end-to-end process for the DL-based social bot detection task.

The main contribution of our work can be summarized as follows:

- We provide a systematic review of various DL approaches which have been used particularly in social media bot detection, including all the related literature found between 2000 and 2021, from pre-defined resources, and based on pre-defined inclusion/exclusion criteria. Out of 1496 obtained publications using the search term, 40 are finally selected for this review. Up to the best of our knowledge, it is the first systematic review conducted on this topic.
- We present a comparative study among DL algorithms and between DL and traditional ML approaches for the bot detection task. It is worth mentioning here that only the comparisons with ML, as reported in the DL articles, were presented and analysed in this review.
- We provide the summary of the datasets, taxonomy of the features and their extraction mechanism that are extensively used for DL approaches in this field.
- We provide the future research direction by informing the researchers of the high potential of these approaches as well as of the current gaps and challenges.

These contributions can enrich DL-based research for the bot detection, especially by directing researchers to the most effective algorithms and their associated features and pre-processing strategies and to the gaps and future potentials of this research field.

This paper is structured as follows: Sect. 2 defines the social bots; Sect. 3 explains the methodologies of this paper; Sect. 4 explains the datasets, features and other pre-processing strategies used for DL methods; Sect. 5 analyses and compares the architecture and performance of DL models for bot detection; Sect. 6 presents a comparative study between DL and traditional ML approaches; Sect. 7 discusses the research gaps and challenges regarding the DL-based social bot detection and, finally, Sect. 8 concludes the paper with directions to future work.

2 Social media bots

Before going into detail studies on 'Social media bots', it is very important to define what they are. The word 'bot' generally means automated software agents that are designed to hold human conversation or to use in command-and-control networks to launch attacks [18, 19]. Still, the concept of bots in social media can be really intricate. By following the very recent systematic review by [17], we will use the term 'social media bots' to clearly differentiate them from more general categories such as social bots or sybils that are defined with the condition of mimicking human behaviour [20].

Now, automated accounts or bots can be benign or malicious. Also, the automated malicious accounts may not mimic human behaviour, known as traditional spam bots. Some authors proposed to define them as semi-automatic agents [21] as they are designed to fulfil a specific purpose and generally controlled by a botmaster or human who manages their activities. Some categorized malicious bots such as spam bots (disseminating spam), political bots (get involved in political discussion) and sybils (fake accounts) use to gain undeserved influence [17]. In our review, we emphasize on the word 'bot' or 'sybil' that represents automated accounts only. Hence, we excluded all those papers which mentioned fake profile or malicious account detection in mixed terms (human + automated) and included those reporting the automated accounts only. In some articles, although it is not clear which kind of bots

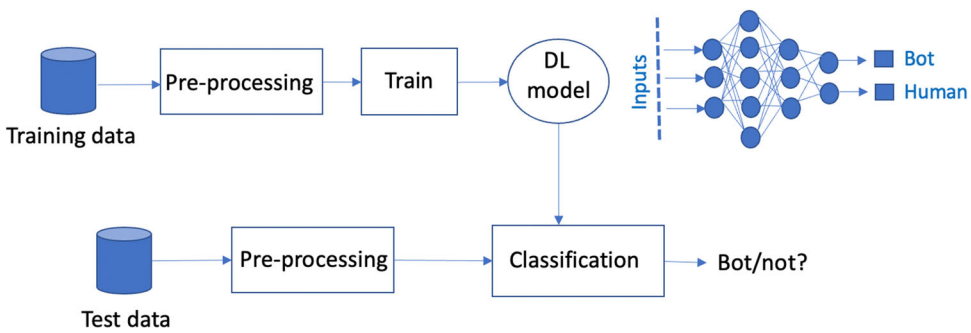

Fig. 1 A generalized workflow diagram of the DL-based social bot detection. The collected data are pre-processed first through labelling, data augmentation and feature extraction and then used to train a DL model. The trained model is used for the classification of human/normal users or social media bots on the unseen test data

they attempted to detect (benign or malicious), we did not exclude any paper based on that criteria. The papers used different terms such as normal or human users or legitimate accounts to define the opposite class of bots; however, we did not set any filtering criteria for that category either. Our search strategies and inclusion/exclusion criteria are distinctly defined in the next section.

3 Methodology

3.1 Research questions

For this review, the authors followed the guidelines suggested by [22] for conducting a systematic literature review (SLR). After identifying the need for such a review as depicted in 'Introduction', we specified the research questions. This work has been conducted precisely to report about the deep learning research in the field of social media bot detection. Hence, the study aimed to answer the following questions:

RQ1 What are the deep learning algorithms that are used for social bot detection or prediction?
RQ2 What are the pre-processing mechanisms needed for the above algorithms?
RQ3 What is the effectiveness (success or failure) of different DL models in association with combinatory feature inputs (multimodal inputs, e.g. user data, posts, etc.) for social bot detection?
RQ4 How is the performance of deep learning algorithms in comparison with traditional machine learning models?
RQ5 What are the gaps and future research directions in this area?

3.2 Search strategy

The search strategy was determined based on its objectives and research questions. It was derived through the consultation among the authors with relevant experience and through trial searches using various combinations of search terms to find the already known primary studies. The search is conducted in several stages: primary term-based search in three of the largest databases including Google Scholar/Scopus/ScienceDirect, crawling in published literature/systematic reviews for any missing papers (journal articles and conference proceedings), manual screening over the title and abstracts and finally applying exclusion and inclusion criteria to generate the targeted set of publications for review. We also informally searched over other databases such as IEEE/ACM to find any missing ones and did not find anything additional. The abstract screening and the outcome of search using scoping criteria were assessed based on the agreement among the researchers on each paper.

We used Rayyan Intelligent Systematic Review software[3] for an organized screening of the papers with the above-mentioned steps.

Figure 2 shows our selection steps from the initial search to finalizing the articles to be analysed and reported in this study.

3.3 Search terms

We generated a very wide variety of search terms and searched over three significant databases to include every published article that used DL for social media bot detection. Therefore, we included different bot terms ('bot', 'fake news', 'botnet' and 'sybil') and also both 'detection' and 'prediction' terms. However, the terms 'detection' and 'prediction' represent the same task, i.e. to classify bots vs. human. The search term was finalized through initial trials and manual inspection on the outcome.

Google Scholar:With all of the words: social bot detection ['in title' box]and articles dated between 2000 and until now (the time range was set considering social media bot detection with DL approaches as a relatively new field of research)Time stamp: 31 March 2021, 9:27 pm: count: 46

Scopus:

TITLE-ABS ('Review' AND 'social bot detection') OR TITLE-ABS (('detection' OR 'detecting' OR 'prediction' OR 'predicting') AND ('Bot' OR 'bots' OR 'fake news' OR 'botnet community' OR 'botnet' OR 'Social network polluting contents' OR 'sybil') AND ('deep' OR 'machine learning' OR 'neural network')) AND PUBYEAR > 2000Time stamp: 31 March 2021, 12:15 am: count: 1232

ScienceDirect:

In advanced search: 'Find articles with these terms: social bot detection, deep learning, machine learning and neural network'.

Year: 2000–2021Time stamp: 31 March 2021, 12:09 am: count: 311

4 Results

4.1 Overview of input features and pre-processing

This section presents the detailed information on the taxonomy of the features and pre-processing strategies including data labelling, balancing and feature extraction for the reviewed studies.

[3] https://www.rayyan.ai/.

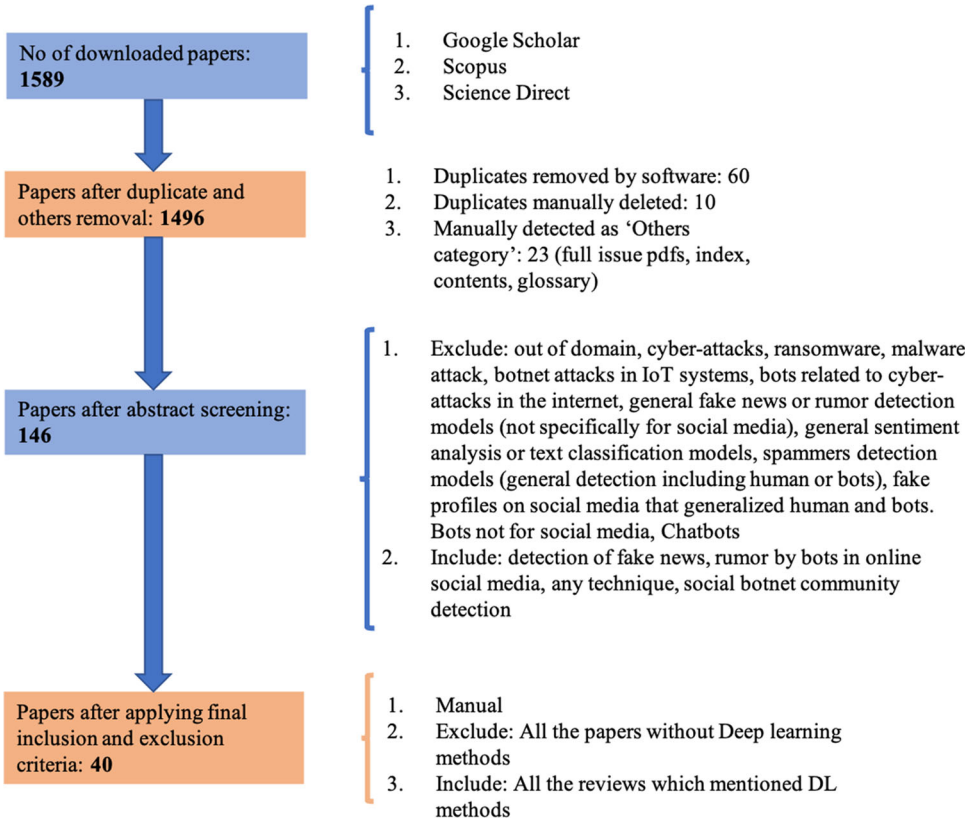

Fig. 2 Systematic steps for the selection of the articles for our final review

4.1.1 Types of features

Based on our detail analysis, the features used for DL models can be classified into two main groups as shown in Fig. 3.—user metadata and tweets/posts. These two main features with their subclasses are described below.**User metadata:** This group generally represents information that come from the user profile information, such as nick name, introduction, location, follower count, friend count, listed count, favourites count and statuses count, as used by several studies [23, 24].

Tweet/posts data: The following features are shown to be extracted from users' posts for classification.

- *Temporal/behavioural features* This group represents behaviour information including posting behaviour/connecting behaviour. Studies like [25] regarded user's history tweets as temporal text data instead of plain text in existing methods, by using two features, timestamps (posting inter-arrival time pattern) and posting type (original/retweet). In addition, they focussed on extracting endogenous (circadian rhythm) and exogenous features (cultural or environment influences) that represent social behaviour of a user. Posting times have been used by other studies too [26]. From the time interval between posts, another study [27] also derived burstiness parameters or information entropy to use them as features. Following a different approach, Lingam and colleagues represented the descriptive features of user's behavioural patterns as states, as inputs to the learning agent for the deep Q-learning model [28].
- *Textual/content-based features* These features can be directly extracted without any intermediary steps such as num_hashtags, num_urls, num_mentions, num_punctuations and num_interjections. In addition, the high-level representation of the textual data such as abuses, angry terms and named entities from the tweets can be derived by ConvNets [25, 29]. Al-Qurishi et al. [26] used quality of the generated content as features and derived emotional reactions from twitter posts to predict bots.
- *Interaction-based features* The previous studies [27] used interaction-based features such as number of comments/reposts/likes in the posts, repost ratio and diversity of sources of posts for the classification. Al-Qurishi et al. [26] mentioned the interactions as graph-based features. [30] also proposed to extract the bipartite graph of the following relationship and the retweet relationship between users to address the challenge of differentiating between anomaly and legitimate accounts.

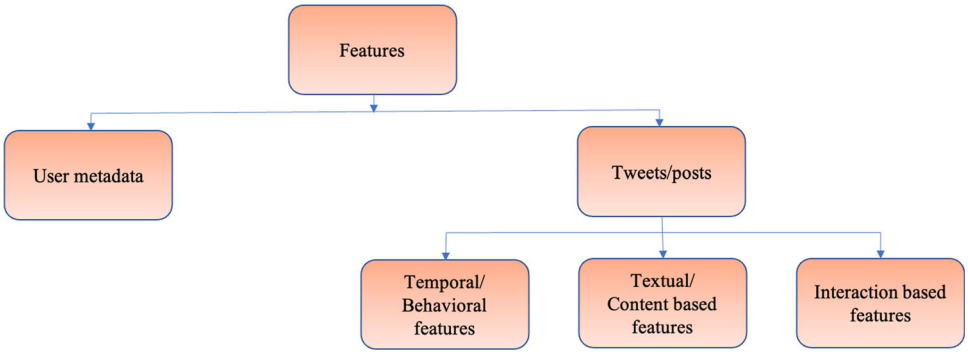

Fig. 3 Taxonomy of features used in the DL research for bot detection

4.1.2 Data collection, labelling and augmentation

Most of the reviewed studies used existing or popular datasets for their model training and testing (see supplementary; Table S2). For example, Katarya, Mehta [29] used the popular MIB datasets that included manually annotated genuine and fake twitter accounts along with tweet data corresponding to these accounts. Cai, Li [25] used a public dataset [31] which was collected with honeypot method. Heidari and Jones [32] used the advantage of BERT's transfer learning approach and used labelled data of movie reviews to detect bots in social media, and thus proposed a new way to use labelled data from other online platforms to detect bots in Twitter.

Wu et al. [27] manually labelled a small portion of user data, collected from Sina Weiboo, based on their proposed six discrimination metrics. They introduced an active learning module composed of a query algorithm, a machine learning classifier and a supervisor, for building a large-scale experimental dataset. Kudugunta and Ferrara [33] proposed an effective technique to generate a large, labelled dataset from a minimal amount of labelled data based on a combination of synthetic minority oversampling (SMOTE) and two under sampling techniques (Edited Nearest Neighbours (ENN) and Tomek Links). Wu et al. [34] proposed another data-augmentation approach by adopting improved CGAN generative methods to extend social media bot samples with three types of oversampling methods: SMOTE, the adaptive synthetic (ADASYN) algorithm and the random oversampling methods.

4.1.3 Feature extraction

Some of the key feature extraction or use strategies are described as follows:

A couple of studies [23, 29] used a number of data processing steps to improve model efficiency, namely null value cleaning and reducing dimensionality, and thus ended up with an appropriate subset of the features for training. In contrary, Wu et al. [27] used the combination of metadata-, interaction-, content- and timing-based features to distinguish between social media bots and normal users and showed that all detection approaches performed best on the feature set that contained all the features.

Ping and Qin [35] used temporal features, content features and relationship between them as inputs to multiple DL models. Cai, Li [25] concatenated text vector, timestamp vector and posting type vector into a single sequence vector as the input to the bot detection classifier. Halvani and Marquardt [36] also performed some low-level preprocessing steps on their dataset including concatenation of all tweets in each XML file into a one long document, lowercasing of the entire text and substitution of noisy elements with a dummy token, for example twitter handles.

Most of the studies used appropriate language modelling technique to convert the textual content to suitable inputs for classification. A number of studies [25, 29, 32, 37, 38] used GloVe word embedding [39] for language modelling. Kudugunta and Ferrara [33] used an additional step by forming a string of tokens from each tweet by a python library that is called ekphrasis, which performed tokenization, word normalization, word segmentation (for splitting hashtags) and spell correction. Then, the tokenized tweets were transformed into an embedding using the GloVE model. Finally, the mean of the embeddings for each tweet had been calculated, and thus, each tweet was represented as a vector. Onose, Nedelcu [40] used pre-trained word2vec word embeddings for text representation. Mou and Lee [41] proposed a novel framework JntL that incorporated both handcrafted features and automatically learnt features by DL methods from various perspectives. They extracted Linguistic Inquiry and Word Count (LIWC), a dictionary that captured the general statistics of each user's profile, activities and their preferences in terms of word usage. Automatic feature learning was divided into two components: text embeddings and temporal behaviour embeddings. They used Universal Sentence Encoder (USE) for sentence-level embedding and BERT-Base for word-level embeddings. Given the sequences of each user's posting timestamp, they applied two methods for mapping those sequences into 3D images: GAFMTF and II Map for

informative pattern recognition. (Heidari & Jones, 2020) used Google Bert for sentiment classification of tweets. BERT and RoBERTa (an adaptive version of BERT) embeddings were also used for bot detection by [42, 43]. While BERT facilitates multilingual embeddings, RoBERTa improves the performance in longer sequences, or when there are vast volumes of data.

4.2 Deep learning methods for social media bot detection: evaluation

Table 1 summarizes all the DL algorithms that the social media bot detection studies have used until now. From the list, we have explained here some groups of high performing or novel DL models and scrutinized them according to their pre-processing strategies, training/test datasets and prediction performance.

The holistic analysis will validate the information that the cutting-edge artificial intelligence techniques are generally working well in this field, and in addition to traditional ML algorithms, the future researchers might want to consider deep learning models for their study. This will help the researchers to find out the best suitable approach in terms of the model architecture, feature extraction, preprocessing mechanism and sample size for their particular dataset and objective of the study. If any model performance is worse or the approach has limitations, we have identified the reason and, in contrast, emphasized on their unique benefits. The key approach here is the same, i.e. the use of deep learning to deal with the highly heterogenous and evolving social media bots.

Here, we present a taxonomy for the most commonly used DL models for the bot detection task in Fig. 4.

Some of the highly performing approaches are as follows:

CNN/LSTM/Bi-LSTM/RNN/GRU: max accuracy 100%

A number of social media bot detection studies used different combinations of deep convolutional neural network (CNN), LSTM, recurrent neural network (RNN) and gated recurrent units (GRU). Ping and Qin [35] achieved the highest performance with 100% accuracy among all. The main characteristic of their proposed model was the use of tweet content and temporal features in a fused way. In general, CNNs work better to classify textual content and identify abuses, angry terms, named entities, etc.; however, representing the textual data in a vectorized form through language models such as Word2Vec was essential. In contrast, LSTMs and RNNS are both highly useful for finding the relationship between the consecutive time point data, where sequence of sentence is important. Their proposed DeBD model included the following layers:

- Joint content feature extraction layer for generating features of tweet contents and the relationship between them through CNNs from the concatenated tweets posted by social bots.
- Temporal feature extraction layer for the extraction of temporal features of tweet metadata using LSTM.
- Fusion layer to fuse the above-mentioned features with certain rules.
- Finally, a fully connected layer and a softmax layer to generate the binary labels: 'social bot or not'.

Although they achieved a very high precision and nearly perfect accuracy on different datasets, it should be noted that it required a large amount of tweet information from the user and removal of the non-English users and their tweets from the dataset.

In comparison, the behaviour enhanced deep model (BeDM) model by [25] fused the tweet content information and behavioural information using deep learning method. The order of a specific tweet was determined by its posting time, and a high-level representation of each tweet was obtained using the shared CNNs. The posting behaviour information in terms of timestamp and posting types was concatenated with the text vector output of CNN. Finally, the sequence of concatenated vectors for all history tweets was fed into an LSTM network to finally label each input as bot or human. Although the accuracy measure was not available, it turned out that this strategy was not as successful as other studies with only 88.41% precision for detecting bots, even with a very large sample size of around 2 million in each of the bot and human class.

Sengar, Kumar [23] used an interesting model architecture with a machine learning classifier for the final prediction labels, from the multimodal feature inputs such as first level predictions by CNN from GloVe word embeddings fused with user account and tweet metadata. Their approach on an extremely large dataset with more than eight thousand user accounts and around 7 million tweets, achieved as good as or better performance when compared to only LSTM/Bi-LSTM model-based studies [33, 39] and the combinatory CNN-LSTM studies as mentioned above. The study by Mou and Lee [41] attained 91.8% precision and 90.1% accuracy with around 37,000 accounts in each class, with 84 handcrafted and automatically learnt features such as text embeddings and temporal behaviour embeddings and using CNN for final decision-making. In a less complicated way, another study [44] showed that bot accounts could be detected with a very high accuracy with a combination of CNN and ANN using a single post on the social media; however, it requires a huge amount of tweets (about 1 million) in each class. Gao et al. [45] used a self-normalizing CNN, adopted to extract lower features from the multi-dimensional input data, that

Table 1 Comparison among DL algorithms: architecture and performance and comparison with traditional ML algorithms [model performance measures are shown as the (min–max) range of values reported by combining the results from the cited studies in each row. Please note that some measures have single values as only one study reported that metric, and a range is not available. A range of values for a single study may represent measures for different test datasets or different combinations of features/models.]

Deep learning methods	Social media	Model performance	Comparison with other ML/DL methods	Dataset	Input features	Used in/mentioned in
CNN + Multilayer Perceptron/ other ML models	Twitter	Accuracy: 90.31–99.54% Precision: 90.76–99.54% Recall: 90.31–99.54% F1-score: 96.73–99.54%	When CNN predictions added with other features for a final classification, performance improved than KNN, support vector, decision tree, RF, AdaBoost, gradient boosting, Gaussian NB, Linear Discriminant and Multilayer Perceptron	Cresci	Profile details, tweets and tweet metadata	[23, 29]
CNN + ANN	Twitter	Accuracy: 93.69–99.43% Precision: 94–99% Recall: 94–99% F-measure: 94–99%	Performance improved over only CNN (tweet features based) and only ANN (profile features based) models CNNs can learn the difference between the text contents of tweets by bots and humans better than ANN	Cresci	Profile details, tweets and tweet metadata	[44]
CNN + LSTM/only Bi-LSTM/CNN + self-normalizing Bi-LSTM + dense layers	Twitter	Accuracy: 96.1–100% Precision: 87.58–99.99% Recall: 86.26–100% Specificity: 93.8–99.8% F1-score: 87.32–99.99% AUC: 99.32% MCC (Mathews correlation coefficient): 0.92–0.998	(CNN + LSTM) performance improved over Stweeler [54], Boosting [55], BoostOR [31] Performance of bi-SN-LSTM improved over commonly utilized RNN methods including vanilla RNN, bidirectional RNN (bi-RNN), gated recurrent unit (GRU), bidirectional GRU (bi-GRU), LSTM and bidirectional LSTM (bi-LSTM) Only Bi-LSTM application on Cresci'17 dataset [56] was either compatible or a little lower in performance in comparison with unsupervised DNA-inspired online behavioural model [57] CNN + LSTM models performed better than Twitter, human annotators, BotOrNot (supervised RF Classifier), decision tree, BayesNet and Decorate classifiers [7] unsupervised stream clustering algorithms (StreamKM + + and DenStream)[58], generic statistical approaches in [59] Unsupervised DNA-inspired online behavioural model proposed in Cresci'16 performed a little better in terms of precision and specificity in one of the test sets but all other measures	Social Honeypot, Cresci, Lee, Gilani, Varol, Midterm, Botometer, Botwiki, Celebrity, Pronbots, Vendor and Verified	Tweet, tweet metadata, timestamps, user profiles, structure of users' local subgraph, LIWC features and word/sentence/ temporal embeddings	[25, 35, 38, 41, 45]

Table 1 (continued)

Deep learning methods	Social media	Model performance	Comparison with other ML/DL methods	Dataset	Input features	Used in/mentioned in
			were lower than the DeBD model (CNN + LSTM) as in [35]			
CNN	Twitter	Accuracy: 85.65% Precision: 97.02% Recall: 94.35% F1-score: 83.67% AUC: 85.65%	Simpler models with less trainable weights work better than complicated ones	Pan19	Tweets	[54]
Bidirectional Encoder Representations from Transformers (Google Bert) + fast forward neural network	Twitter	Accuracy: 94.8–97.9% F1-score: 94.2–97.6%. MCC (Mathews correlation coefficient): 89.1–96.2%	Accuracy and overall performance improved over BotOrNot [61], RF/ decision tree/BayesNet/Decorate [7], StreamKM + + (based on K-means) and DenStream (based on DBSCAN) [58], Naïve Bayes, Jrip, decision tree [59] and unsupervised sequence clustering [57]	Cresci	BERT embeddings	[32]
A novel deep neural network model called RGA = a residual network (ResNet) + a bidirectional gated recurrent unit (BiGRU) + attention mechanism	Twitter	Accuracy: 98.87%. Precision: 98.40% Recall: 99.33% F1-score: 98.86%	Performance improved over LR/SVM/ ELM/RF/MLP/LSTM/ComNN/CNN and performance improved over only Resnet, only BiGRU, Resnet + BiGRU, Resnet + attention, BiGRU + attention models	Data collected from Sina Weibo	Metadata-, interaction-, content- and timing-based features	[27]
Contextual LSTM/Bi-LSTM/ LSTM	Twitter	For tweet-level bot prediction: Accuracy: 79.64–98.88% Precision: 96% Recall: 96% F1-score: 96% ROC/AUC: 87.04%–96.43%	Contextual LSTM performance improved over only metadata-based LR/SGD/RF/ AdaBoost classifiers Bi-LSTM performance improved over MLP, AdaBoost, LSTM, RNN, Bi-GRU and GRU classifiers	Cresci and pan19	User profile features, tweet metadata and temporal features	[33, 37, 46]
Embedding layer + two Bi-LSTM layers + attention mechanism at the top of the second Bi-LSTM layer + three dense layers	Twitter	Accuracy: 82.61–99.06% Precision: 83.65% Recall: 89.41% F1-score: 86.43% ROC/AUC: 88.70%	Performance improved over MLP/CNN/ SVM/RF/LR/BNB/KNN/decision tree on the same dataset	Social Honeypot and Cresci	User profile features, tweet metadata and word embeddings from tweets	[47]

Table 1 (continued)

Deep learning methods	Social media	Model performance	Comparison with other ML/DL methods	Dataset	Input features	Used in/mentioned in
Feed-forward neural network/fully connected based neural network/MLP	Twitter, VKontakte	Precision: 92.59% Recall: 50% F1-score: 64.94% Accuracy: 86%–92% AUC: 73% (in VKontakte)	Accuracy is better for English tweet inputs in comparison with Spanish tweets Bot-DenseNet [43] was either compatible or a little lower in performance when compared to previous supervised (RF, one-class classifier, AdaBoost, logistic regression and decision tree classifiers) and unsupervised algorithms (LSTM autoencoder for feature extraction + hierarchical density-based algorithm)	Verified, Botwiki, Midterm, Cresci, Botometer, Vendor, Celebrity, Pronbots, Gilani, Varol, pan19 and collected datasets from Twitter, VKontakte	Profile-based, content-based features, tweet metadata, BERT contextualized text embeddings and emoji embeddings	[24, 26, 42, 43, 55, 56]
RNN	Twitter	88%	Performance improved over GRU/MLP/AdaBoost classifier	Cresci	Temporal features from tweets	[46]
Gated recurrent unit (GRU)/bidirectional GRU	Twitter	Accuracy: 93.1%	Performance improved over MLP/AdaBoost Classifier; Bi-GRU improved over RNN	Cresci	Temporal features from tweets	[46]
Deep Q-learning based on particle swarm optimization (DQL-PSO)	Twitter	Precision: $\sim$ 95% Recall: $>$ 90% F-measure: $>$ 90% G-measure: $>$ 90%	Performance improved over adaptive deep Q-learning, FFNN, RDNN, C-DRL and SNA-DRL	Social Honeypot and the Fake project	User behavioural patterns as states	[28, 57]
Improved conditional generative adversarial network (improved CGAN)	Twitter	Accuracy: 97% Precision: 97% Recall: 97% F1-score: 97% G-mean: 93%	Performance improved over the traditional CGAN algorithm	Cresci and Varol	User metadata, sentiment, friends, content, network and timing	[34]
Deep forest algorithm	Twitter	Accuracy with SMOTE: 97.55% Accuracy without SMOTE: 94.26%	Performance improved over RF algorithms	Varol and Social Honeypot	Metadata of user profile and posts	[58]
Hierarchical attention networks (HAN)	Twitter	Accuracy: 89.43%	Accuracy is better for English tweet inputs in comparison with Spanish tweets	Pan19	Tweets	[40]

were generated from user profiles and the structure of users' local subgraph and a bidirectional self-normalizing LSTM network (bi-SN-LSTM) to extract higher features from the compressed feature map sequence. Their unique technical contribution lied in proposing bi-SN-LSTM network with SELU as the activation function of its recurrent step, which provided unbounded changes to the state value and using alpha dropout to prevent overfitting while training neural networks. With 1950 sybils and 1950 human users' data, the model achieved 99.99% precision.

Thus, overall, it implies that the efficient fusion of textual, temporal and user account features with ensembled modelling is the key to the higher success for the bot detection.

By using GloVe word embeddings of tweets and Stanford CoreNLP toolkit for sentence splitting and tokenization to a bidirectional LSTM model, Wei and Nguyen [38] achieved lower performance (up to 96.1% accuracy and 94% precision) than the highest CNN + LSTM measures. However, their model required no prior knowledge or assumption about users' profiles, friendship networks or historical behaviour on the target account and thus no handcrafted features. On the contrary, using GloVe embeddings and a Bi-LSTM network, Luo et al. [37] achieved only 79.64% accuracy which could be rendered to their small sample size with 412,000 annotated tweets posted by 2060 bots and 2060 humans, respectively. Rajendran et al. [46] were able to distinguish the tweeting rate and frequency of bot accounts from the genuine accounts with a good classification accuracy of 98.88% using a Bi-LSTM model only. Although this is also lower than the maximum CNN + LSTM limit, they showed an important comparison among Bi-LSTM, Multilayer Perceptron (MLP), RNN, LSTM, GRU and bidirectional GRU models which revealed superiority of Bi-LSTM models with the highest accuracy among all. These results implied that the failure of RNN models to collect information from a long sequence can be overcome by replacing them with LSTM and GRU units.

Interestingly, the study by Kudugunta and Ferrara [33] proposed a contextual long short-term memory (LSTM) architecture that exploited both tweet content and user metadata to detect bots from a single tweet. Using a minimal amount of labelled data (roughly 3000 examples of sophisticated Twitter bots), they demonstrated a high classification accuracy (AUC $>$ 96%), which could limit the model generalizability and usability. However, the study also supported the idea that utilizing tweet metadata in addition to tweet content can improve the model performance. Another work by Ilias and Roussaki [47] also got successful using a variant of LSTM model (with embedding layer, two Bi-LSTM layers and attention mechanism at the top of the second Bi-LSTM layer and finally three dense layers). This model outperformed earlier works with 99.06% accuracy on Cresci'17 dataset (legitimate tweets vs. malicious ones), however, failed on Social Honeypot dataset with 82.615% accuracy (humans and content polluters). Future studies may investigate more hybrids of LSTMs like sentence-state LSTMs in association with other architecture as described above.

Multilayer perceptron/feed-forward neural network: max accuracy: 99.92%

A novel but much simpler Bot-DenseNet model has been proposed by Martín-Gutiérrez, Hernández-Peñaloza [43] to address the bot identification task by a multilingual approach using BERT and RoBERTa to generate tweet embeddings. They are later on concatenated with the rest of the metadata to build a potential input vector on top of a dense network. The model achieved a lower F1-score than the previous LSTM autoencoder performance [48]. However, certainly, the Bot-DenseNet model had a unique capability for analysing text features from multiple languages and due to its low-dimensional representation, was suitable for use in any application within the information retrieval (IR) framework. In contrast, Al-Qurishi, Alrubaian [26] achieved a high accuracy of 99.92% on cleaned (outliers removed) by using a data harvesting module, a (profile-based, graph-based and content-based) feature extracting mechanism and a deep feed-forward neural network to detect sybil vs. honest accounts. Again, the model accuracy was only 86% on noisy/unclean datasets, thus limiting its real-time applicability.

(ResNet) + a bidirectional gated recurrent unit (BiGRU) + attention mechanism: max accuracy: 98.87%

Using a novel deep neural network model called RGA, consisted of a residual network (ResNet), a bidirectional gated recurrent unit (BiGRU) and an attention mechanism, and using an active learning module to efficiently expand the labelled data, Wu, Fang [27] achieved 98.87% accuracy for bot detection from the data from Sina Weibo, one of the most popular Chinese OSNs in the world. This model was successfully shown to be more effective than the state-of-the-art DL solutions (MLP [49]/LSTM [50]/ComNN [51]) and thus could be more explored in future studies, especially with limited datasets.

Graph convolutional neural network: F1-score: 67–91%

New types of solutions are a necessity to face the open challenge of identifying more advanced bots. Some of the studies addressed this issue by developing bot detection algorithms based on the social network graph in addition to the user profile features. Alhosseini et al. [52] proposed a model based on graph convolutional neural networks (GCNN) for spam bot detection, as the first attempt of this

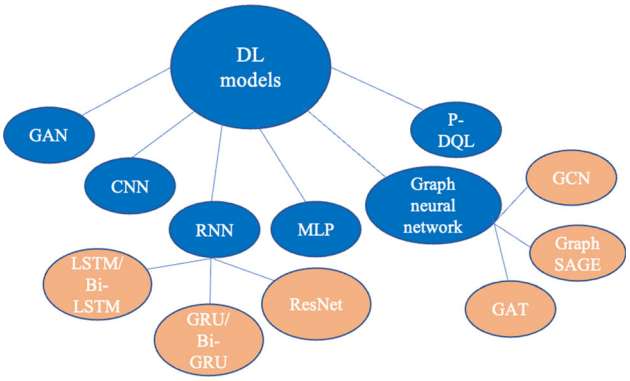

Fig. 4 Taxonomy of the DL models used for social media bot detection

kind. Their inductive representation learning approach included both the features of a node and a node's neighbourhood to better detect bots. They achieved 89% precision, 80% recall and 84% F1-score for detecting malicious accounts and social media bots. This approach outperformed (94% AUC) other classification algorithms (86% AUC with MLP and 79% with BP) with at least ~ 8% improvement. Another study by Aljohani, Fayoumi [53] achieved only 71% classification accuracy by employing a GCNN, combining Twitter's social network properties in novel altmetrics data, with other user profile features.

In the same domain, Zhao et al. [30] presented a new semi-supervised graph embedding model based on a graph attention network for spam bot detection. By learning a complex method to integrate the different neighbourhood relationships between nodes to operate the social graph, the model achieved 88% recall (15% higher than MLP), 93% precision, 91% accuracy and outperformed other graph neural networks such as GCNN, GraphSAGE and GAT. This model was shown to be stable and robust even on the imbalanced dataset. However, this approach needs to be further nourished to be compatible with the more popular user metadata, tweet content and timestamp-based CNN studies for the bot detection.

Deep Q-learning: F1-score > 90%

We found that Lingam, Rout [28] proposed a significantly different approach to combat the slower convergence of deep reinforcement models for the spambot detection task. They developed a particle swarm optimization (PSO)-based deep Q-learning algorithm (P-DQL) through generating an optimal sequence of actions associated with a goal state. The features such as spam content in the tweet and average number of tweets posted per day were represented as set of states. The model achieved around 95% precision and more than 90% F1-score, recall and G-measure. More importantly, they showed that the performance improved over adaptive deep Q-Learning, feed-forward neural network (FFNN), regularized deep neural network (RDNN), content-based deep reinforcement learning (C-DRL) and social network analysis-based deep reinforcement learning (SNA-DRL), thus could prove superiority over some of the deep learning approaches.

4.3 Comparison between DL and traditional ML techniques

Deep learning is basically a specialized subset of machine learning; hence, we can categorize the algorithms as traditional ML and DL. A traditional ML can be as simple as linear regression in which the algorithm learns to make a prediction based on patterns and inference. Deep learning algorithms can be regarded as more sophisticated and mathematically complex evolution of traditional ML algorithms. These algorithms can analyse data with a logical structure similar to how human brain makes intelligent decisions, by using a layer-by-layer processing strategy.

In the past few years, DL approaches begin to be more wildly used in social media bot detection and came up with some advantages over traditional ML approaches with better generalization performance, especially in case of big data.

Here, we followed up some of the DL-based social media bot detection studies that showed a comparison with the ML methods, either by applying the ML models on their data or by comparing with the previous studies, wherever appropriate, as also shown in Table 1. We, ourselves, did not search for related ML studies for such comparison as this was out of scope for the review.

By concatenating handcrafted and automatically learnt features from posting contents and temporal behaviours, the CNN + Bi-LSTM models (~ 90% accuracy) as proposed in [41] outperformed RF models (~ 87% accuracy) proposed by [55, 61], with a sufficiently large dataset of 37,497 malicious bots and 37,941 legitimate accounts. Wei and Nguyen, 2019, also showed that CNN + LSTM models performed better than Twitter, human annotators, BotOrNot model with supervised RF classifier [60], decision tree, BayesNet and Decorate classifiers [7], unsupervised stream clustering algorithms (StreamKM + + and DenStream) [58] and generic statistical approaches in [59]. Kudugunta and Ferrara [33] showed a detailed comparison of their proposed contextual LSTM architecture performance for social media bot detection at tweet level, with a number of ML methods. For account level bot detection with user metadata, RF and AdaBoost classifier with data enhancement techniques achieved 6–7% higher performance than a 2-layer neural network (NN) model. For tweet-level bot detection with textual content and tweet metadata, the LSTM model achieved ~ 4–20% higher

accuracy in comparison with metadata-based LR/SGD/RF/AdaBoost classifiers.

Ilias and Roussaki [47] showed that, a DL architecture that integrated an attention mechanism at the top of the Bi-LSTM layer, outperformed (in terms of all of the recall, F-measure, accuracy and AUROC scores) the existing approaches (SVM, RF, LR, DNA inspired behavioural modelling or digital DNA compression methods), to classify tweets. More specifically, recall score increased by at least 12.6% and up to 223.5%, F-measure score by 8.5–53.5%, accuracy by 7.7% and AUROC by at least 12.1%. A similar study by Rajendran et al. [46] showed that a Bi-LSTM network performed the best with 98.88% accuracy among RNN, LSTM, GRU, Bi-GRU and MLP models and outperformed the AdaBoost classifier by $\sim$ 21%. A different approach by [23, 29] using ML methods for the final classification of human vs. bots showed that by combining user data with tweet metadata and CNN predictions from tweet contents, RF and gradient boosting classifier outperformed KNN, support vector, decision tree, AdaBoost, Gaussian NB, Linear Discriminant and Multilayer Perceptron models.

Bot-DenseNet, suggested by [43], was either compatible to previous supervised (RF, one-class classifier, AdaBoost, LR and decision tree classifiers) [6, 61, 62] or showed slightly lower performance when compared to LSTM autoencoder-based feature extraction followed by hierarchical density-based algorithm[47]. Using BERT and neural network, Heidari and Jones, 2020, showed that accuracy and overall performance improved over BotOrNot [60], RF/decision tree/BayesNet/Decorate [7], StreamKM + + (based on K-means) and DenStream (based on DBSCAN) [58], Naïve Bayes, Jrip, decision tree [59] and unsupervised sequence clustering [57].

The study by Wu et al. [34] proposed a DL framework combining a residual network (ResNet), a bidirectional gated recurrent unit (BiGRU) and an attention mechanism and the performance improved over RF, LR and SVM by $\sim$ 1–2% accuracy measures, from metadata-based, interaction-based, content-based and timing-based features.

A comparison of the semi-supervised graph embedding model for spam bot detection, as proposed by C. Zhao et al., (2020), with the classical machine learning algorithms showed an improvement of 15% in F1-score than RF models, an improvement of 15% in recall than MLP models and more than 60% improvement in F1-score, recall and precision than the BP models. The authors rendered this higher performance of such models with combined features and graph structure inputs to two reasons, 1. ML models considered either the feature set or graph structure and 2. The influence of imbalanced dataset was increased when traditional ML algorithms were used.

The deep Q-network architecture designed by Lingam et al. [28] showed approximately 5–20% improvement of precision, recall, F1-score and G-measures over the baseline algorithms (FFNN, RDNN, SNA-DRL, C-DRL and ADQL), from a combination of tweet-based, user profile-based and social graph-based features. The deep forest algorithm proposed by Daouadi et al. [58] yielded an accuracy of 97.55% information from the metadata of user profiles and posts, which was $\sim$ 2% higher than the RF model and outperformed other traditional ML algorithms such as bagging (B), AdaBoost (AB), random forest (RF) and simple logistic (SL), in terms of AUC measures.

We found only one study by Dukic et al. [42] that reported that LR models (weighted F1-score of 83%) outperformed deep neural networks (DNN) (81.78%) by using contextualized BERT embedding from tweets with some additional features that represent emojis, retweets, URLs, mentions and hashtags. They also demonstrated better interpretability of LR models over DNNs.

Considering all these facts together, we concluded that DL models outperformed ML models in almost all of the cases we reviewed that attempted to do such a comparison, with $\sim$ 5% exceptions. Data suggest that this high performance can be mainly rendered to the higher capability of processing complex data such as textual content and large volume and imbalanced datasets by DL algorithms.

5 Discussion

The objective of the current review study was to figure out the efficacy of deep learning techniques to mitigate current and future challenges in social media bot detection, considering this field is going through a very fast pace of evolution. Our review study showed that deep learning techniques in general are very successful in discriminating human and bot accounts in highly heterogeneous real datasets, in most cases over traditional machine learning models. We detected some special DL subfields/models/architecture that should direct the future bot research, such as ensembled models with fusion of various handcrafted and automatically learnt features, graph neural networks for bot cluster detection over feature-based model, latest transformer models, generative adversarial networks with an anomaly detection approach and combination of real and synthetic data-based models. We provided an accumulated information of all the datasets available for this type of study and analysed the significance and efficacy of currently used features over one another. Finally, we identified the significant challenges and improvement areas in DL-based bot detection research, irrespective of the social media platforms.

The fusion of features extracted by different types of DL models (e.g. CNN and LSTM extracted features from tweets) for the final classification has been proved effective in a number of studies [29, 35]. Therefore, future studies should explore more ensembled models and new strategies for the fusion of two or more model predictions. In contrast, the study by Halvani and Marquardt [36] suggested from their preliminary experiments that tweet-based simple feed-forward neural network outperformed advanced approaches such as CNN and LSTM models. However, generation of useful input vectors from tweets with the fast space of change in bot strategies can be highly challenging. Combination of handcrafted and automatically learnt features (content, behavioural and temporal) worked well in the previous studies [25, 41]. In addition, new type of features should be investigated to boost up the accuracy for real-world implementation and ensure generalizability. Graph neural network-based methods should be explored more in addition to individual feature-based model, to facilitate the detection of social media bot cluster or coordinated attack or campaigns [45].

Latest NLP models have been highly successful to generate more meaningful input vectors from tweets for a DL model [36, 43, 45]. Fine tuning of such transformer parameters and exploring the improvement scopes may further boost up these models performance [42, 52].

Most of the studies showed the superior performance of DL techniques over ML ones. This means that layered network has the property of understanding the pattern in the input data, as the data traverse through the hidden layers. This basic characteristic is missing in traditional ML models which rely on human-fed features. Still, at least one study by [42] found that deep neural network models are only redundant to shallow ML models by a BERT-based bot detection approach along with exploratory data analysis of tweets. These findings should be investigated more with advanced or stacked DL models and by cross-validation through different datasets or platforms. In addition, future research should invest on the explainability of DL model outputs in general, to make it as advantageous and interpretable as ML models. In addition, the current review particularly reported the benefits and pitfalls of deep learning approaches and presented a comparison with ML algorithms, if reported on the DL articles. Hence, future reviews may want to inform the researchers of the overall scenario of the artificial intelligence research including all the classical machine learning and deep learning studies available in social bot detection, identify their respective potential and gain deeper insights.

Generating large dataset from a minimal amount of labelled data by leveraging state-of-the-art oversampling techniques can potentially flourish this research field, by reducing overfitting and increasing bot detection accuracy. Transfer learning from another domain also may assist to overcome the limitation of labelled data.

Finally, one of the major challenges in the social media bot detection research is that the attributes and behaviour pattern of the social media bots are continuously evolving and are growing more complex, sophisticated and optimized for particular activities. Thus, new and adaptable approaches to the evolution of social media bots are highly required. Real-time processing by leveraging GANs or other anomaly detection approaches, unsupervised techniques or novel learning modules, can be keys to combat the evolution of bots. Considerable efforts are required to develop generalized models among social media platforms such as Twitter Instagram, Facebook and electronic mail services such as Gmail or so. Application of bot detection techniques to other areas such as fake news detection or enriching the quality of news events is a potential area to explore. Future research should be directed towards bot-type classification in addition to bot classification, which is expected to increase the overall efficacy.

From the methodical perspective, for a systematic review, we could possibly start with a literature review in this field; however, there is risk of not being thorough or fair. Therefore, we decided to do a systematic literature review in the first place that followed a pre-defined and fair search strategy. Also, other than the topic level inclusion and exclusion criteria, we did not introduce a study quality assessment criterion in this review, which could be a limitation. However, due to limited number of studies in this area so far and for this might be more appropriate in clinical research reviews, we did not put an extra ‘quality’ filter.

We could do a quantitative study, where the data from the reviewed studies using meta-analytic techniques are combined, and more real and unknown effects are possibly detected that individual smaller studies are unable to detect. Hence, the heterogeneity of the input data, features and the modelling approaches made it difficult to present a useful and concrete statistical analysis outcome on the results. Our study presents a baseline quantitative synthesis on the results though, which should enable the future researchers to present the next level of analysis with more statistical power, with the inevitable additional publications in this field. There is a concept of tertiary review which is a systematic review of systematic reviews, in order to answer wider research questions. As there are only six review articles we found in this field, we did not go forward in that direction either. However, any future study directed to tertiary review might figure out some interesting aspects of the social bot detection evolution path. and how the technical approaches are changing with the evolution.

6 Conclusions and future direction

In this study, we followed a systematic approach to review the DL applications, as a state-of-the-art and highly advanced technology, in the social media bot detection research to assess the current status and critical challenges. Our review shows that DL-based techniques can be much effective and may potentially outperform traditional ML approaches, with few exceptions that definitely represent a great room for further research. In general, DL methods can make better use of textual features than other ML methods and in many cases fusion of different DL models produced a better performance such as hybrids of CNN and LSTMs exhibited a consistently reliable performance across studies. Sophisticated NLP models such as Google Transformers and combination of multiple types of features in the intermediary pre-processing steps may further boost up the classification performance. However, research suggests that feature-based models may not be always the appropriate one, for example to detect co-ordinated attacks, and graph neural network models could be an effective and alternative solution in that regard.

Deep learning algorithms especially generative adversarial networks or semi-supervised techniques may play an important role to leverage anomaly detection approach to address the major challenge of continuous change in the overall social media environment and rapid evolution of social media bots. Retrainable models through real-time processing would be another solution to this issue. Finally, most of the models are confined on twitter now. Therefore, leveraging the DL solutions to overcome similar issues in other platforms may potentially increase the usability and impact of this research to a great extent.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00521-023-08352-z.

Acknowledgements This work was supported by Zayed University, UAE, under the RIF research grant R20132.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article and agree to the publishing of its content.

Data availability All data generated or analysed during this study are included in this published article (and its supplementary information files).

References

1. Smith C (2017) 388 amazing twitter statistics and facts. DMR (February 2017)
2. Alothali E, Hayawi K, Alashwal H (2020) Characteristics of similar-context trending hashtags in Twitter: a case study. In: International Conference on Web Services. 2020. Springer
3. Gao H et al (2011) Security issues in online social networks. IEEE Internet Comput 15(4):56–63
4. Rathore S et al (2017) Social network security: issues, challenges, threats, and solutions. Inf Sci 421:43–69
5. Gupta A, Lamba H, Kumaraguru P (2013) $1.00 per rt# bostonmarathon# prayforboston: analyzing fake content on twitter. In: 2013 APWG eCrime researchers summit. 2013. IEEE
6. Varol O, et al. (2017) Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the international AAAI conference on web and social media
7. Yang C, Harkreader R, Gu G (2013) Empirical evaluation and new design for fighting evolving twitter spammers. IEEE Trans Inf Forensics Secur 8(8):1280–1293
8. Cresci S (2020) A decade of social bot detection. Commun ACM 63(10):72–83
9. Kantepe M, Ganiz MC (2017) Preprocessing framework for Twitter bot detection. in 2017 International conference on computer science and engineering (ubmk). 2017. IEEE
10. Alarifi A, Alsaleh M, Al-Salman A (2016) Twitter turing test: identifying social machines. Inf Sci 372:332–346
11. Chu Z et al (2012) Detecting automation of twitter accounts: Are you a human, bot, or cyborg? IEEE Trans Dependable Secure Comput 9(6):811–824
12. Goodfellow I et al (2020) Generative adversarial networks. Commun ACM 63(11):139–144
13. Alothali E, et al. (2018) Detecting social bots on twitter: a literature review. In: 2018 International conference on innovations in information technology (IIT). 2018. IEEE
14. Balaji T, Annavarapu CSR, Bablani A (2021) Machine learning algorithms for social media analysis: a survey. Comput Sci Rev 40:100395
15. Collins B, et al. (2020) Method of detecting bots on social media. A literature review. In: International conference on computational collective intelligence. Springer
16. Latah M (2020) Detection of malicious social bots: a survey and a refined taxonomy. Expert Syst Appl 151:113383
17. Orabi M et al (2020) Detection of bots in social media: a systematic review. Inf Process Manage 57(4):102250
18. Yang Z et al (2014) Uncovering social network sybils in the wild. ACM Trans Knowl Discov Data (TKDD) 8(1):1–29
19. Geiger RS (2016) Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. Inf Commun Soc 19(6):787–803
20. Stieglitz S, et al. (2017) Do social bots dream of electric sheep? A categorisation of social media bot accounts. arXiv preprint arXiv:1710.04044.
21. Grimme C et al (2017) Social bots: human-like by means of human control? Big data 5(4):279–293

22. Brereton P et al (2007) Lessons from applying the systematic literature review process within the software engineering domain. J Syst Softw 80(4):571–583
23. Sengar SS et al (2020) Bot detection in social networks based on multilayered deep learning approach. Sens Transducers 244(5):37–43
24. Zegzhda PD, Malyshev E, Pavlenko EY (2017) The use of an artificial neural network to detect automatically managed accounts in social networks. Autom Control Comput Sci 51(8):874–880
25. Cai C, Li L, Zengi D (2017) Behavior enhanced deep bot detection in social media. In: 2017 IEEE international conference on intelligence and security informatics (ISI). IEEE
26. Al-Qurishi M et al (2018) A prediction system of Sybil attack in social network using deep-regression model. Futur Gener Comput Syst 87:743–753
27. Wu Y et al (2021) A novel framework for detecting social bots with deep neural networks and active learning. Knowl-Based Syst 211:106525
28. Lingam G et al (2020) Particle swarm optimization on deep reinforcement learning for detecting social spam bots and spam-influential users in twitter network. IEEE Syst J 15(2):2281–2292
29. Katarya R, et al. (2020) Bot detection in social networks using stacked generalization ensemble. In: The international conference on recent innovations in computing. Springer.
30. Zhao C et al (2020) An attention-based graph neural network for spam bot detection in social networks. Appl Sci 10(22):8160
31. Morstatter F, et al. (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). IEEE
32. Heidari M, Jones JH (2020) Using bert to extract topic-independent sentiment features for social media bot detection. In: 2020 11th IEEE annual ubiquitous computing, electronics & mobile communication conference (UEMCON). IEEE
33. Kudugunta S, Ferrara E (2018) Deep neural networks for bot detection. Inf Sci 467:312–322
34. Wu B et al (2020) Using improved conditional generative adversarial networks to detect social bots on Twitter. IEEE Access 8:36664–36680
35. Ping H, Qin S (2018) A social bots detection model based on deep learning algorithm. In: 2018 IEEE 18th international conference on communication technology (icct). IEEE
36. Halvani O, Marquardt P (2019) An unsophisticated neural bots and gender profiling system. In: CLEF (Working Notes)
37. Luo L, et al. (2020) Deepbot: a deep neural network based approach for detecting Twitter bots. In: IOP Conference Series: Materials Science and Engineering. 2020. IOP Publishing
38. Wei F, Nguyen UT (2019) Twitter bot detection using bidirectional long short-term memory neural networks and word embeddings. In: 2019 First IEEE international conference on trust, privacy and security in intelligent systems and applications (TPS-ISA). IEEE
39. Pennington J, Socher R, Manning CD (2014) Glove: global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)
40. Onose C, et al. (2019) A hierarchical attention network for bots and gender profiling. In: CLEF
41. Mou G, Lee K (2020) Malicious bot detection in online social networks: arming handcrafted features with deep learning. In: Social informatics: 12th International Conference, SocInfo 2020, Pisa, Italy, October 6–9, 2020, Proceedings. 2020, Springer-Verlag: Pisa, Italy. p. 220–236
42. Dukić D, Keča D, Stipic D (2020) Are you human? Detecting bots on Twitter using BERT. In: 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). pp. 631–636
43. Martín-Gutiérrez D et al (2021) A deep learning approach for robust detection of bots in twitter using transformers. IEEE Access 9:54591–54601
44. Mohammad S, et al. (2019) Bot detection using a single post on social media. In: 2019 third world conference on smart trends in systems security and sustainablity (WorldS4)
45. Gao T et al (2020) A content-based method for sybil detection in online social networks via deep learning. IEEE Access 8:38753–38766
46. Rajendran G et al (2020) Deep temporal analysis of Twitter bots. Springer Singapore, Singapore
47. Ilias L, Roussaki I (2021) Detecting malicious activity in Twitter using deep learning techniques. Appl Soft Comput 107:107360
48. Mazza M, et al. (2019) RTbust: exploiting temporal patterns for botnet detection on Twitter. In: Proceedings of the 10th ACM Conference on Web Science
49. Lian Y et al (2019) An internet water army detection supernetwork model. IEEE Access 7:55108–55120
50. Makkar A, Kumar N (2020) An efficient deep learning-based scheme for web spam detection in IoT environment. Futur Gener Comput Syst 108:467–487
51. Pei W, Xie Y, Tang G (2018) Spammer detection via combined neural network. In: Machine Learning and Data Mining in Pattern Recognition. Springer International Publishing. pp. 350–364
52. Alhosseini SA, et al. (2019) Detect me if you can: spam bot detection using inductive representation learning. In: Companion proceedings of the 2019 world wide web conference. 2019, Association for Computing Machinery: San Francisco, USA. p. 148–153
53. Aljohani NR, Fayoumi A, Hassan S-U (2020) Bot prediction on social networks of Twitter in altmetrics using deep graph convolutional networks. Soft Comput 24(15):11109–11120
54. Färber M, Qurdina A, Ahmedi L (2019) Identifying twitter bots using a convolutional neural network. In: CLEF
55. Braker C et al (2020) BotSpot: deep learning classification of bot accounts within twitter. Internet of things, smart spaces, and next generation networks and systems. Springer, pp 165–175
56. Staykovski T (2019) Stacked bots and gender prediction from twitter feeds. In: CLEF (Working Notes)
57. Lingam G, Rout RR, Somayajulu DV (2019) Deep Q-learning and particle swarm optimization for bot detection in online social networks. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE
58. Daouadi KE, Rebaï RZ, Amous I (2019) Bot detection on online social networks using deep forest. In: Computer science on-line conference. Springer
59. Ahmed F, Abulaish M (2013) A generic statistical approach for spam detection in online social networks. Comput Commun 36(10–11):1120–1129
60. Davis CA, Varol O, Ferrara E, Flammini A, Menczer F (2016) Botornot: a system to evaluate social bots. In: WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 273–274
61. Yang K-C, Varol O, Hui P-M, Menczer F (2020) Scalable and generalizable social bot detection through data selection. In: Proceedings of the AAAI conference on artificial intelligence, vol. 34, pp. 1096–1103
62. Rodrguez-Ruiz J, Mata-Snchez JI, Monroy R, Loyola-Gonzalez O, pez-Cuevas AL, (2020) A one-class classification approach for bot detection on twitter. Comput Secur 91:101715