Are You a Cyborg, Bot or Human? - A Survey on detecting fake news spreaders

WAJIHA SHAHID, YIRAN LI, DAKOTA STAPLES, GULSHAN AMIN, SAQIB HAKAK, and Ali Ghorbani

Canadian Institute for Cybersecurity, Faculty of Computer Science, University of New Brunswick, Fredericton, Canada (e-mail: wajiha.shahid@unb.ca, yli30@unb.ca, dstaples@unb.ca, gulshan.amin@unb.ca, saqib.hakak@unb.ca, ghorbani@unb.ca)

Corresponding author: Wajiha Shahid (e-mail: wajiha.shahid@unb.ca).

ABSTRACT
One of the major components of Societal Digitalization is Online social networks (OSNs). OSNs can expose people to different popular trends in various aspects of life and alter people’s beliefs, behaviors, and decisions and communication. Social bots and malicious users are the significant sources for spreading misinformation on social media and can pose serious cyber threats in society. The degree of similarity of user profiles of a cyber bot and a malicious user spreading fake news is so great that it is very difficult to differentiate both based on their attributes. Over the years, researchers have attempted to find a way to mitigate this problem. However, the detection of fake news spreaders across OSNs remains a challenge. In this paper, we have provided a comprehensive survey of the state of art methods for detecting malicious users and bots based on different features proposed in our novel taxonomy. We have also aimed to avert the crucial problem of fake news detection by discussing several key challenges and potential future research areas to help researchers who are new to this field.

INDEX TERMS
Cyborg, Deep fake, Deceptive content, Fake news detection, malicious user, Misinformation, news propaganda, Social bots, Social media,

I. INTRODUCTION

In the present era, our society is gradually getting digitalized as the Internet is the main source of information, entertainment and communication. The central part of societal digitalization is the online social networks (OSNs), such as Facebook, Twitter etc. OSNs are nowadays an integral part of people’s daily life, giving users the platform to interact, express themselves and access news [1] [2]. Facebook has 1.88 billion daily users and Twitter has 199 million monetizing daily users [3]. The convenience of social networking has brought the world together due to ease of communication and access to information [4]. At the same time, this easy access to information comes with its drawback such as the excessive propagation of fake news in the form of propaganda, misinformation etc. [5]. More than 40% of traffic to websites spreading fake news is redirected through links on Facebook, Instagram, and Twitter [6] due to their easy access and rapid dissemination [7]. Spread of fake news has even been listed as a major threat to society by the World Economic Forum [8]. Fake news can be described as a kind of news story involving intentional false information to alter users’ minds on social media [9]. The dissemination of fake news significantly affects personal reputation and public trust. A survey of 92,000 consumers on a variety of digital topics in 46 markets to see the trust ratio in online news all over the world was conducted by Reuters in 2021 1 as summarized in Figure 1. Results show that Finland had the highest share of respondents agreeing "you can trust news most of the time" at 65% which marks a 9%-point increase since the last edition of the report. The United States made little progress and only 29% of people trusted the news most of the time based on previous experiences.

Although the topic of fake news is not new, the study of fake news spreaders’ on social media is a developing topic [10]. There are currently numerous challenging issues [11] which currently require further investigation such as differentiating a user account from automated accounts. Automated accounts are controlled by algorithms known as social bots [12]. Multiple social bots can take the form of a social botnet. Social botnet is a group of social bots created and

1https://www.statista.com/chart/7248/where-people-trust-the-news-most-and-least/
controlled by a botmaster. They perform malicious activities, such as creating multiple fake accounts, spreading spam, manipulating online ratings, and so on [13].

A recent study estimated that there are 321 million Twitter accounts out of which 48 million are bot accounts, i.e., 15% of all Twitter accounts [14]. The automated nature of bots makes it easy to achieve a large scale impact when spreading misinformation [15]. Analyzing large-scale social data collected during the Catalan referendum for independence on October 1, 2017, consisting of nearly 4 million Twitter posts generated by almost 1 million users revealed that bots produced 23.6% of the total number of posts during the event. A Barracuda report reveals that Automated traffic makes up 64% of internet traffic. Just 25% of it was made up by good bots, while 39% of all traffic was from bad bots as shown in Figure 2. Figure 3 shows the bad bot traffic in North America accounts for 67% of bad bot traffic.

Despite the efforts to detect social bots, it is still difficult to distinguish them from legitimate users which makes it a challenge [16]. The process of social bots identification and their detection is cyclic. New bots are created which spread fake news. And then new social bots filters are derived to tackle them, while old bots mutate into advanced ones [17]. Sometimes automated accounts show human characteristics giving birth to “Cyborg” [18]. These bots can even interact as legitimate users when the human takes over the bot profile from time to time.

The aforementioned statistics clearly state the need to come up with an effective solution to identify and detect the fake news spreaders. Various research studies have been carried out in the past to identify the nature of the fake news spreaders accounts.

Many surveys have been conducted that reviewed bot detection, human-based detection and cyborg detection along with their taxonomy separately but none of the surveys reviewed all three of them together. The purpose of our survey is to review the recent research done on the subject of bot detection, human detection and cyborg detection. The existing surveys on this topic are summarized in Table 1. Recently in one of the studies, the authors reviewed bot and cyborg detection algorithms, whereas and reviewed only bot detection algorithms, with the former contributing a taxonomy of their work. Authors discussed human user-based detection algorithms. As clearly shown from Table 1, our survey is different from the existing surveys as it not only deals with bot and human-based detection methods but also hybrid-based methods, which include the detection of cyborgs.

To conduct this study, due to an extensive volume of literature on this topic, we used keywords such as Bot detection on social media, Fake news users, Human and Bot detection on twitter and other similar keywords. Based on these keywords, relevant papers were extracted published...
within the last three years from reputed databases such as IEEE, Springer, Elsevier and ACM. From the list of extracted papers, we excluded conference articles, book chapters and shortlisted technical journal articles with a reasonably good number of citations. Furthermore, popular journals on human Psychology were found to search for papers on Fake news targets and characteristics of people who are impacted by fake news. We have also listed the acronyms used in this work in Table2.

All of our contributions in this paper are summarized as follows:

- An extensive survey over the current state of art methods in detecting bot, human and hybrid-based accounts.
- A novel taxonomy on fake news spreaders detection approaches.
- Identify and discuss existing and emerging new challenges, and future research agendas.

The remaining part of the paper is organized as follows: Section II describes what the fake news is, further discussing its various components and features. In Section III, we propose a taxonomy, and based on that, explain the existing studies. Section IV explains all the methods used to detect the fake news spreaders. In Section V, the challenges and issues in detecting fake news spreaders are discussed. Section VI outlines potential future directions leading to concluding this paper in Section VII.

II. FUNDAMENTALS OF FAKE NEWS

In this section, we discuss the fundamental concepts of fake news. The major fundamental concepts discussed are the definition, components, types, and features of fake news.

A. DEFINITION OF FAKE NEWS

The spread of fake news has become a global issue that needs to be addressed immediately [25]. This concept drew major attention after the US elections of 2016 [8]. Fake news is defined by [11], as misleading content including conspiracy theories, rumors, clickbaits, fabricated news, and satire. [26] defines fake news as misinformation and disinformation both, including false and forged information, that is spread on social media. 

In order to clearly understand the spread of fake news, it is important components need to be discussed. These components can be divided into four main categories including creator/spreader, target victims, content and social context [30]. Figure 5 shows how fake news is spread on social media. [31]

1) Creators/spreaders

Creators generate fake news and spreads propagite it by re-sharing. They can either be humans or non-humans. Non-humans include social bots and cyborgs. [32]. Social bots are algorithms that are programmed to engage autonomously on social media. They can create content as well as increase its reach [33]. Cyborgs are a hybrid between human accounts and social bots [34].

2) Target victims

Target victims are the group of people or organizations that are impacted by fake news. They are specifically identified and targeted by fake news spreaders. Voters can be targeted in case of smear campaigns during elections [35]. Vulnerable populations include online customers being exposed to scams, patients being exposed to wrong medical information and non-digital natives who don’t have enough exposure to differentiate false news from the truth [36].

3) News content

News content comprises of non-physical and physical contents. Physical contents may include headings and visual features to attract users. Clickbait and hashtags are examples of physical content that catches viewers’ attention initially [37]. Non-physical contents contain opinions and sentiments. This is the content that results in creating polarity and change of views. Authors use strong positive or negative emotions to make their content more sensational and easily exploitable [38].

4) Social context

Social context refers to the overall social environment in which the news is being spread. Social context includes the interaction of online users with each other and the content they are interested in [39]. The social environment of online communities and users, along with the social context, determines how fast fake news is propagated through various channels of social networks [40].

C. TYPES OF FAKE NEWS

Fake news is of multiple kinds. They may be stance, satire, multi-modal, deep fake, and disinformation. Stance can be classified into four types, i.e. agree, disagree, discuss and unrelated [41]. An agreeing stance would be related to...
TABLE 1. Previous surveys on the topic vs our survey

| Survey   | Year | Bot Detection | Human User-based Detection | Cyborg Detection | Taxonomy |
|----------|------|---------------|-----------------------------|------------------|---------|
| [19]     | 2019 | ✓             | x                           | ✓                | x       |
| [23]     | 2019 | ✓             | x                           | x                | x       |
| [22]     | 2020 | x             | ✓                           | x                | x       |
| [24]     | 2021 | ✓             | x                           | x                | ✓       |
| [20]     | 2021 | ✓             | x                           | x                | x       |
| [21]     | 2021 | ✓             | x                           | x                | x       |
| Our Survey | 2021 | ✓             | ✓                           | ✓                | ✓       |

FIGURE 4. Bird eye’s view of Fake news

the headline of fake news, whereas a disagreeing stance holds contradictory information. Satire involves humor and mockery [42], it usually includes some political message or criticism in the form of humor and the tone used is generally sarcastic. Multi-modal involves spread of fake news using multiple means such as videos, images, audio, text etc. [43]. Deep fake is a type of fake news that is spread through manipulated video clips, images and recordings [44]. Deep fake is generated by using deep learning techniques. It lets a computer to generate fabricated media content. A startup named Deep Trace reported 7964 deep fake videos in 2019 and that number doubled within nine months and continues to grow exponentially [45]. Disinformation is misleading and false information that is spread in order to deceive people [46]. Disinformation has various sociopolitical repercussions [47]. Sources may spread manipulated information to deceive the audience in order to achieve political agendas or to create social havoc.

D. FEATURES USED IN DETECTING FAKE NEWS

Research work is being carried out in order to help online users uncover and recognize fake news and to develop au-
New solutions need to be developed in order to detect fake news sources [50]. In order for the early detection methods to be efficient, features of fake news need to be identified and extracted first [51]. The key features in detecting fake news are user-based features, temporal analysis, sentiment features, linguistic analysis, social context-based analysis, and network features. Details of these features are discussed below:

(i) **User-based features** include unique characteristics of users profiles that can be analyzed to find out if the person happens to be a fake news spreader [52]. User-based features can be divided into profile analysis and credibility analysis. In **user profile analysis**, user profile can be analyzed on the basis of username, age, profile picture, Geo-location and account verification status [53]. **User credibility analysis** includes information about the number of friends and followers. For example, bot accounts generally have more users in their follower list and follow very less users themselves [54].

(ii) **Temporal analysis** includes timing and frequency of the posts as well as user engagement. This helps to identify bot accounts as they have specific set patterns of online engagement [55]. Bot accounts are programmed to have a more active engagement at certain times.

(iii) **Sentiment features** involve analyzing sentiments that trigger emotional response. Bot accounts use out-of-context misleading facts to provoke emotions. A lot of the content created by fake news propagators is highly polarized and exaggerated. [56]

(iv) **Linguistic analysis** involves determining the writing patterns and formats [57]. Most fake news creators have a specific writing style and format. Fake news content can be identified by the excessive use of bold letters in headings and paragraphs. The presence of suspicious tokens such as URLs, tags, and excessive uppercase words is also a fake news feature that can be used for detection.

(v) **Knowledge-based analysis** involves verifying suspicious information...
from credible resources such as using verified websites. This can be done manually or by using AI algorithms. (vi) Social context-based analysis includes user network analysis and distribution network analyses. User network analysis is used to study the engagement patterns between online accounts. The distribution pattern focuses on the distribution of information [58]. (vii) Network features have two types of networks analyzed: Homogeneous networks and heterogeneous networks. Homogeneous networks have singular nodes and include stance networks and propagation networks [59]. Stance-based modelling determines the users’ stance on a specific idea or news. The classification is based on the agreement or disagreement between the main headline and body of the news [60]. Propagation networks analyze the relationship between posts and re-posts. Generally fake news gets re-posted excessively and faster compared to authentic ones [61]. Heterogeneous networks have multiple nodes. It involves analyzing relationships between multiple nodes, including articles, publishers, users and posts [62].

III. FAKE NEWS DISSEMINATION STUDIES BASED ON SPREADERS ACCOUNTS FEATURES

In this section, we will explore state-of-the-art approaches for fake news spreader detection based on our taxonomy as highlighted in Figure 6. All the existing studies are categorized based on three categories, i.e., source, propagation and target based features. We have identified the commonly used datasets in these studies in Table 3

| Category          | dataset          | Studies |
|-------------------|------------------|---------|
| Bot Detection     | PyTrawler 2021   | [63]    |
|                   | BadBots          | [64]    |
|                   | Cresci           | [65], [66], [67] |
|                   | PAN 2020         | [68], [69], [70] |
|                   | Social Honeypot  | [71], [66] |
| User Detection    | Twitter 15       | [72], [73], [1] |
|                   | Twitter 16       | [72], [73], [1] |
|                   | FakeNewsNet      | [74]    |
|                   | BuzzFeed         | [75], [76] |
|                   | Politifact       | [75], [76], [77], [78], [79] |
|                   | Gossipcop        | [77]    |

1) Personality feature

The personality feature includes the qualities of a fake news spreader. It is further divided into the linguistic feature, posting frequency and login interval [1]. Linguistic features include the writing style and grammar of the post/tweet [82]. Fake news is generally written in capital letters with typing errors, poor sentence structure, and many exclamation marks. All these come under the linguistic feature of detecting fake news [67]. Posting frequency means how many posts or tweets are posted in a day and the time gap between consecutive posts. The posting frequency of such accounts is visibly high with repetitive posts being posted after a fixed interval of time [83] [84]. The login interval means the time duration of each session and the gap between two consecutive sessions. Fake news spreaders will have longer login intervals, and login time is likely to be the same each day [69] [85].

2) Historical feature

Historical feature means analyzing the account’s metadata and identifying the trends such as login time of the day, posting history and posting time [66]. Metadata provides enough details about the user’s profile which helps to identify if it is a real human’s account or a bot or cyborg. Similarly, by detecting the user pattern of logging into his social media account with respect to time and analyzing the posted tweets, it can detect if the account is a fake news spreader. Spreaders generally spam by posting same fake posts after a fixed interval of time and posting many tweets at a time. Time-based analysis can provide a great insight in distinguishing fake news spreader from a normal account [77]. For example, a bot working as fake news spreader will usually post false content all day without break whereas its impossible for a human post all day as humans have other activities too [20].

3) Credibility feature

The credibility feature takes into account the authenticity of the publishing source and originality of content posted by a user by analyzing the previous posts from that account [86]. Credibility is one of the major features to identify fake news spreaders among other users [75]. Credibility can be assessed on two bases, namely, publishing source and content source. Posts from unverified or malicious URLs are less authentic and chances are more that they will contain false information or manipulated news [68]. Similarly, the content source is also an important feature to distinguish a fake news spreader from other accounts [87]. Content source means the platform where information/news is shared. Social platforms like WhatsApp, Facebook, Twitter, Instagram are all accessible to common people and people freely post about literally anything without any justification of their authentication [88]. News posted on these forums is equally likely to be manipulated, if not completely fake [89]. Following studies have used the source features in their detection models.

Lingam et al. [71] proposed SBCD and DA-SBCD methods which can detect social bots and identify social botnet...
FIGURE 6. Taxonomy on Fake News Dissemination Features based on Fake News Spreaders’ Accounts

communities in online social networks (OSNs). The efficiency of proposed algorithms gives better performance than existing schemes in terms of normalized mutual information (NMI), precision, recall and F-measure. The generated dataset was pre-processed, and machine learning algorithms determined the fake accounts. This paper aimed to identify bots effectively with the minimum possible collection of attributes on the Twitter social network. The authors used the Random Forest classifier to train the data. This study was able to find the percentage of bot accounts in each cluster. For this purpose, a previously trained classifier was used to label the data for bot accounts. This study also concluded that bots had a low follower growth rate as compared to normal accounts. On the other hand, they had high friends/followers ratio. Furthermore, the screen names of bots usually had more digits than a normal account indicating automated behaviors. Apuke et al. [90] study the underlying reason on who shares fake news, and why they share it. Cardaioli et al. [67] used the machine learning approach which could be used in evaluating the stylistic consistency of social network posts and to accomplish other kinds of analyses based on authors style, it can distinguish when posts are posted by cyborgs or bots with statistical evidence. Wu et al. [88] used DABot to detect bots and cyborgs by increasing the efficiency of the model by labeling user data and obtaining a large-scale dataset at a small cost. This paper designed a new deep neural network model RGA for detection. Kaliyar R. K. et al. [75] concluded that not only the content of the news articles is an important factor for fake news detection, but also the existence of echo chambers, a group of users with the same interests grouped together to form a community. Shu K. et al. [77] focus on understanding and exploiting user-profile features on social media for fake news detection. The authors measure users sharing behaviors, and the group representative user sets who are more likely to share news. Khaund T. et al. [20] discuss different methods to detect bots like Early Sybil, Miso’s algorithm and BotGraph. Orabi et al. [23] studied the behaviors and features of social media bots to detect the bots who spread fake news online.

B. PROPAGATION-BASED ACCOUNTS DETECTION

A propagator disseminates fake news widely to increase its reach to maximum victims [68]. The features of propagator have been classified into three main categories i.e., user engagement, time dynamics and platform-based features. Table 5 enlists the related existing research mentioning the features of fake news propagator along with the algorithms used.

1) User engagement feature

The user engagement features include the user network details and circulation of fake news between these networks [83]. Generally fake news propagators have a network of spam accounts [1]. Most of the propagator profiles have a list of spam or bot accounts in their followers and followees list [91]. Furthermore, the content of comments under posts and articles can be analyzed for relevance and to identify fake news propagators [75]. At times, the comments section contain totally irrelevant comments amidst a series of relevant comments [92]. These comments can include malicious URLs or website links advertising something, or even starting a comment war [93]. Most of these irrelevant comments are made to propagate fake news [72]. Moreover, huge number of retweets or re-shares in a small span of time is a good measure to identify fake news propagators as well.
2) Time dynamics feature

The timelines of fake news propagators consist mainly of fake news with a few real news to make them look authentic [88]. There is usually a set pattern following which posts are shared and re-shared on a daily basis [67]. The accounts used by fake news propagators are generally over active and show spam posts and comments at random times. On the other hand, genuine users usually have specific active and inactive login pattern based on their daily schedule. The fake news propagators are most active during specific important events, such as, election campaigns and social movements [69].

3) Platform-based feature

Fake news propagators accounts are found on all popular social media platforms such as Facebook, Instagram, Twitter etc., and even on other community websites like blogs, discussion platforms, fake news websites etc. These accounts are involved in “hashtag wars” on social media to propagate fake news and spread propaganda. Moreover, the fake news propagators share irrelevant URLs containing malware links, and unwanted information in comment sections and user inboxes [66]. The following papers discuss the features of fake news propagators.

[94] focused on defining the degree to which bots can exploit hashtags. The machine learning algorithms determined the fake accounts based on preprocessing datasets. This paper aimed to identify bots effectively with the minimum possible collection of attributes on the Twitter social network. Authors have used the Random Forest classifier to train the data. This study was able to find the percentage of bot accounts in each cluster. For this purpose, a previously trained classifier was used to label the data for bot accounts. This study also concluded that bots have a low follower growth rate as compared to normal accounts. They have high friends/ followers ratio. Furthermore, the screen names of bots usually have more digital numbers as compared to a normal account indicating automated behavior. Sansonetti G et al. [79] proposed a text content and social context-based model for human-based fake news detection. De Nicola et al. [21] proposed a profile feature and timeline based classification of bots, and concluded that feature-based classification deemed to perform well with detecting social and sophisticated bots. Lu et al. [1] proposed a graph-based method to classify a fake news from real one, and highlighted the suspicious re-tweeters. Mendoza et al. [93] has proposed a leveraging graph-based representation approach which can learn a network-based representation of users, and is suitable for effectively detecting social bots. This method defines a semi-supervised algorithm which accurately detects groups of social bots, by performing an in-order traversal of the proximity graph. Vogel et al. [68] focused on the detection of fake news on Twitter in English and Spanish, and have followed the approach of identifying the fake news propagators by extracting emotions behind the tweets. Pozzana et al. [83] have analyzed two Twitter datasets: the collection from French presidential election 2017 and the hand-labeled tweets from three groups of bots active in as many viral campaigns, to detect different type of bots. Bello et al. [69] proposed a multilingual approach of identifying fake news spreaders on Twitter data. They manually engineered domain-specific features covering behavioural, lexical and psycho-linguistic aspects and evaluated them using traditional machine learning models. The focus of this paper was to test domain-specific features on different types of classifiers first, and finally evaluating a pure multilingual approach on a combined English and Spanish dataset. The authors extended their experiment to a multilingual model design using less preprocessing and feature selection. Their study has demonstrated the importance of selecting domain-specific features in the domain of fake news identification. They concluded that it was possible to detect fake news spreaders based on a limited dataset of 300 Twitter users, by applying gradient boosting to a set of lexical, behavioural and psycho-linguistic features.

| Study                  | Source | Historical | Personality | Credibility | ML | DL | NLP |
|------------------------|--------|------------|-------------|-------------|----|----|-----|
| Lingam et al. (2020)   | ✓      |            |             | ✓           | ✓  |    | x   |
| Apuke et al. (2021)    | x      |            |             | ✓           | ✓  |    | x   |
| Cardiallo et al. (2021)|       | ✓          |             | ✓           | x  |    | x   |
| Wu et al. (2021)       | x      | ✓          |             | x           | ✓  | x  | x   |
| Kaliyar et al. (2021)  | x      | ✓          |             | ✓           | ✓  | x  | x   |
| Khairud et al. (2021)  | ✓      | ✓          |             | x           | ✓  | x  | x   |
| Orabi et al. (2020)    | ✓      | ✓          | ✓           | ✓           | ✓  | x  | x   |

C. TARGET-BASED ACCOUNTS DETECTION

The target features identify end users that are affected by the fake news. A target can be a human, bot or cyborg depending on the nature and domain of fake news [69]. Although fake news can reach almost all the users through social media, an easy target will be those people that are more vulnerable and prone to getting influenced by the fake news [87]. In order to understand and identify a potential victim of fake news, we can make use of victim dynamics feature as described below. Table 6 shows related studies which have made use of target victim features to detect fake news.
1) Victim Dynamics feature

Victim dynamics mean thoroughly understanding the details of the end user. The details can include age, gender, education history, account creation history, network of followers, location etc. From the study [77], we find that generally new users with limited exposure to social media are targets of fake news spreaders, as they tend to believe anything presented to them due to lack of exposure. Teenagers and aged people with limited knowledge of possibilities of fake news on social media are an easy target [72]. Similarly, people with low qualifications and coming from rural areas are more prone to be the victims of fake news [95]. Following papers have discussed the features of a target-based account. Yuan et al. [73] have proposed a human-based fake news detection technique SMAN which can detect fake news within 4 hours with an accuracy of over 91 percent, which is much faster than the state-of-the-art models. Chowdhury et al. [74] have proposed a credibility score-based model which detects the fake news by observing credibility of both publishers of the news and its users, making the model useful in fake news detection. Zhang et al. [30] propose a combination of cyborgs and bot detection scheme and discuss practical solutions versus research-based solutions. Ahmed et al. researched about who inadvertently shares fake news [96] and introduced a bias which enabled users to self-report whether they shared deep fakes or not. The model also took into perspective the users who may have shared deep fakes and did not realize it. Albadi et al. [72] proposed a regression model which detect bots spreading hateful messages against various religious groups on Arabic Twitter. Ajesh et al. [87] detected fake news user profiles using random forest, optimised Naïve Bayes and support vector machine algorithms. Rodriguez-Ruiz et al. [66] took a hybrid, one-class classification approach to decide between bad bots and humans without the requirement of anomalous behavior examples. Shu et al. [7] has addressed the increasing fake news propagation on social media and addressed the features of target victims that are more prone to be affected by fake news.

IV. DISCUSSION

Fake news dissemination and identification of fake news spreaders, propagators and targets is a challenging task. The use of Artificial Intelligence has proved fruitful in this regard. Most of the existing studies have used Deep Learning, Machine Learning and Natural Language Processing methods to detect fake news spreaders through feature extraction and classification.

Machine Learning is an application of AI that enables systems to learn and identify patterns which leads to decision-making without the intervention of humans. Machine Learning algorithms have specifically seen a boost in the field of fake news feature determination. During our extensive study, we have come across various ML algorithms that have been used for this purpose. The ML algorithms are trained using large datasets so that they can automatically detect the fake news spreaders [97]. Once fake news is shared on the internet, ML algorithms check its contents and detect fake news spreaders based on different features. Researchers have been trying to train machine learning classifier to detect with higher accuracy [98]. The better trained a classifier is, the more accurate it is [99]. Within ML framework, the common algorithms that have achieved better results include Neural Networks, Naive Bayes, Decision Trees and SVM.

Natural Language Processing deals with the interactions between computer systems and languages that enables computers to understand speech and text. NLP-based algorithms are used to detect linguistic and semantic patterns in fake news [100]. NLP supports AI in performing language related tasks such as creation of dialogues and interpreting words and sentences with ease [101]. Commonly used NLP techniques that achieved remarkable results compared to others include TF-IDF, LSA and LDA.

Deep learning is a branch of AI that comprises of artificial neural networks. The detection of fake news spreaders is complex and there are few shortcomings when it comes to using NLP alone for detection. DL and NLP techniques can be used in conjunction to improve automatic detection [102]. DL involves systematic representation of data and text analytics. The learning can be supervised or unsupervised. Some common DL techniques used for fake news spreaders
detection with good performance include CNN, LSTM, Bi-
GRU etc.

During our critical literature review, we found that the re-
searchers have encountered many limitations out of which we
have listed the most common challenges and future directions
in our next section which can be addressed to build a more
accurate fake news spreaders detection system.

V. RESEARCH CHALLENGES
Table 7 summarizes the most common challenges found by
the researchers in creating an efficient fake news spreaders
detection model.

| Study                 | Target Dynamics | Techniques |
|-----------------------|-----------------|------------|
| Yuan et al. (2020) [73]| ✓               | ✓          |
| Chowdhury et al. (2020) [74]| ✓         | ✓          |
| Zhang et al. (2020) [30]| ✓               | ✓          |
| Ahmedet al. (2021) [96]| ✓               | ✓          |
| Albadi et al. (2019) [72]| ✓             | ✓          |
| Ajesh et al. (2021) [87]| ✓               | ✓          |
| Rodrigo-Ruiz et al. (2020) [66]| ✓       | ✓          |
| Shu et al. (2019) [7]| ✓               | ✓          |

prominent in deep learning models. So, researchers should
provide their datasets publicly so that other researchers can
contribute to keep it updated [104]. Research that implements
existing detection models and tests them on the some real
public dataset is also needed. [23]

In addition, based on the studies we surveyed so far, we
believe that the fake news detection lacks comprehensive
dataset. Most of the datasets available, are based on political
news. On the other hand, there are other aspects of social
media concerning health, education, religion etc. Not much
has been done in this regard. It is also challenging to create
datasets of that caliber because of lack of accessibility to in-
formation of users due to privacy and confidentiality aspects.
Finally, papers using hybrid-based fake news detection, such
as [75] and [78] show that the accuracy of the experimental
result when using both news content and social context, is
more accurate as compared to using either news content
or social context. However, collecting characteristics of the
users can be very challenging and need regular updates to
their status. Besides, Shu K. et al. [77] indicates that there are
both explicit and implicit features contained in user profile
and the implicit features such as personalities are very useful
for analysis. The collection of implicit features of a large
number of users would be even more difficult than the explicit
features.

C. BIASES IN SURVEYS
One research challenge that was encountered is biases in the
survey, and the way metrics were evaluated in the surveys
created and conducted. For example, when researching who
shared fake news, Apuke et al. [90] only sampled people from
Nigeria, which means that the data may not generalize to
other countries. [96] introduced a bias in the way that users
had to self report whether they shared deep fakes or not,
meaning some users may have shared deep fakes and did not
realize it.

D. MALICIOUS HUMAN AUTHORS
A fake news spreader can be a simple human user sometimes.
This person can write their posts in such a way to avoid
detection. They can meticulously craft a post which looks
very real and not so different than a real post. In addition, a
Table 7. Open challenges in the detection of fake news spreaders

| Challenges                  | Causes                                                                                     |
|-----------------------------|-------------------------------------------------------------------------------------------|
| Cyborg Vs Real User         | Cyborg are partially controlled by humans and partially by bots with fresh content on their profiles every now and then. |
| Limitations in training data| Most datasets are relatively small and are not comprehensive which impacts deep learning greatly. In addition, collecting user features, especially implicit features, is very challenging, and must be updated frequently. |
| Biases in Surveys           | Surveys Conducted by other researchers can include biases when polling users               |
| Malicious Human Authors     | Malicious Humans are extremely hard to detect as they can create a legitimate account with good metrics and often wait to share fake news to fool ML models. |
| Platform difference         | Most research models focus on Twitter only for fake news detection meaning other platforms on which fake news is posted such as Instagram, Facebook may be hard to detect |
| Cross Platform              | Fake news may be spread on lesser known social media platforms whereas legitimate users share on known ones |

malicious actor may choose to wait some time before posting fake news. By doing this they may post regular content that goes undetected and in doing so gain real followers, friends, comments, and more. These metrics are all used by ML models to detect fake news accounts and posts. In the case of a human posting fake news and deliberately crafting posts maliciously, the challenge becomes much greater to detect it.

E. PLATFORM DIFFERENCE

Many papers focus on fake news detection method on a specific platform. For example, Twitter is chosen by many studies. The datasets used for their experiments are also Twitter users. As social media have various features and functions which are sometimes not similar to other social media; for example, Facebook and Instagram host a short story feature which exists for 24 hours and can be exploited to share fake news, whereas other platforms like Twitter just focus on feeds. So in order to detect fake news on other social media like Facebook, Instagram Youtube, LinkedIn, etc., the detection system may need some changes.

F. CROSS PLATFORM

Another major challenge in this area is number of available social media platforms besides the popular ones. Every region in the world has its own non popular social media platform which fake news spreaders target to spread fake news and then the genuine users take this information across different platforms, because they believe it to be true. It is highly unlikely to identify such accounts. There must be some cross platform control mechanism to verify links and articles before sharing them.

VI. FUTURE RESEARCH DIRECTIONS

From the survey, we conclude that malicious users and bot detection can be further improved by working on following areas.

A. PLATFORM INDEPENDENT CLASSIFIERS

Most of the literature focuses on detecting users and bots on Twitter platform, and most close-to-real datasets are normally available for Twitter platform. While Twitter is a popular platform, there are other platforms which are popular places of discussions for users around the world and so the spread of fake news is undiscovered and under researched on those platforms. A study shows that Facebook has surpassed all other social media platforms for users now with WhatsApp and YouTube leading after it. Therefore, bots may have different features based on platforms and platform dependent models will render it impossible to detect bots on other platforms which are gaining popularity now. A potential direction would be to create platform independent datasets that can be used in building detection models that caters to other platforms.

B. MULTIPLE TYPES OF BOT DETECTION

The research shows that many bot detectors fail to detect other types of bot as bot masters are constantly changing features of bots making them more difficult to be detected online. A good future direction would be to have a classifier that could detect multiple type of bots separately instead of one. One way to do it can be to design an unsupervised method to cluster similar bot accounts based on dataset automatically, and then assign homogeneous accounts to the specialized bot classifiers. [65]

C. MULTILINGUAL DETECTION

In the literature, we can see a distinct lack of models trained in languages other than English. This presents a good future direction of study. If a model was trained with numerous languages, then it could be generalized to other countries that have a different native language. The style of fake news and how it is written may differ country to country as well, so a dataset from a country speaking that language would be a good contribution, instead of translating existing datasets into languages, then it could be generalized to other countries.

[105] others have detected general fake news in English, Spanish and Portuguese only. [106]

8https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
The research also shows models which detect fake news bots after they have posted fake news on the internet. A future direction which researchers may choose, is a real time detection. This means that a social media platform can implement a real time model, which would flag users and posts when they try to post fake news. Even with existing system such as the one Varshney [107] describes, it is only described as a real time system with limited abilities to detect fake news spreaders in real time.

E. COLLECTION OF IMPLICIT FEATURES OF USERS

Implicit features are not directly shown from user profiles. Shu K. et al. [77] shows that implicit features perform better than explicit features. However, implicit features are hard to obtain because they are the inference of user behaviors. For example, the inference of the personality of a user is hard, specially when that user does not have too much information on his account. If people start inferring the implicit features, this may result in biasness. One future research direction would be finding a good way of extracting both explicit and implicit features from user accounts.

F. SOCIAL BOT WITH GOOD USES

Bots are used as a way of spreading fake news. They can heavily influence on human behavior by manipulating their emotions. For example, bots can create fake news with content which look true to invoke human fear or surprise of a fake fact or a group of people. Then, they would be more likely to share this news with others. This is how attackers are using bots to spread fake news. One future direction is that whether it is possible for bots to be used to encourage people or spread facts and positive emotions to make the society better. One direction of future research could be figuring out how can bots be used for positive utilities and conducting a study on cyborg coordination and communication.

VII. CONCLUSION

OSNs has nowadays become the most integral part of everyone’s life, and has become the major source of information for everyone. While it has lot of benefits, it has also shown some serious drawbacks in the form of spread of fake news, done to manipulate users minds and their decisions [108]. Both human and bots share fake news, and the bots can mimic human features very closely. There are numerous challenging issues which currently require further investigation such as differentiating a user account from automated accounts. In this survey, we have reviewed all the state of art methods used by researchers in detecting a malicious human user, and a bot based on source-based, propagator-based and target-based features of user accounts. Lastly, we have mentioned common challenges and future research directions from our survey which will help future researchers come up with a sophisticated classifier with better accuracy rate.

REFERENCES

[1] Y.-J. Lu and C.-T. Li, “Gcan: Graph-aware co-attention networks for explainable fake news detection on social media,” arXiv preprint arXiv:2004.11648, 2020.
[2] C. Grimme, M. Preuss, L. Adam, and H. Trautmann, “Social bots: Human-like by means of human control?,” Big data, vol. 5, no. 4, pp. 279–293, 2017.
[3] C. Snider, “Social media statistics,” Sep 2021. [Online]. Available: https://chrissniderdesign.com/blog/resources/social-media-statistics/
[4] Y. Roth and D. Harvey, “How twitter is fighting spam and malicious automation,” Twitter [blog], June, vol. 26, 2018.
[5] K.-C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, “Arming the public with artificial intelligence to combat social bots,” Human Behavior and Emerging Technologies, vol. 1, no. 1, pp. 48–61, 2019.
[6] H. Allcott and M. Gentzkow, “Social media and fake news in the 2016 election,” Journal of economic perspectives, vol. 31, no. 2, pp. 211–36, 2017.
[7] K. Shu and H. Liu, “Detecting fake news on social media,” Synthesis lectures on data mining and knowledge discovery, vol. 11, no. 3, pp. 1–129, 2019.
[8] R. R. Mourão and C. T. Robertson, “Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information,” Journalism Studies, vol. 20, no. 14, pp. 2077–2095, 2019.
[9] H. Rashkin, E. Choi, J. Y. Jang, S. Volkova, and Y. Choi, “Truth of varying shades: Analyzing language in fake news and political fact-checking,” in Proceedings of the 2017 conference on empirical methods in natural language processing, 2017, pp. 2931–2937.
[10] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, “Fake news detection on social media: A data mining perspective,” ACM SIGKDD explorations newsletter, vol. 19, no. 1, pp. 22–36, 2017.
[11] C. Shao, G. L. Ciampaglia, O. Varol, A. Flammini, and F. Menczer, “The spread of fake news by social bots,” arXiv preprint arXiv:1707.07592, vol. 96, p. 104, 2017.
[12] C. Wardle, H. Derakhshan et al., “Thinking about ‘information disorder’: formats of misinformation, disinformation, and mal-information,” Iteron, Chelshin; Posetti, Julie. Journalism, ‘fake news’ & disinformation. Paris: Unesco, pp. 43–54, 2018.
[13] M. Alldairi and A. Alwahedi, “Detecting fake news in social media networks,” Procedia Computer Science, vol. 141, pp. 215–222, 2018.
[14] O. Varol, E. Ferrara, C. Davis, F. Menczer, and A. Flammini, “Online human-bot interactions: Detection, estimation, and characterization,” in Proceedings of the international AAAI conference on web and social media, vol. 11, no. 1, 2017.
[15] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, “The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race,” in Proceedings of the 26th international conference on world wide web companion, 2017, pp. 963–972.
[16] M. Spradling, M. Allison, T. Tsogbadrakh, and J. Strong, “Toward limiting social botnet effectiveness while detection is performed: A probabilistic approach,” in 2019 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, 2019, pp. 1388–1391.
[17] S. Cresci, M. Petrocchi, A. Spognardi, and S. Tognazzini, “Better safe than sorry: An adversarial approach to improve social bot detection,” in Proceedings of the 10th ACM conference on Web Science, 2019, pp. 47–56.
[18] K. Warwick, “Cyborg morals, cyborg values, cyborg ethics,” Ethics and information technology, vol. 5, no. 3, pp. 131–137, 2003.
[19] S. Castillo, H. Allende-Cid, W. Palma, R. Alfaro, H. S. Ramos, C. Gonzalez, C. Elorregui, and P. Santander, “Detection of bots and cyborgs in twitter: A study on the chillean presidential election in 2017,” in International Conference on Human-Computer Interaction. Springer, 2019, pp. 311–323.
[20] T. Khaund, B. Kirdemir, N. Agarwal, H. Liu, and F. Mormorste, “Social bots and their coordination during online campaigns: A survey,” IEEE Transactions on Computational Social Systems, 2021.
[21] R. De Nicola, M. Petrocchi, and M. Praitelli, “On the efficacy of old facts and positive emotions to make the society better. One future direction is that whether it is possible for bots to be used to encourage people or spread facts and positive emotions to make the society better. One direction of future research could be figuring out how can bots be used for positive utilities and conducting a study on cyborg coordination and communication.

D. REAL TIME DETECTION

The research also shows models which detect fake news bots after they have posted fake news on the internet. A future direction which researchers may choose, is a real time detection. This means that a social media platform can implement a real time model, which would flag users and posts when they try to post fake news. Even with existing system such as the one Varshney [107] describes, it is only described as a real time system with limited abilities to detect fake news spreaders in real time.
Evolving and Adaptive Intelligent Systems (EAIS). IEEE, 2020, pp. 1–8.

[23] M. Orabi, D. Mouheb, Z. Al Aghbari, and I. Kamel, “Detection of bots in social media: A systematic review,” Information Processing & Management, vol. 57, no. 4, p. 102250, 2020.

[24] S. Zad, M. Heidari, H. James Jr, and O. Uzuner, “Emotion detection of textual data: An interdisciplinary survey,” in 2021 IEEE World AI IoT Congress (AIoIT). IEEE, 2021, pp. 0255–0261.

[25] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycooke, D. Rothschild et al., “The science of fake news,” Science, vol. 359, no. 6380, pp. 1094–1096, 2018.

[26] C.-C. Wang, “Fake news and related concepts: Definitions and recent research development,” Contemporary Management Research, vol. 16, no. 3, pp. 145–174, 2020.

[27] J. M. Burkhardt, “History of fake news,” Library Technology Reports, vol. 53, no. 8, pp. 5–9, 2017.

[28] S. B. Naeem, R. Bhatti, and A. Khan, “An exploration of how fake news is taking over social media and putting public health at risk,” Health Information & Libraries Journal, vol. 38, no. 2, pp. 143–149, 2021.

[29] J. Prier, “Commanding the trend: Social media as information warfare,” Strategic Studies Quarterly, vol. 11, no. 4, pp. 50–85, 2017.

[30] X. Zhang and A. A. Ghorbani, “An overview of online fake news: Characterization, detection, and discussion,” Information Processing & Management, vol. 57, no. 2, p. 102025, 2020.

[31] S. Hakak, W. Z. Khan, S. Bhattacharya, G. T. Reddy, and K.-K. R. Choo, “Propaganda of fake news on social media: Challenges and opportunities,” in CSOnet, 2020.

[32] T. Schuster, R. Schuster, D. J. Shah, and R. Barzilay, “Are we safe yet? the limitations of distributional features for fake news detection,” arXiv preprint arXiv:1908.09805, 2019.

[33] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” Commun. ACM, vol. 59, no. 7, p. 96–104, Jun. 2016. [Online]. Available: https://doi.org/10.1145/2818717

[34] R. Gorwa and D. Guilbeault, “Unpacking the social media bot: A typology to guide research and policy,” Policy & Internet, vol. 12, no. 2, pp. 225–248, 2020.

[35] R. Varão, “Patrícia campos mello (2020), the hate machine: notes from a reporter on fake news and digital violence [a máquina do ódio: notas de uma repórter sobre fake news e violência digital].” São paulo: Companhia das letras. language: Portuguese brazilian. isbn-10: 853593362x. pbk, 296 pages, 2021.

[36] N. M. Lee, “Fake news, phishing, and fraud: a call for research on digital media literacies education beyond the classroom,” Communication Education, vol. 67, no. 4, pp. 460–466, 2018.

[37] J. F. Wanda, B. S. Chipanjilo, G. Gondwe, and J. Kerunga, “Clickbait-style headlines and journalism credibility in sub-saharan africa: Exploring audience perceptions,” Journal of Media and Communication Studies, vol. 13, no. 2, pp. 50–56, 2021.

[38] A. Devitt and K. Ahmad, “Sentiment polarity identification in financial news: A cohesion-based approach,” in Proceedings of the 45th annual meeting of the association of computational linguistics, 2007, pp. 984–991.

[39] A. Olteanu, E. Kıcıman, and C. Castillo, “A critical review of online disinformation and opportunities,” Procedia Computer Science, vol. 121, pp. 817–825, 2017.

[40] S. Osoughi, D. Roy, and S. Aral, “The spread of true and false news online,” Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[41] M. Gupta, P. Zhao, and J. Han, “Evaluating event credibility on twitter,” in Proceedings of the 2012 SIAM International Conference on Data Mining. SIAM, 2012, pp. 153–156.

[42] C. Braker, S. Shiaeles, G. Bendiab, N. Savage, and K. Limniotis, “Botspot: Deep learning classification of bot accounts within twitter,” in Proceedings of the 29th ACM International Conference on Information & Knowledge Management. SIAM, 2020, pp. 2725–2732.
F. Ajesh, S. Aswathy, F. M. Philip, and V. Jeyakrishnan, “A hybrid method for fake news detection: a hybrid cnn-rnn based deep learning approach,” International Journal of Information Management Data Insights, vol. 1, no. 1, p. 100007, 2021.

O. D. Apuke and B. Omar, “Fake news and covid-19: modelling the predictors of fake news sharing among social media users,” Telematics and Informatics, vol. 56, p. 101475, 2021.

S. Martinez-Bea, S. Castillo-Pérez, and J. García-Alfaro, “Real-time malicious fast-flux detection using dns and bot related features,” in 2013 Eleventh Annual Conference on Privacy, Security and Trust. IEEE, 2013, pp. 369–372.

R. U. Rahman and D. S. Tomar, “New biostatistics features for detecting web bot activity on web applications,” Computers & Security, vol. 97, p. 102001, 2020.

M. Mendoza, M. Tesconi, and S. Cresci, “Bots in social and interaction networks: detection and impact estimation,” ACM Transactions on Information Systems (TOIS), vol. 39, no. 1, pp. 1–32, 2020.

S. Barhate, R. Mangla, D. Paniwani, S. Gatkal, and F. Kazi, “Twitter bot detection and their influence in hashtag manipulation,” in 2020 IEEE 17th India Council International Conference (INDICON). IEEE, 2020, pp. 1–7.

W. Antoun, F. Baly, R. Achour, A. Hussein, and H. Haji, “State of the art models for fake news detection tasks,” in 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT). IEEE, 2020, pp. 519–524.

S. Ahmed, “Who inadvertently shares deepfakes? analyzing the role of political interest, cognitive ability, and social network size,” Telematics and Informatics, vol. 57, p. 101508, 2021.

A. A. Ahmed, A. Aljahoub, P. K. Donepudi, and M. S. Choi, “Detecting fake news using machine learning: A systematic literature review,” arXiv preprint arXiv:2102.04458, 2021.

E. M. Mahir, S. Akhter, M. R. Huq et al., “Detecting fake news using machine learning and deep learning algorithms,” in 2019 7th International Conference on Smart Computing & Communications (ICSCC). IEEE, 2019, pp. 1–5.

P. Kiehniewicz, M. Choraś, R. Kozik, and M. Woźniak, “Machine learning methods for fake news classification,” in International Conference on Intelligent Data Engineering and Automated Learning. Springer, 2019, pp. 332–339.

C. Busioc, S. Ruseti, and M. Dascalu, “A literature review of nlp approaches to fake news detection and their applicability to romanian-language news analysis,” Revista Transilvania, no. 10, 2020.

A. Hamid, N. Shiekh, N. Said, K. Ahmad, A. Gul, L. Hassan, and A. Al-Fuqaha, “Fake news detection in social media using graph neural networks and nlp techniques: A covid-19 use-case,” arXiv preprint arXiv:2012.07517, 2020.

S. Deepak and B. Chitturi, “Deep neural approach to fake-news identification,” Procedia Computer Science, vol. 167, pp. 2236–2243, 2020.

S. Qi, L. AliKulaib, and D. A. Broniatowski, “Detecting and characterizing bot-like behavior on twitter,” in International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation. Springer, 2018, pp. 228–232.

K. Lee, B. D. Eoff, and J. Caverlee, “Seven months with the devils: A long-term study of content polluters on twitter,” in Fifth international AAAI conference on weblogs and social media, 2011.

G. Guibon, L. Ermakova, H. Seffih, A. Firsov, and G. Le Noé-Bienvenu, “Multilingual fake news detection with satire,” in CICLing: International Conference on Computational Linguistics and Intelligent Text Processing, 2019.

H. Q. Abonizio, J. I. de Morais, G. M. Tavares, and S. Barbon Junior, “Language-independent fake news detection: English, portuguese, and spanish mutual features,” Future Internet, vol. 12, no. 5, p. 87, 2020.

D. Varshney and D. K. Vishwakarma, “Hoax news-inspector: a real-time prediction of fake news using content resemblance over web search results for authenticating the credibility of news articles,” Journal of Ambient Intelligence and Humanized Computing, pp. 1–14, 2020.

S. R. Sahoo and B. B. Gupta, “Multiple features based approach for automatic fake news detection on social networks using deep learning,” Applied Soft Computing, vol. 100, p. 106983, 2021.
