Fake Profile Detection on Social Networking Websites: A Comprehensive Review

Pradeep Kumar Roy and Shivam Chahar

***Abstract*—This article aims to summarize the recent advancement in the fake account detection methodology on social networking websites. Over the past decade, social networking websites have received huge attention from users all around the world. As a result, popular websites such as Facebook, Twitter, LinkedIn, Instagram, and others saw an unexpected rise in registered users. However, researchers claim that all registered accounts are not real; many of them are fake and created for specific purposes. The primary purpose of fake accounts is to spread spam content, rumor, and other unauthentic messages on the platform. Hence, it is needed to filter out the fake accounts, but it has many challenges. In the past few years, researchers applied many advanced technologies to identify fake accounts. In the survey presented in this article, we summarize the recent development of fake account detection technologies. We discuss the challenges and limitations of the existing models in brief. The survey may help future researchers to identify the gaps in the current literature and develop a generalized framework for fake profile detection on social networking websites.**

***Impact Statement*—Social networking platforms have become an essential part of today's human life—almost every individual is associated with at least one of the online social networking websites today. Hence, a huge crowd is always active on these platforms; a large number of user engagements attracted spammers and unauthentic users on online social networking. To spread unauthentic messages such as rumors, hate speech, bullied text, and others, users create a fake profile. Researchers proposed several techniques to limit this issue using machine-learning- and deep-learning-based models, but many fake accounts are still present. However, for a good social networking platform, these fake accounts are not acceptable. This article summarizes the recent advancement of social networking's fake account detection, which helps the future researcher build a robust model to prevent and identify fake accounts on online social networking.**

***Index Terms*—Deep learning (DL), fake profile, machine learning (ML), online social network, spammer, twitter.**

I. Introduction

IN RECENT years, with the growth of technologies and easy Internet connections, online social networking (OSN) websites have attracted huge crowds worldwide. Facebook (FB), Twitter, Instagram, YouTube, and LinkedIn are a few famous OSN websites with a large number of registered users from every corner of the world [1]–[5]. These websites are free to join and share personal or career interests and activities with their connections without paying any cost. They can exchange messages, photos, videos, diary entries online (blogs), etc., with their contacts as well [6], [7]. Today, social media as a medium is used to communicate with people and as a vital resource for business and marketing.[1] OSN websites are interactive computer-mediated technologies that manage the creation and sharing of information, ideas, career interests, and other forms of expression via virtual communities and networks [8], [9]. FB attracts 2.13 billion monthly active users and an average of 1.4 billion daily active users around the world in 2017.[2] Current statics revealed that FB has 260 million registered users from India, which is the largest number, followed by 180 million registered users from America.[3] Another OSN website Twitter has 330 million monthly active users and 145 million daily users.[4] The statistics of worldwide users on two popular OSN websites, such as Twitter and FB, are shown in Fig. 1. The same crowds are present on YouTube and Instagram also.

Helping people of every age group and community to stay in contact with old and new friends is one of the significant benefits of OSN websites. Making new connections with the people of the same interests around the globe is also made very easy by OSN websites. These capabilities and the many other ways to communicate with friends make social networking sites very appealing. For these users, their social life, even their practical life, has become interrelated [10]–[12]. On the other hand, OSN websites face many issues, such as security, privacy, spamming, rumor, and fake profiles. Due to no restriction over how many and who can create a profile on OSN websites, it provided a huge opportunity for unethical users to create fake profiles and misuse the platform for their personal/organizational benefit. The popular OSN websites such as Twitter and FB continuously look at these issues and delete the fake profiles from time to time.[5]

Manuscript received September 22, 2020; revised November 30, 2020, January 15, 2021, and March 3, 2021; accepted March 5, 2021. Date of publication March 9, 2021; date of current version May 25, 2021. *(Corresponding author: Pradeep Kumar Roy.)*

Pradeep Kumar Roy is with the Vellore Institute of Technology, Vellore 632014, India, and also with the Indian Institute of Information Technology, Surat 395007, India (e-mail: pkroynitp@gmail.com).

Shivam Chahar is with the Vellore Institute of Technology, Vellore 632014, India (e-mail: shivam.chahar@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TAI.2021.3057027.

Digital Object Identifier 10.1109/TAI.2021.3064901

[1][Online]. Available: https://www.rightmixmarketing.com/onlinemarketing/fake-profile-affects-online-marketing/

[2][Online]. Available: (https://about.fb.com/company-info/) [Retrieved on April 12, 2018]

[3][Online]. Available: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ [Retrieved on March 8, 2020]

[4][Online]. Available: https://www.brandwatch.com/blog/twitter-stats-and-statistics/ [Retrieved on April 02, 2020]

[5][Online]. Available: https://techcrunch.com/2020/02/03/twitter-suspends-large-network-of-fake-accounts-used-to-match-phone-numbers-to-users/

2691-4581 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

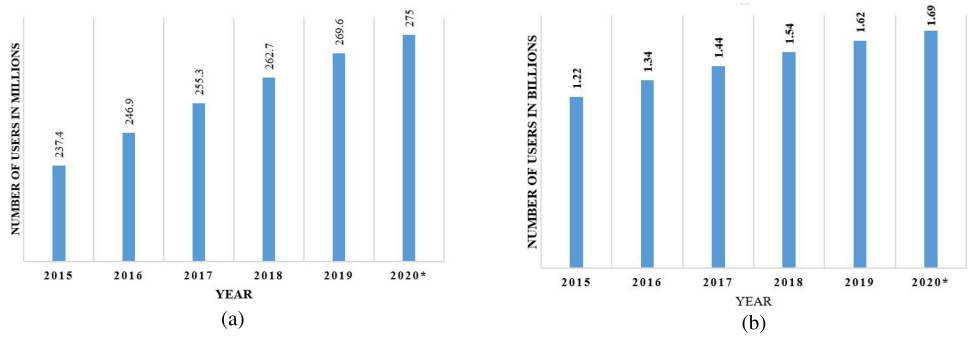

Fig. 1. Statistics of the registered users on Twitter and FB OSN platform. (a) Twitter. (b) FB.

The fake profile issue is present on all popular OSN websites such as FB having 1.3 billion fake profiles. Fake profiles are generally used to post spam messages, rumors, promotional posts, and others to make money illegally [13]. Due to restrictions on the length of the message on Twitter, spammers attach the external malicious link in the post to complete it, which is sometimes being dangerous; the attached link may have some virus or other malicious program, which stole the user's information [14]. The spam message spreading using Twitter is 0.13%, whereas 0.0003–0.0006% spam messages spread using the emails [15]. From the above analysis, it is confirmed that among all OSN websites, Twitter is the most favorite website for the spammer to spread unethical, untruthful, rumors, and spam messages [16].

The unethical, untruthful, rumor, and spam messages generally posted using a fake account/profile. There were many pieces of research that have been done to detect the spam messages [17]–[24]. But the detection of fake profiles on OSN websites is still a big issue. Fake profiles are created by stealing the data from other existing profiles on the network, which is easily available such as profile name, profile photo, age, sex, and others [25]. This results in exposing incorrect information to their friends and contacts, which are connected to them through social media. This situation can result in huge damage in the real world, including citizens, business entities, and others. Fake accounts/profiles can present fake news, fake web rating, and spam. OSN operators currently spend determined resources to detect, physically confirm, and close fake profiles [1], [8], [16], [26]–[38].

This survey aims to present up-to-date research work on fake account detection. To do this, quality research works are collected from different digital libraries, including IEEE Xplore,[6] Springerlink,[7] Sciencediret,[8] and ACM.[9] We have also included the papers presented in top-tier conferences such as the International Conference on Machine Learning, the International Conference on Data Mining, and Neural Information Processing Systems to name a few. This survey covers an explanation of the existing methodology and used feature sets in subsequent sections. This will help new researchers identify the limitations in the existing work and provide new models that will detect fake accounts with higher accuracy efficiently.

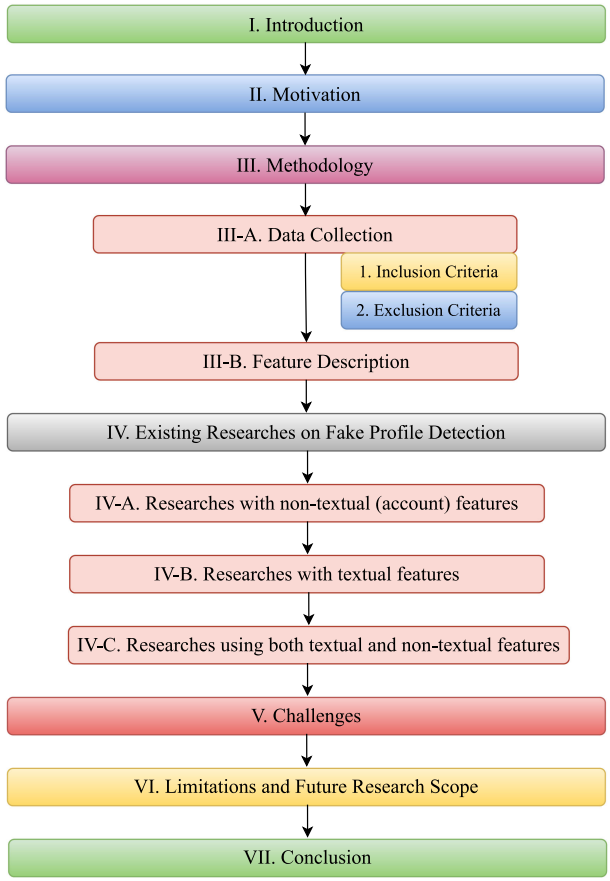

Fig. 2. Organization of the survey.

The researchers used traditional machine learning (ML)-based models, but the benefits of deep neural-network (NN)-based models have attracted more. The deep NN models are capable of handling complex data as well, giving a promising performance on many other domains of research such as text classification [24], [39], image processing [40]–[42], video captioning [43]–[46], and others. The complete road map of this survey is shown in Fig. 2.

[6][Online]. Available: https://ieeexplore.ieee.org/Xplore/home.jsp
[7][Online]. Available: https://link.springer.com/
[8][Online]. Available: https://www.sciencedirect.com/
[9][Online]. Available: https://dl.acm.org/

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

This survey categorized existing research into three categories: 1) research using nontextual (account-based) features; 2) research using textual features; and 3) research using both the nontextual and textual features. More significantly, to the best of our knowledge, this is the first comprehensive survey on fake account detection; previous surveys [33], [47]–[57] mainly discussed various networking attacks, and other ones did not cover many insights of the OSN websites. Camacho *et al.* [58] published a review article that summarizes four dimensions of social media: pattern and knowledge discovery, scalability, information fusion and integration, and visualization in brief. Each dimension is well studied and addressed. Our article is different in such a way that we focused on the contributions of ML and deep learning (DL) to handle one of the above-mentioned dimensions in brief. Existing surveys [48], [50], [51] covered very few research works for the said problem and left many important pieces of research untouched. Others are also not focused on the importance of detection techniques such as ML and DL. Our main contributions are as follows.

1) We collected and organized the research on the fake profile detection in three modules.
2) We highlighted the main issues, current status, and future research scope for each module in brief.
3) We provided the feature's information, metrics, and datasets used by researchers.

II. Motivation

The number of online social networking website users is increasing with time. The number of worldwide users active on social media as of April 2020 is 3.81 billion, and among them, 3.76 (98.687% of total) billion users are actively using through the mobile phone.[10] As shown in Fig. 1(a), currently, Twitter has 275 million users, which is 2% higher than the number of registered users (267.4) in 2019. Similarly, FB has 1.69 billion users, which is 1.04% higher than the number of registered users in 2019. Twitter, FB, Instagram, LinkedIn, Tumblr, and other social networking websites receive similar worldwide users' responses. One of the issues of these social networking websites is that they allow users to create multiple accounts. For example, if you have an account on FB, Twitter, or a similar platform, you can easily create another profile. Hence, it can easily find many duplicate profiles on these social networking websites.

Currently, social networking websites do not provide any notification to their users about profile authenticity. Hence, it is tough for naive users to differentiate the fake and real profiles. Moreover, many threats, such as cloning of profile information and monitoring of the user's activity and others, also increase—privacy of the users' data being a sensitive issue as it introduces many cybercrimes. In the past few years, researchers have developed many models to address these issues, but, still, the problem remains open. As per our knowledge, there is no up-to-date review article that summarizes the existing research on fake account detection. In 2018, Ramalingam and Chinnaiah [33] published a review article that mainly focused on different networking attacks; another survey has been recently published by Joshi *et al.* [57], which has not covered many insights into social networking websites. This motivates us to develop an up-to-date survey aiming to discuss the fake-account-detection-related research only with the used features in detail.

[10][Online]. Available: https://www.statista.com/statistics/617136/digital-population-worldwide /

III. Method and Materials

This section highlights the working principle of ML- and NN-based models that were frequently used by the researcher to identify the fake profiles on social networking platforms.

A. Data Collection

There are many online–offline platforms available today, which are publishing the research article in online and offline modes. For this study, we have collected articles from online digital libraries such as IEEE Xplore, Springerlink, Sciencedirect, ACM digital library, Scopus, and Web of Science database. The key phrases "Fake Profile," "Fake Identity," "Spammer," "Unauthorized Users," and similar terms are used to search the articles in these databases. While collecting the data from these sources, few duplicate research articles were collected, which were removed during the articles' categorization. To limit the amount of quality research among the searched articles, we have adopted the inclusion–exclusion criteria, which are defined in the subsequent subsections.

1) Inclusion Criteria:

1) The articles published in reputed journals.
2) The article having the proper explanation of the proposed methodology with experimental proof.
3) The articles published in top-tier conferences.
4) The article having the main focus on social networking websites.
5) The article having a performance comparison with the existing models.
6) The articles using either ML models such as random forest (RF), decision tree (DT), support vector machine (SVM), logistic regression (LR), gradient boosting, naive Bayes (NB), and other or DL models such as convolutional neural network (CNN), recurrent NN, long short-term memory (LSTM), and similar ones.

2) Exclusion Criteria:

1) The articles published in the conferences having no experimental proof.
2) The articles having no comparison with baseline models.
3) The articles that were not focused on fake account detection.
4) The articles that used a non-English dataset for the research.

B. Feature Description

To detect the fake profiles on Twitter, FB, and other OSN websites, various features related to the user's account and their posts are used (Table I). The features are defined as follows.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

TABLE I
FREQUENTLY USED FEATURES FOR FAKE ACCOUNT DETECTION

Feature Types	Features
Account	Username [59], [60], [8], [61], [6], [62], [63], [64], [65]
	Biography [66], [67], [68], [59], [60], [8], [61], [6], [29], [64], [69], [70], [71], [65]
	Profile Photo [68], [59], [72], [73], [60], [8], [74], [6], [61], [75], [64], [76], [63], [65]
	Header Photo [73], [6], [29], [76]
	Homepage [59], [6]
	Theme Color [77], [76], [65]
	Birth date[78], [71]
	Location [59], [60], [8], [6], [62], [64], [63], [71], [70], [65]
	Creation [67], [8], [61], [74], [64], [63], [70], [65]
	Number of tweets [66], [6], [75], [69], [79], [80], [71], [65]
	Number of followers [66], [82], [59], [72], [60], [8], [61], [74], [6], [75], [62], [29], [79], [69], [64], [76], [71], [70], [65]
	Following count [66], [60], [75], [69], [81], [71], [65]
	Number of likes [82], [83], [79]
	Listed count [84], [59], [74], [6], [29], [69], [64], [76], [70]
Textual	Sender [78], [85], [29]
	Mentions [67], [82], [84], [74], [29], [69], [80], [71]
	Hashtags [66], [67], [84], [59], [8], [74], [6], [86], [29], [69], [71], [70]
	Link [66], [67], [78], [85], [84], [59], [87], [88], [74], [86], [89], [29], [62], [69], [80], [79], [71]
	Number of retweets [84], [59], [69], [80], [71], [70]
	Number of replies [70]
	Sent date[85], [83], [75], [90], [81]
	Location [85], [6]

1) *Username [6], [8], [59]–[65]:* The username is the unique identifier of the user's account. Hence, on the social networking website, each user has a unique name.
2) *Biography [6], [8], [29], [59]–[61], [64]–[71]:* Social networking platforms allow their user to add a short introduction; it is termed as user's biography. The users prefer to write about their achievements, expertise, and other important information in the biography.
3) *Profile Photo [6], [8], [59]–[61], [63]–[65], [68], [72]–[76]:* One of the main components of every social networking website is a profile photo of the users' account. The profile visitor quickly recognizes the person by seeing the photo and sometimes getting trapped by believing that it is real.
4) *Header Photo [6], [29], [73], [76]:* Apart from the profile photo, the websites such as FB and Twitter also allow their users to add the header photo to make the profile more attractive.
5) *Theme Color [65], [76], [77]:* To make the profile view more attractive, users have the option to choose their favorite theme color for the account.
6) *Birthdate [71], [78]:* Social networking websites ask their users to add the date of birth with an option to publish the birth year publicly or keep it private. The birth date reminds them of their birthday every year, invoking a personal connection.
7) *Homepage [6], [59]:* Homepage of the user profile.
8) *Location [6], [8], [59], [60], [62]–[65], [70], [71]:* At the time of profile creation on any social networking website, users have to provide their location, but there is no authenticity check about the real and provided location information. Hence, sometimes, it misleads.
9) *Creation [8], [61], [63]–[65], [67], [70], [74]:* The date when the users created their account on the OSN platform is termed as the creation date. Researchers used the information of the profile creation date also for fake profile detection.
10) *Number of Tweets [6], [65], [66], [69], [71], [75], [79], [80]:* The users' posts are termed as tweets on Twitter, whereas post on FB. How many tweets or posts are published by the users on OSN is an essential feature for fake account detection. Generally, fake users posted a large number of posts in a short time.
11) *Following Count [60], [65], [66], [69], [71], [75], [81]:* It is defined as the number of existing profiles on Twitter that are following the user's profile.
12) *Number of Followers [6], [8], [29], [59]–[62], [64]–[66], [69]–[72], [74]–[76], [79], [82]:* It is defined as the number of other existing profiles who are following the user.
13) *Number of Likes [79], [82], [83]:* If a profile visitor likes the profile content or their achievement, they can like their profile. The total number of likes in an account is also an important feature.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

14) *Number of Retweets [59], [69]–[71], [80], [84]:* If a user posted any post or tweets, their connections could be shared either on the same or different social networking platforms. The number of times a post is shared will be counted and treated as a feature to achieve the objective.
15) *Listed Count [6], [29], [59], [64], [69], [70], [74], [76], [84]:* The number of public lists of which the account owner is a member.
16) *Sender [29], [78], [85]:* The author of the message creator.
17) *Mentions [29], [67], [69], [71], [74], [80], [82], [84]:* The names or brands that are mentioned in the post.
18) *Hashtags [6], [8], [29], [59], [66], [67], [69]–[71], [74], [84], [86]:* It is used to describe the topic of the message posted on Twitter.
19) *Link [29], [59], [62], [66], [67], [69], [71], [74], [78]–[80], [84]–[89]:* The URL mentioned in the post.
20) *Number of Retweets [59], [69]–[71], [80], [84]:* The number of times a message is shared on Twitter.
21) *Number of Replies [70]:* The responses received on a post.
22) *Sent Date [75], [81], [83], [85], [90]:* Date at which the tweet was sent.
23) *Location [6], [85]:* Location such as city name, country, or similar information from where the tweet was originated.

The necessary information on the widely used classification models with their limitations may be obtained by reading the original paper. Additionally, the future researcher may also follow the survey on text classification written by Kowsari *et al.* [91] to know the working principle of various classification techniques.

IV. Existing Research on Fake Profile Detection

We categorized the fake profile detection research into three categories: 1) research with account-based features; 2) research with text-based features; and 3) research with both account- and text-based features. In the subsequent sections, we highlight the key finding, limitations, and future scope of the proposed models.

A. Research With Account-Based Features

Ahmad and Abulaish [92] suggested a Markov clustering (MCL)-based approach for spam profile detection in OSN websites. They used the FB dataset consisting of 320 user profiles, out of which 165 profiles were of spam category, and 155 profiles were normal. They extracted mainly three features from the user's profile information: 1) active friend: it captures how frequently a user on FB is communicating with their connection; 2) page-like: it captures how often users like the shared page; and 3) URL: it captures how often users share the URL. Their analysis confirmed that the fake/spammer account shares the URL more frequently than the authentic user. The social graph was created by using users and their connections to exploit the behavior similarity of profiles and unearthed the clusters present in the profile dataset.

Conti *et al.* [78] developed a model called FakeBook to detect the fake profile on OSN websites. They said that the current research on OSN websites mainly focused on protecting the privacy of an existing online profile. However, there was more risk if a person does not have their account on the fancy social networking website. The risk is that someone may create a fake profile on behalf of the real user. The fake profile may use to build a relationship with their trusties by aiming to get more personal information. To identify such fake profiles, they collected FB user's data and developed the FakeBook model. They did the temporal evaluation in real time using the characteristics received from the real user account.

Fire *et al.* [97] developed software for FB called Social Privacy Protector (SPP) to detect fake profiles and improve the security and privacy of OSN websites. The SPP works in three layers: the first layer identifies the friends who may pose a threat, the second layer monitors FB's different privacy settings, and the third layer informs users about the installed application on FB and has access to the user's private information. They used three datasets: 1) fake-profile dataset: 141 146 links; 2) friends-restriction dataset: 144 431 links; and 3) the all-link dataset: 147 291 links. The best classification result was obtained using the RF classifier, where the precision values of the friends-restriction, fake accounts, and all-link datasets are 0.98, 0.93, and 0.90, respectively.

Adikari and Dutta [68] developed a model to identify the fake accounts on LinkedIn. They said that to differentiate the fake and real profiles on OSN websites, account-based features are required. However, the restriction of private information access makes the task more complicated. They experimented with a small set of data having a total of 74 samples, in which 40 samples were legitimate and 34 samples were fake. They split the dataset into a 1:1 ratio for training and testing. The features were extracted using the principal component analysis and fed into classifiers to classify legitimate and fake profiles. The SVM classifier achieved the highest accuracy of 84.04%. Another model was developed by Kontaxis *et al.* [103] to identify the cloned profile on LinkedIn. They used 1120 LinkedIn public profiles for the research and designed a prototype that can be employed by the user. The profile hunter component returned at least one clone for 7.5% of the user profile.

Xiao *et al.* [87] also used LinkedIn social networking data and developed a model to cluster the users into two groups, i.e., legitimate and fake users in real time. They said that the existing models mainly identified the fake users based on their social activity, and hence, fake profiles allowed them to stay on the network and perform the activity. However, the need is to identify and block the fake profiles before they start the communication. To address this issue, they collected 260 644 labeled accounts, of which 153 019 were fake profiles and 107 625 were legitimate. They used three binary classifiers, i.e., RF, SVM, and LR, and found that the RF classifier performed best, whereas LR gave the worst performance.

Gurajala *et al.* [8] proposed a model using a pattern-matching algorithm for fake profile detection. They collected 62 million user profiles using the social web-crawling technique. In the filtering process, 724 494 groups containing a total of 6 958 523

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

accounts were generated. For further refining of 724 494 groups, their screen names were analyzed, and near-identical ones were identified. Among all, a highly reliable subset of fake user accounts was identified by using the map-reducing techniques and the pattern recognition approach.

In 2016, Gurajala *et al.* [61] investigated the profile characteristics of the fake profiles on Twitter. They clustered user's profiles into fake and real accounts based on the profile name and other attributes. Their algorithm successfully predicted a subset of users as fake, and with manual inspection, all predicted fake profiles are verified. Their analysis revealed that fake accounts are created in a bunch on weekdays on selected dates and times by following the time interval of lesser than 40 s. On a 2016-user Twitter database, the model achieved 93% of the accuracy value for detecting the fake profiles.

BalaAnand *et al.* [93] identified fake users on the social networking website using the user's nonverbal behavior. The user's nonverbal behavior data on social networking websites help detect multiple accounts and fake identity deception. They used Wikipedia's publicly available logs of blocked users' information as a dataset and applied the traditional ML-based classifiers such as SVM, RF, and adaptive boosting algorithms with the cross-validation technique. Using the adaptive boosting classifier, they achieved the best performance accuracy.

Cresci *et al.* [60] suggested a model for detecting fake Twitter followers. They used multiple datasets, such as The fake project, # elezioni2013, Fast followerz, Inter Twitter, and Twitter technology having a sample of 469, 1481, 1169, 1337, and 845, respectively. Furthermore, the dataset was categorized into two baseline datasets: baseline human dataset (1950 accounts) and baseline fake dataset (1950 accounts). Set of ML algorithms is employed for fake and spam accounts detection. With the SVM classifier, more than 95% of accounts were accurately classified into respective classes.

Boshmaf *et al.* [73] developed a model for fake account detection in OSN websites. Two datasets were used: the first dataset contained 8888 public user profiles of FB and the second one contained 60 000 real users, collected from Tuenti's production servers. They used user-level activities and graph models with weights. The lower weights toward the node were indications of the victim. Alowibdi *et al.* [77] collected 194 292 user profiles from Twitter during January and February 2014. On the user's profile information such as first name, username, profile layout color, and others, the clustering algorithms were applied and categorized the fake account with 90% of accuracy. The model was not tested with other platforms, which may include a limitation of the article.

Xuan *et al.* [95] developed a model that detected the malicious accounts on location-based social networks (LBSN) like Twitter, FB, etc. They collected data from February 15, 2015, to May 27, 2015, from a popular LBSN in China called Dianping. The dataset contained 398 malicious accounts and 3997 legitimate accounts. To make the dataset balanced, they have randomly picked 398 legitimate user accounts from the collected 3997 legitimate accounts. They applied many ML-based classifiers on the balanced dataset, including NB, DT, RF, LR, and SVM, and achieved the best F1-score value of 0.89 with the SVM classifier.

Meligy *et al.* [94] proposed a fake profile recognizer model to verify the identity and detect the fake profiles on social networking websites. They collected datasets from three platforms: Twitter, FB, and Google+. The dataset contains a total of 4000 (Twitter), 4039 (FB), and 3045 (Google+) samples. They used regular expression (RE) and deterministic finite automaton (DFA) approaches for the detection. The RE was used for both the authentication and representation of profile identities into a set format, whereas DFA was used to identify the trusted profiles. Their model achieved the F1-Score on Twitter, FB, and Google+ datasets as 0.818, 0.897, and 0.769, respectively, for the best case.

Jia *et al.* [96] used the random walk technique to identify the fake account on OSN websites named as SybilWalk. They used four social network data for their research: 1) FB (4039 nodes and 88 234 edges); 2) Enron (33 696 nodes and 180 811 edges); 3) Epinions (75 877 nodes and 811 478 edges); and 4) Twitter (41 652 230 nodes and 1 202 513 046 edges). In the dataset, the users were termed as nodes, and their mutual relations were termed as edges. On the Twitter dataset, SybilWalk achieved the best performance with a false positive rate of 1.3% and a false negative rate of 17.3%. Sandy *et al.* [102] investigated human capabilities for judgment of fake profiles on the social network. The authors selected the profile pictures from the face database created at the University of Michigan in 2004. They used a human-based approach for the experiment.

Khaled *et al.* [76] proposed an NN-based algorithm to identify fake profiles and bots on Twitter. Feature selection and dimensionality reduction techniques were applied. The MIB dataset was used to perform the study [60]. For feature reduction, many models were used. They found that SVM and NN model's combination yielded the best accuracy of 98.3% with the features obtained using the Spearman's rank-order correlation technique.

Caruccio *et al.* [64] proposed a novel technique to discriminate real accounts on social networks from fake ones. They collected data from 9019 Twitter accounts. An algorithm was used to discover relaxed functional dependencies (RFDs) from data. The extracted RFDs were used to differentiate fake, real, and verified accounts. Singh *et al.* [63] used supervised ML models to detect fake profiles in OSN websites. The proposed mechanism worked in the following manner: First, the data were gathered and cleaned and fictitious accounts were created; next, data (psychology, stats) were validated with the user's social status and then fictitious accounts were injected. To distinguish the fake and real profiles, they used an average number of followers of the users. They found that if a user's profile has more than 30 followers, it is not a fake profile. They also found that the average ages of the fake profiles' owners lie between 18 and 19 years and the profile images are taken from an Internet source. Apart from this, fake profiles' location was also not valid; such users use random locations to become untraceable. The authors did not provide any experimental results of their research; the effectiveness of the extracted features to trace the

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

fake profiles on the social network can be implemented and evaluated.

Revathi and Suriakala [100] used profile similarity communication matching approaches for the detection of duplicate profiles on OSN websites. Datasets were gathered from various pages from a single group to accomplish find cloned profile users in the dataset. They achieved the best accuracy of 93.87% using node similarity communication matching (NSCM). Suriakala and Revathi [101] used data mining techniques to identify cloned social network profiles. The dataset for the research was collected from the GitHub repository. They used NSCM, network description logic rule generation technique, RF, SVM, and privacy-protected methods to detect fake accounts. Among all, the privacy-protected system model yielded the best accuracy value of 97.3%.

Rahman *et al.* [65] used the *fakeproject* dataset to detect the fake Twitter profile. The dataset contained 6825 numbers of instances. A list of classifiers was used; the experimental results confirmed that the RF classifier performed the best and achieved an accuracy value of 99.20%. The limitation of their approach was that the dataset contained 6825 user accounts, but, currently, Twitter has around 326 million active users. So, the achieved results may vary in real time. Agarwal *et al.* [3] developed a model to detect fake accounts by analyzing users' emotions on FB using a supervised model. Emotion-based features such as *anger*, *sadness*, *fear*, *joy*, *trust*, *anticipation*, *positive_frac*, *negative_frac*, and others were considered for the research. Their analysis confirmed that fake profile users mainly used emotions like *hate*, *kill*, *ugly*, etc. They applied many ML-based classifiers on selected features and obtained the best accuracy value of 90.91% with RF classifiers.

Zarei *et al.* [98] gave a model on how to detect politician impersonating accounts in social media. The dataset of three politicians was collected from Instagram, which included Donald Trump (@realdonaldtrump), Barack Obama (@barackobama), and Emmanuel Macron (@emmanuelmacron). Data collected in the three months included 80 Instagram posts, 9 million likes, 350 000 comments, and 1.5 million profiles. Using these data, their model was able to distinguish crowds of impersonators and political bots. The authors reported that this was the first paper that conducts such analysis on Instagram Data. They extend the work in [104] and collected data from three separate types of profiles: politician, news agency, and sports star. A total of 550 posts, 1.3 million views, 20 million likes, and 6 million profiles have all been crawled and stored in JSON format. They used the term frequency-inverse document frequency (Tf-Idf) technique to detect the profiles that had similar profile information, while the CNN model was used to suit the profile images. The other characteristics were analyzed, such as the number of comments and likes to get a post based on age. Finally, by using the clustering methods, the profiles were clustered to observe obscure behavior and unusual profile characteristics.

Wanda and Jie [16] proposed a DL model called *DeepProfile* to detect fake accounts on OSN websites. They modified the pooling layer of the CNN, which was a novelty in their work. On the OSN dataset, their model achieved the precision, recall, and F1-score values of 94.00, 93.21, and 93.42, respectively. The ROC values were ranging from 0.950 to 0.959, and the model loss was 0.214. Pourghomi *et al.* [105] developed a model to detect the bots on Twitter using ML-based classifiers. They extracted various features and grouped them into six types: users, friends metadata, network pattern, activity, tweet content, and sentiment. With these features, the developed system achieved the ROC values of 0.950. However, the miss-classification rate was also high, which requires the re-analysis of the selected features. Another model was proposed by Pourghomi *et al.* [106] to identify fake accounts on FB. Their analysis answered the following: 1) the effectiveness of the FB fake account detection algorithms; 2) FB's artificial intelligence learning capabilities to differentiate fake and real accounts; and 3) ethical impacts of FB's policy changes. Their study provided a deep analysis of the various security and privacy policies and suggested an effective way to detect possible fake accounts. The key researches using account based features are shown in Table II.

B. Research With Text-Based Features

This section highlights the research that was done using the textual features of the OSN websites to detect fake, malicious, or spam accounts. Swe and Myo [80] proposed a model to detect the fake account of OSN websites using a blacklist instead of a traditional spam words list. The blacklist is created by using the topic modeling approach and a keyword extraction approach. The evaluation is done on the 1KS-10KN dataset and also on the Social Honeypot dataset. The traditional spam-word-list-based approach on the 1KS-10KN dataset achieved the precision value of 0.854, recall of 0.904, and F-measure of 0.879, whereas using the proposed methods, the precision, recall, and F1-measure were 0.958, 0.950, and 0.954, respectively. The fake account detection rate using the proposed model was 95.4%. The false positive rate of the spam-word-list-based approach is 0.154, and the false positive rate of the blacklist-based approach using the social honeypot dataset is 94.9%. The detection rate of the spam-word-list-based approach using the social-honeypot-based approach is 91.1%. The blacklist-based approach achieves acceptable accuracy and reduces the false positive rate. Their model does not require profile- and network-based features, and hence, it reduces the time and cost overhead for extracting these features.

Clark *et al.* [88] used natural language processing (NLP) to detect automation on Twitter. Their model uses natural language text from humans to provide a criterion for identifying accounts with automated messages. Two datasets were collected: first, geo-tweets from the most active 1000 users, referred to as the Geo-Tweet dataset to classify humans and robots. The second collection of data was obtained from the social honeypot experiment. They found that the accuracy of the model on the Geo-Tweet dataset increased by increasing the tweets bin's size. The model achieved a true positive rate of 86%, which means that most robots have been identified successfully. The model uses textual data on its own for prediction purposes, so it is flexible and can be applied to any text data beyond the Twitter sphere.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

TABLE II
KEY RESEARCH USING ACCOUNT-BASED FEATURES

Source	Dataset	Problem	Method	Result	OSNs covered
[3]	FB	Fake user detection	SVM, NB, JRip and RF	Accuracy: 90.91%	Only FB
[8]	Twitter	Fake Twitter account	Activity-based		Only Twitter
[60]	Multiple datasets[11]	Fake Twitter followers	ML algorithms	Accuracy: 95%	Multiple
[61]	Twitter	Fake Twitter accounts	Activity-based	Accuracy: 93%	Only Twitter
[64]	Twitter	Fake account detection	RFDs	–	Only Twitter
[65]	Fake Project	Detection of fake identities	ML algorithms and NN	Accuracy: 99.20%	Twitter and others
[68]	LinkedIn	Identifying fake profiles	NN and SVM	Accuracy: 84.04%	Only LinkedIn
[73]	FB and Tuenti	Fake account detection	–	AUC: 1.0	FB and Tuenti
[76]	Twitter and MIB	Detecting fake profiles	SVM, NN	Accuracy: 98%	Twitter and MIB
[77]	Twitter	Deception detection	K-means	Accuracy: 90%	Only Twitter
[78]	FB	Detecting fake profiles	Adversarial Model	Privacy analysis	Only FB
[92]	FB	Spam profile detection	MCL	FP: 0.88	Only FB
[93]	Wikipedia	Identifying fake users	ML algorithms	–	Wikipedia only
[94]	Multiple dataset[12]	Fake account detection	ML algorithms	F-measure: 0.83	Multiple
[95]	Dianping	Malicious account detection	ML	F1-Score: 0.89	Dianping
[96]	Twitter	Fake account detection	Random Walk	AUC: 0.96	Only Twitter
[97]	Three datasets	Fake account identification	SPP with ML	F1: 0.896	Not specified
[98]	Instagram	Detecting politician fake accounts	Similarity measure	Analysis based	Only Instagram
[99]	OSN	Detection of profile cloning	Similarity index	Similarity measure	Not specified
[100]	FB	Detection of duplicate profiles	KNN, SVM	Accuracy: 93.87%	Only FB
[101]	GitHub	Cloning Profile Detection	ML	Accuracy: 97.30%	GitHub only
[102]	Manually created	Fake profile detection	Statistical analysis	Similarity measure	Manual dataset

Khan *et al.* [86] segregated spammers and bloggers from real experts on Twitter. Approximately 0.4 million tweets were collected from approximately 3200 user profiles active in disseminating health-related information on Twitter. They used the Hyperlink-Induced Topic Search (HITS) approach to classify spammers and bloggers and isolated them from experts in a specific area. Their model does not require a large amount of preclassified data to differentiate bloggers from true experts. The top 30% of the 3200 profiles were marked as bloggers with a precision score of 0.70. Galán-Garciá *et al.* [83] have developed a model for identifying fake profiles on Twitter. They collected 1900 tweets referring to 19 separate tweeter accounts. The tweets were modeled using the vector space model, and then, supervised classification models were applied. In the best case, the accuracy of the model was 68.47% using the sequential minimal optimization technique.

Egele *et al.* [85] suggested a way to identify compromised accounts on social networks. They used Twitter and FB datasets for their work. The Twitter dataset contained 1.4 billion tweets. The stream includes live tweets that are sent to Twitter. The FB dataset included 106 373 952 wall posts obtained from five regional networks (i.e., London, New York, Los Angeles, Monterey Bay, and Santa Barbara). On the Twitter dataset, their model gave 3.6% false positives, while on FB, the false positive rate was 3.3%. The drawback of this model is that if an attacker is aware of the proposed model's workings, then they have a range of options to prevent his compromised accounts from being identified.

Phad and Chavan [90] suggested a model for identifying a compromised profile on social networks. They gathered data from Twitter accounts for their study using the Twitter archive. The dataset contained 26 363 tweets from 48 high-profile accounts. Out of these, 25 363 tweets are legitimate, and 1000 were malicious. The program creates a user history profile and determines whether or not the account is compromised based on this profile. Seven features used were time, message source, message body text, message subject, message connection, direct user interaction, and proximity, and the accuracy value of 88.29% was obtained for the best case. The key researches using textual features are shown in Table III.

C. Work With Account- and Text-Based Features

This section addresses the research done to detect fake profiles on OSN websites using account- and text-based features (the key researches using hybrid features are shown in Table IV). The GAIN measure technique [107] was applied to the training dataset to produce weighting for all attributes. The weighting

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

TABLE III
KEY RESEARCH USING TEXTUAL FEATURES

Source	Dataset	Problem	Method	Result	OSNs covered
[80]	Twitter and honey pot	Fake account detection	Detection by using blacklist	F-measure: 0.954	Only Twitter
[85]	Twitter and FB	Detecting compromised accounts	COMPA	3.6% false positives	Twitter and FB
[86]	Twitter	Segregating spammers	HITS approach	Precision: 0.70	Only Twitter
[88]	Twitter	Detecting automation on Twitter	NLP	TP: 86% FP:22%	Only Twitter
[90]	Twitter	Detecting compromised high-profile accounts	Behavioral analysis	Precision:99.07%	Only Twitter

TABLE IV
KEY RESEARCH USING HYBRID FEATURES

Source	Dataset	Problem	Method	Result	OSNs covered
[6]	Twitter	Fake account detection	RF, DT, NB, NN, and SVM	F-measure: 0.85	Only Twitter
[29]	Twitter	Fake account detection	NB	F1-score: 0.90	Only Twitter
[62]	Twitter	Malicious account detection	ML	F-measure: 0.93	Only Twitter
[59]	Twitter	Fake account detection	Optimized ML algorithms	F1-score: 0.99	Only Twitter
[74]	Twitter	Malicious accounts detection	SVM, DT, NB, RF, ANN	Accuracy: 94.0%	Only Twitter
[66]	Twitter	Real time phishing detection	PhishAri	Accuracy: 92.52%	Only Twitter
[67]	Twitter	Abuse monitoring	SVM	F1-score:0.70	Only Twitter
[69]	Twitter	Spammers detection	C-Means	Accuracy: 97.98%	Only Twitter
[71]	Twitter	Spam account detection	ML	Precision: 0.933	Only Twitter
[72]	Twitter	Detecting malicious users	BayesNet, NB, SMO, J48, RF	Accuracy: 99.85%	Only Twitter
[75]	Twitter	Spam profile detection	C4.5, KNN, NB, MLP	Precision: 0.946	Only Twitter
[79]	Twitter	Detecting spam accounts	KNN, DT, NB, RF, LR, SVM	F-measure: 0.91	Only Twitter
[82]	Twitter	Spam profile detection	Classification model	Accuracy: 79.26%	Only Twitter
[84]	Twitter	Detecting Sybil accounts	ML classifiers	F1-score: 0.87	Only Twitter

attributes determine the efficiency of the attribute in the classification algorithm. Five classification algorithms were used: RF, DT, NB, NN, and SVM. First, they applied the listed classifiers to all attributes defined and obtained the F1-score values of 81.71%, 78.25%, 73.38%, 77.85%, and 79.13%, respectively. Second, they replicate the experiment using all the weighted attributes calculated and obtained an F1-score of 83.38%, 83.47%, 85.54%, 84.94%, and 85.40%, respectively. In the third experiment, the same series of classifiers were implemented using a minimized set of weighted attributes and obtained an F1-score of 82.76%, 85.03%, 85.36%, 84.87%, and 85.06%, respectively.

Aggarwal *et al.* [66] proposed a model called *PhishAri* to detect phishing on Twitter in real time. *PhishAri* detects phishing on Twitter in real time. The data were collected from Twitter and then labeled into two classes: phishing and legitimate. The Twitter streaming application programming interface (API) and the "Filter" feature offered by the API was used to collect 309 321 tweets from February 1 to April 19, 2012. Two blacklists of PhishTank and Google Safe Browsing were used to mark the collected dataset in the phishing or legitimate class. URL-based features, WHOIS-based features, Tweet-based features, user attributes, and network-based features have been used for phishing detection. The API offers a POST form for the submission of tweets for review. If a tweet is sent to the API, it classifies the URL as "phishing" or "safe" with the help of a collection of defined features using a trained classifier model preloaded to the server. The model was able to detect 80.6% more URLs than common blacklists like PhishTank and Google Safe Browsing at zero hours with 92.52% accuracy. The detection mechanism also works better than the Twitter defense system by 84.6% at zero hour. Since the model did not reach 100% accuracy, there is always a possibility of false negatives.

Chakraborty *et al.* [67] have proposed a framework called social profile abuse monitoring. They gathered information from the Twitter profile of 5000 users along with their 200 latest tweets. The SVM classifier was used to analyze the dataset. They introduced a four-class classification model for calculating profile similarity indexing based on fine-grained interface similarity characteristics. The F1-score of the proposed approach was 70%, with a precision value of 0.60. Hua and Zhang [82] proposed a spam profile identification interface on Twitter. The dataset included 173 spam accounts and 285 nonspam account screen names for a total of 458 screen names from emails sent by Twitter to followers. The threshold and associative classification techniques were used to achieve the accuracy value of 79.26%. This model is slower, but an iterative version of the model can be developed to improve performance in the future. Singh *et al.* [72] suggested a model for identifying the malicious, nonmalicious, and celebrity users on Twitter. The dataset comprised 7500 users of the website, and it was divided into a 70:30 ratio for training and testing. A total of five classifiers, namely, BayesNet, NB, social media optimization (SMO), J48, and RF, were used for model development. For the best case, their model achieved an accuracy of 99.80% using the RF classifier.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

Cresci *et al.* [59] developed a framework for the identification of fake profiles on Twitter. They divided the work into three phases: first, they studied the existing features and rules for the identification of anomalies in various contexts, such as the Academy and the Media. Second, it developed a dataset for human and fake profiles detection, and finally, it designed machine-based classifiers designed over the collection of rules and revised features. Their experiment showed that the rules used in the Media domain were not useful for the detection of a fake follower, whereas the rules and features associated with Academia produced a good result. The built-in model successfully detected 95% of fake profiles on training data of Twitter. Alsaleh *et al.* [84] have developed a model to detect Sybil accounts on Twitter. 1.8 million Twitter accounts were collected, of which 48.05% human, 44.42% Sybil, and the remaining 7.53% were hybrid accounts. Seventeen features were extracted from the dataset, and five ML-based classifiers were used to classify the accounts into three classes. The Weka tool was used for the experiment and found the best results with the combination of MLP and gradient descent, where the human, hybrid, and Sybil F1-scores were 0.94, 0.22, and 0.93, respectively.

El Azab *et al.* [6] used the minimum weighted function collection to detect fake profiles on Twitter. They defined a minimized set of key factors that affect the identification of fake Twitter accounts and then used various classification methods to assess the factors. They used a dataset called: the fake project.[13] The dataset consists of 1481 human accounts and 3000 fake profiles. David *et al.* [74] suggested a model for selecting the appropriate features to identify fake accounts. The dataset of 853 bot profiles and the most recent 1000 tweets in each timeline was collected over a week. This was complemented by 791 manually labeled human accounts between April and June 2016, most of which were Mexican users. User accounts and part of their timelines were extracted via the Twitter API. SVM, DT, NB, RF, and single-layer feedforward artificial neural networks (ANNs) have been used to classify the accounts. The highest average accuracy was 94% achieved with an RF classifier operating on 19 features. The relatively low variation of results across classifiers supports the belief that consistency features and subset selection of features play a significant role in incorrect predictions. Despite the convergence toward 91–92%, the remaining methods have not been similarly effective in terms of growth.

Ercsahin *et al.* [29] have built a fake account identification model on Twitter. A total of 501 fake and 499 real account data were collected manually from Twitter. Sixteen features were obtained from the Twitter API information to differentiate between legitimate and fake profiles. They believed that the predictive attributes were conditionally independent and thus used the NB classifier to characterize the Twitter accounts. First, they applied NB to the dataset using all attributes without discretization and obtained a weighted average F-measure value of 0.860. In the second experiment, the discretization was applied and achieved the F-measure's weighted average value of 0.909. The performance indicated the importance discretization technique, as the model output was enhanced by preprocessing the dataset using a discretization technique for the selected features.

Venkatesh *et al.* [62] proposed a model for detecting malicious accounts on Twitter. They collected 4230 user information and 380 suspicious accounts from Twitter. They also collected user tweets and used hashtags. The dataset was divided into a training and test sample at a ratio of 80:20. They used the Weka tool for implementation. They calculated the trust score for all users and set a threshold value of $\theta = 0.5$; if the calculated trust score was greater than the threshold value, then the user was classified as legitimate. Otherwise, the user was not trustworthy. The DT classifier detected 75% of malicious users correctly, while 25% of malicious users were misclassified.

Al-Zoubi *et al.* [75] detected spam profiles using public features on Twitter. The dataset contained 82 Twitter profiles in which users posted in English and Arabic languages. Features such as *suspicious word*, *default image*, *text-to-link ratio*, *comment ratio*, *tweet time*, and others were extracted from the 82 profiles. ML-based classifiers, namely, DT, C4.5, KNN, NB, and MLP, were used to classify the data into spammer and nonspammer profiles. All classifiers were applied using tenfold cross-validation with stratified sampling as a training/testing methodology. NB delivered the highest performance and obtained 95.70% accuracy, 94% accuracy, and 96% recall.

Alom *et al.* [79] have suggested a model for detecting spam accounts on Twitter. They used a combination of graphics- and content-based features that are sufficient for spam account identification. Several ML classifiers such as KNN, DT, NB, RF, LR, SVM, and XGBoost have been used on selected features to distinguish spam and legitimate users' account. They found that the RF classifier provided a better result than the other classifiers and obtained an accuracy of 91% and an F1-score of 0.91 for the best case. Aswani *et al.* [69] suggested a model for the identification of spammers on Twitter. They collected 1 844 701 tweets from 14 235 Twitter profiles on 13 statistically relevant factors. The factors were extracted from social media analytic. They used a bioinspired algorithm, namely, Firefly, to detect the spammer and nonspammer and achieved a 97.98% accuracy value with the help of tweet-based features like polarity and diversity hashtag frequency, unique words, and others.

Adewole *et al.* [71] proposed a model that detected both the spam message and the spam account in the OSN websites. For the identification of spam messages, a dataset was used, which was compiled from three sources: SMS collection V.1, SMS Corpus V.0.1 Big, and Twitter Spam Corpus with a total of 5574, 1324, and 18 000 samples, respectively. 20 998 twitter accounts and 3 755 367 tweets were used to detect spam accounts. Eighteen features were collectively extracted in both projects, including content/behavior-based features in identifying spammers on Twitter. They used ML-based algorithms and obtained a precision value of 0.933 and an area under the curve (AUC) value of 0.977 in the best case using the RF classifier. Moreover, we have listed the top ten models developed for fake profile detection in Table VI. The popular datasets that the researchers used for fake account detection are listed in Table V. Researchers

[13] [Online]. Available: http:/wafi.iit.cnr.it/theFakeProject

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

TABLE V
DATASETS USED FOR FAKE PROFILE DETECTION WITH SOURCE ADDRESS

Dataset Name	Source
TFP (the fake project): 100% humans	Data Source[17]
E13 (elections 2013): 100% humans	
INT (intertwitter): 100% fake followers	
FSF (fastfollowerz): 100% fake followers	
TWT (twittertechnology): 100% fake followers	
#Elezioni2013 dataset:	Data Source[18]
Deceptive Opinion Spam Corpus	Data Source[19]
BibSonomy	Data Source[20]
Tweets2011 Corpus	Data Source[21]

TABLE VI
TOP TEN MODELS FOR FAKE ACCOUNT DETECTION

Source	Method	Accuracy
[72]	RF	99.80%
[65]	ML algorithms and NN	99.20%
[76]	SVM, NN	98.00%
[69]	C-Means Clustering	97.98%
[101]	ML algorithms with Optimization	97.30%
[60]	ML algorithms	95.00%
[74]	SVM, DT, NB, RF, ANN	94.00%
[100]	KNN, SVM, Similarity matching	93.87%
[61]	Activity-based profile detection approach	93.00%
[66]	PhishAri	92.52%

may also follow the link available on GitHub[14] to get the implemented models for fake profile detection.[15] There were many prototype models are available on GitHub, which may be helpful to initiate the research for the fake profile detection.[16]

V. CHALLENGES

Today's online social networking websites such as FB, Twitter, Tumblr, LinkedIn, and others have become one of the favorites platforms for all age group users to share their achievements, activities, feelings, and life stories community. Openness, friendly interface, and fewer limitations on registration on the websites have created many community developers' challenges. There are currently several challenges associated with these sites that need to be addressed soon. Fig. 3 lists a few challenges of the social networking websites such as security and privacy [108], cyberbullying [109]–[112], identification of fake profiles [16], [33], [113]–[115], event detection, link prediction [116], rumor detection [117]–[119], spam detection [24], [39], [120], [121], and abusive content filtering [122]–[124] to name a few.

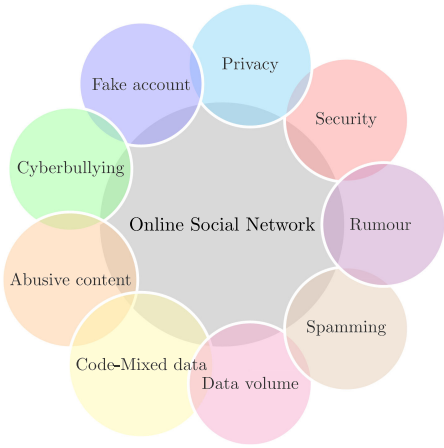

Fig. 3. Challenges of online social networks.

This survey addressed issues related to fake profile identification using several approaches: Section IV-A discussed the research using the account-based features, Section IV-B discussed the research using the textual features, and Section IV-C discussed the works that were developed using both the textual and account-based features. We addressed the works that were done using account- and text-based features to detect the fake/bot/spammer accounts on the OSN websites. Existing work mainly used traditional ML classifiers such as RF, NB, SVM, KNN, XGBoost, and others. Previous research confirmed that, apart from traditional ML algorithms, DL models such as the CNN, LSTM, bidirectional LSTM, and their variants have achieved remarkable success. However, to identify fake profiles, these deep models have not been widely used so far. For this purpose, a future researcher should take advantage of deep models and build systems that can easily discern bot (a software program) and real human profiles on these platforms. Other challenges, such as handling *code-mixed* data [125]–[127], filtering abusive content [122]–[124], handling data volume, have also not been further explored and may, therefore, be considered by future researchers. Although little research has been done in recent years, issues such as cyberbullying [109], [111] and security and privacy [128], [129] need a more robust and generalized framework in the current time.

1) *Security and Privacy [128], [129]:* OSN websites are an open platform to share information, and hence, sometimes, people unknowingly share any private information, such as contact details, personal details, and others on the website. To make such information confidential and hide it from worldwide users is itself a big challenge for OSN websites.
2) *Rumor Detection [117]–[119]:* An unofficial interesting story or piece of news that might be true or invented and quickly spreads from person to person is termed as a rumor. Social networking websites become a favorite medium to spread the rumor at the current time. It is a big challenge for the OSN websites to authenticate the rumors messages before it is spread over the network. Another challenge may include backtracking the source of the rumor message.

[14][Online]. Available: www.github.com
[15][Online]. Available: https://github.com/harshitkgupta/Fake-Profile-Detection-using-ML
[16][Online]. Available: https://github.com/radheysm/Fake-Profile-Detection
[17][Online]. Available: http://mib.projects.iit.cnr.it/dataset.html
[18][Online]. Available: https://github.com/lnyaki/BigDataProject3
[19][Online]. Available: https://myleott.com/op-spam.html
[20][Online]. Available: http://www.kde.cs.uni-kassel.de/ws/rsdc08
[21][Online]. Available: http://trec.nist.gov/data/tweets

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

3) *Spam Detection [24], [39], [120], [130]:* The social networking platform allows users to post an unlimited number of posts, and hence, it is the first choice for the spammer. Earlier, the spammer uses email to spread the spam, but due to more user engagement on the OSN website, the spammer also moved from email to OSN. Managing the spam post on the OSN is a major challenge.
4) *Data Volume [33], [129]:* Social networking websites are the key to future research because they have huge information shared by their users worldwide. To extract the information, it is needed to process the complete data. But the size of the data is one of the major issues. Every day a huge volume of data is transferred through the OSN medium in different formats such as files, audio, video, images, etc. If you want to extract any information from these data, you need to process the data first, which is a big challenge.
5) *Code-Mixed Data [125]–[127]:* On social networking websites, users are connected from every corner of the world, and hence, they received the data in different formats and languages. Currently, the model processes the data if it is written correctly in any language; however, the system will raise an issue if it receives code-mixed data as input.
6) *Abusive Content Detection [122]–[124]:* There is no barrier to joining and posting content on the OSN websites, and hence, along with the informative posts, some abusive posts are also posted by the users. To stop or mitigate this, it needs a very robust system that must understand different languages, because the abusive contents are generally posted in locally spoken languages.
7) *Cyberbullying Detection [109]–[112]:* One of the burning issues across the different online social networking websites is cyberbullying. Many incidents happened in recent years where the victim (bullied) hang themselves and end their lives. A model that can trace the bully post on the OSN websites is very much needed at the current time to early stop future damages.
8) *Fake Account Detection [16], [33], [113]–[115]:* A fake account is used on social networking websites to propagate unauthenticated messages such as spam, rumor, or brand promotions and similar event. A large number of fake accounts currently exist on popular OSN platforms. To filter and remove it needs a generalized model. In recent years, researchers have been contributed to detect fake accounts using various ML and DL models. The current research mainly focused on Twitter; however, FB has more fake accounts. Future research may collectively look at this issue and develop a generalized model that identifies the fake profile in the early stage and stops the unauthenticated communications.

VI. Limitations and Future Research Scope

The online social networking platform's importance has increased in recent years as it provides a common platform for users across the globe to communicate with each other. At the same time, a large number of challenges were also raised, as discussed in Section V. This survey's drawback includes the scope of issues, while a large number of issues exist; this survey only covers fake profile detection. Other issues, such as security, privacy, and so on, remain untouched. The other topics are not included in this study because each of the topics listed is relevant and requires a separate discussion to cover it. Another limitation of this survey is the avoidance of similar topic research like spam and nonspam message detection. This research mainly talked about the methodologies of fake profile detection, and hence, similar research works remain untouched.

As mentioned before, OSN websites have several problems that have not yet been addressed, specifically, the modeling of expression. The social networking platform used English for correspondence, but these websites have now provided content from users in mixed languages, such as Hindi–English and Hindi, which is a difficult task for the program to fully understand the context of the message as it is trained in English corpora. Another problem is finding the message path. OSN websites allow users to exchange the content, and then, without validating the message context, it is forwarded to the network and, thus, flooded. If the message is a rumor, it is going to be a huge problem. To avoid spreading the rumor on OSN websites, the message's context must be verified before it is posted. Finding the route of the message is very critical in this circumstance. The current system needs to be strengthened in order to monitor the message back and determine the source of the message. As we know, no pre-evaluation of the message is being used by any OSN websites. Hence, a future researcher may develop a model that validates the user's post before it is posted to the OSN websites, such as validation of whether the post is abusive, bullied, hate speech, rumor, or similar. If the user's post passes all listed categories, then only it allows to be posted on OSN websites. This mechanism will help to reduce unwanted and unauthorized content from the website.

VII. Conclusion

We summarized the studies that were conducted on fake account identification on online social networking websites. A large number of relevant research published using the global community were compiled and classified into three categories: 1) research using nontextual (account-based) features; 2) research using textual features; and 3) research using both textual and nontextual features. The textual and account-based features have been explained in detail. The researcher may either use existing features or add new features after reviewing the current collection of features to improve the prediction accuracy. OSN websites suffer from a range of huge data volumes, the privacy of user account information, cyberbullying, hate speech, fake profile, and many others. Nonetheless, only fake profile identification issues were discussed in this study. The other topic can also be addressed in depth in the future, which will enable potential researchers to recognize research issues. Another problem that needs to be tackled is language modeling. OSN websites are receiving posts in local languages such as Hindi, Marathi, Telugu, Tamil, and the like, but most of the current

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

models have been trained in English only. They have, therefore, not been able to process languages written in languages other than English. Issues such as sentiment analysis may need to include the non-English text and become important issues.

REFERENCES

[1] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "Towards detecting compromised accounts on social networks," *IEEE Trans. Dependable Secure Comput.*, vol. 14, no. 4, pp. 447–460, Jul./Aug. 2017.

[2] P. V. Savyan and S. M. S. Bhanu, "Behaviour profiling of reactions in Facebook posts for anomaly detection," in *Proc. 9th Int. Conf. Adv. Comput.*, 2017, pp. 220–226.

[3] M. A. Wani, N. Agarwal, S. Jabin, and S. Z. Hussain, "Analyzing real and fake users in Facebook network based on emotions," in *Proc. IEEE 11th Int. Conf. Commun. Syst. Netw.*, 2019, pp. 110–117.

[4] K. Chakraborty, S. Bhattacharyya, and R. Bag, "A survey of sentiment analysis from social media data," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 2, pp. 450–464, Apr. 2020.

[5] P. Chunaev, "Community detection in node-attributed social networks: A survey," *Comput. Sci. Rev.*, vol. 37, 2020, Art. no. 100286.

[6] A. El Azab, A. M. Idrees, M. A. Mahmoud, and H. Hefny, "Fake account detection in twitter based on minimum weighted feature set," *Int. Scholarly Sci. Res. Innov.*, vol. 10, no. 1, pp. 13–18, 2016.

[7] E. Karunakar, V. D. R. Pavani, T. N. I. Priya, M. V. Sri, and K. Tiruvalluru, "Ensemble fake profile detection using machine learning (ML)," *J. Inf. Comput. Sci.*, vol. 10, pp. 1071–1077, 2020.

[8] S. Gurajala, J. S. White, B. Hudson, and J. N. Matthews, "Fake twitter accounts: Profile characteristics obtained using an activity-based pattern detection approach," in *Proc. Int. Conf. Social Media Soc.*, 2015, pp. 1–7.

[9] S. R. Sahoo and B. Gupta, "Real-time detection of fake account in twitter using machine-learning approach," in *Advances in Computational Intelligence and Communication Technology*. New York, NY, USA: Springer, Jun. 2020, pp. 149–159.

[10] Z. Wang, L. Sun, W. Zhu, S. Yang, H. Li, and D. Wu, "Joint social and content recommendation for user-generated videos in online social network," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 698–709, Apr. 2013.

[11] H. Zhao, H. Zhou, C. Yuan, Y. Huang, and J. Chen, "Social discovery: Exploring the correlation among three-dimensional social relationships," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 77–87, Sep. 2015.

[12] T. Ji, C. Luo, Y. Guo, Q. Wang, L. Yu, and P. Li, "Community detection in online social networks: A differentially private and parsimonious approach," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 1, pp. 151–163, Feb. 2020.

[13] K. S. Adewole, N. B. Anuar, A. Kamsin, K. D. Varathan, and S. A. Razak, "Malicious accounts: Dark of the social networks," *J. Netw. Comput. Appl.*, vol. 79, pp. 41–67, 2017.

[14] S. Lee and J. Kim, "WarningBird: A near real-time detection system for suspicious URLs in twitter stream," *IEEE Trans. Dependable Secure Comput.*, vol. 10, no. 3, pp. 183–195, May/Jun. 2013.

[15] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: Survey of new approaches and comparative study," *Comput. Secur.*, vol. 76, pp. 265–284, 2018.

[16] P. Wanda and H. J. Jie, "Deepprofile: Finding fake profile in online social network using dynamic CNN," *J. Inf. Secur. Appl.*, vol. 52, pp. 1–13, 2020.

[17] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," *IEEE Trans. Neural Net.*, vol. 10, no. 5, pp. 1048–1054, Sep. 1999.

[18] A. H. Wang, "Don't follow me: Spam detection in twitter," in *Proc. Int. Conf. Securi. Cryptography*, 2010, pp. 1–10.

[19] M. Cha, F. Benevenuto, H. Haddadi, and K. Gummadi, "The world of connections and information flow in twitter," *IEEE Trans. Syst., Man, Cybern. A: Syst. Humans*, vol. 42, no. 4, pp. 991–998, Jul. 2012.

[20] Z. Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang, "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol. 260, pp. 64–73, 2014.

[21] C. Chen *et al.*, "A performance evaluation of machine learning-based streaming spam tweets detection," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 65–76, Sep. 2015.

[22] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Statistical features-based real-time detection of drifted twitter spam," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 4, pp. 914–925, Apr. 2016.

[23] C. Chen *et al.*, "Investigating the deceptive information in twitter spam," *Future Gener. Comput. Syst.*, vol. 72, pp. 319–326, 2017.

[24] P. K. Roy, J. P. Singh, and S. Banerjee, "Deep learning to filter SMS spam," *Future Gener. Comput. Syst.*, vol. 102, pp. 524–533, 2020.

[25] M. Bilge, T. Strufe, D. Balzarotti, and E. Kirda, "All your contacts are belong to us: Automated identity theft attacks on social networking," in *Proc. 18th Int. World Wide Web Conf.*, 2009, pp. 551–560.

[26] A. Bodhani *et al.*, "The A to Z of fakes," *Eng. Technol.*, vol. 7, no. 1, pp. 41–60, 2012.

[27] Y. Boshmaf, M. Ripeanu, K. Beznosov, and E. Santos-Neto, "Thwarting fake OSN accounts by predicting their victims," in *Proc. 8th ACM Workshop Artif. Intell. Secur.*, 2015, pp. 81–89.

[28] A. J. Sarode and A. Mishra, "Audit and analysis of impostors: An experimental approach to detect fake profile in online social network," in *Proc. 6th Int. Conf. Comput. Commun. Technol.*, 2015, pp. 1–8.

[29] B. Ersahin, O. Aktas, D. Kilinc, and C. Akyol, "Twitter fake account detection," in *Proc. IEEE Int. Conf. Comput. Sci. Eng.*, 2017, pp. 388–392.

[30] A. Gupta and R. Kaushal, "Towards detecting fake user accounts in Facebook," in *Proc. ISEA Asia Secur. Privacy*, 2017, pp. 1–6.

[31] Y. Hao and F. Zhang, "Detecting shilling profiles in collaborative recommender systems via multidimensional profile temporal features," *IET Inf. Secur.*, vol. 12, no. 4, pp. 362–374, 2018.

[32] E. V. D. Walt and J. Eloff, "Using machine learning to detect fake identities: Bots vs humans," *IEEE Access*, vol. 6, pp. 6540–6549, 2018.

[33] D. Ramalingam and V. Chinnaiah, "Fake profile detection techniques in large-scale online social networks: A comprehensive review," *Comput. Elect. Eng.*, vol. 65, pp. 165–177, 2018.

[34] D. Yuan *et al.*, "Detecting fake accounts in online social networks at the time of registrations," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2019, pp. 1423–1438.

[35] Y. Elyusufi, Z. Elyusufi, and M. A. Kbir, "Social networks fake profiles detection based on account setting and activity," in *Proc. 4th Int. Conf. Smart City Appl.*, 2019, Art. no. 37.

[36] H. Hazimeh, E. Mugellini, and O. A. Khaled, "Reliable user profile analytics and discovery on social networks," in *Proc. 8th Int. Conf. Softw. Comput. Appl.*, 2019, pp. 496–500.

[37] J. Kaubiyal and A. K. Jain, "A feature based approach to detect fake profiles in twitter," in *Proc. 3rd Int. Conf. Big Data Int. Things*, 2019, pp. 135–139.

[38] G. Suarez-Tangil, M. Edwards, C. Peersman, G. Stringhini, A. Rashid, and M. Whitty, "Automatically dismantling online dating fraud," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1128–1137, 2020.

[39] A. Makkar and N. Kumar, "An efficient deep learning-based scheme for web spam detection in IoT environment," *Future Gener. Comput. Syst.*, vol. 108, pp. 467–487, 2020.

[40] M. Badar, M. Haris, and A. Fatima, "Application of deep learning for retinal image analysis: A review," *Comput. Sci. Rev.*, vol. 35, pp. 1–18, 2020.

[41] R. Jain, N. Jain, A. Aggarwal, and D. J. Hemanth, "Convolutional neural network based Alzheimer's disease classification from magnetic resonance brain images," *Cogn. Syst. Res.*, vol. 57, pp. 147–159, 2019.

[42] M. Vardhana, N. Arunkumar, S. Lasrado, E. Abdulhay, and G. Ramirez-Gonzalez, "Convolutional neural network for bio-medical image segmentation with hardware acceleration," *Cogn. Syst. Res.*, vol. 50, pp. 10–14, 2018.

[43] E. Daskalakis, M. Tzelepi, and A. Tefas, "Learning deep spatiotemporal features for video captioning," *Pattern Recognit. Lett.*, vol. 116, pp. 143–149, 2018.

[44] M. Nabati and A. Behrad, "Video captioning using boosted and parallel long short-term memory networks," *Comput. Vis. Image Understanding*, vol. 190, 2020, Art. no. 102840.

[45] M. Nabati and A. Behrad, "Multi-sentence video captioning using content-oriented beam searching and multi-stage refining algorithm," *Inf. Process. Manage.*, vol. 57, no. 6, 2020, Art. no. 102302.

[46] H. Xiao and J. Shi, "Video captioning with text-based dynamic attention and step-by-step learning," *Pattern Recognit. Lett.*, vol. 133, pp. 305–312, 2020.

[47] M. Verma and S. Sofat, "Techniques to detect spammers in twitter—A survey," *Int. J. Comput. Appl.*, vol. 85, no. 10, pp. 27–32, 2014.

[48] K. Anand, J. Kumar, and K. Anand, "Anomaly detection in online social network: A survey," in *Proc. IEEE Int. Conf. Inventive Commun. Comput. Technol.*, 2017, pp. 456–459.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

[49] M. Al-Qurishi, M. Al-Rakhami, A. Alamri, M. Alrubaian, S. M. M. Rahman, and M. S. Hossain, "Sybil defense techniques in online social networks: A survey," *IEEE Access*, vol. 5, pp. 1200–1219, 2017.

[50] A. Karatas and S. Sahin, "A review on social bot detection techniques and research directions," in *Proc. Int. Secur. Cryptol. Conf.*, 2017, pp. 156–161.

[51] S. Gheewala and R. Patel, "Machine learning based twitter spam account detection: A review," in *Proc. IEEE 2nd Int. Conf. Comput. Methodol. Commun.*, 2018, pp. 79–84.

[52] R. Kaur, S. Singh, and H. Kumar, "Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches," *J. Netw. Comput. Appl.*, vol. 112, pp. 53–88, 2018.

[53] E. Alothali, N. Zaki, E. A. Mohamed, and H. Alashwal, "Detecting social bots on twitter: A literature review," in *Proc. IEEE Int. Conf. Innovations Inf. Technol.*, 2018, pp. 175–180.

[54] S. P. Velayudhan and M. S. B. Somasundaram, "Compromised account detection in online social networks: A survey," *Concurrency Comput.: Pract. Experience*, vol. 31, no. 20, 2019, Art. no. e 5346.

[55] M. Apte, G. K. Palshikar, and S. Baskaran, "Frauds in online social networks: A review," in *Social Networks and Surveillance for Society*. Berlin, Germany: Springer, 2019, pp. 1–18.

[56] R. Krithiga and E. Ilavarasan, "A comprehensive survey of spam profile detection methods in online social networks," *J. Phys.: Conf. Ser.*, vol. 1362, no. 1, 2019, Art. no. 012111.

[57] S. Joshi, H. G. Nagariya, N. Dhanotiya, and S. Jain, "Identifying fake profile in online social network: An overview and survey," in *Proc. Int. Conf. Mach. Learn., Image Process., Netw. Secur. Data Sci.*, 2020, pp. 17–28.

[58] D. Camacho, A. Panizo-LLedot, G. Bello-Orgaz, A. Gonzalez-Pardo, and E. Cambria, "The four dimensions of social network analysis: An overview of research methods, applications, and software tools," *Inf. Fusion*, vol. 63, pp. 88–120, 2020.

[59] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "A fake follower story: Improving fake accounts detection on twitter," IIT-CNR, Pisa, Italy, Tech. Rep. TR-03, 2014.

[60] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Fame for sale: Efficient detection of fake twitter followers," *Decis. Support Syst.*, vol. 80, pp. 56–71, 2015.

[61] S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, "Profile characteristics of fake twitter accounts," *Big Data Soc.*, vol. 3, no. 2, 2016, Art. no. 2053951716674236.

[62] R. Venkatesh, J. K. Rout, and S. Jena, "Malicious account detection based on short URLs in twitter," in *Proc. Int. Conf. Signal, Netw., Comput., Syst.*, 2017, pp. 243–251.

[63] N. Singh, T. Sharma, A. Thakral, and T. Choudhury, "Detection of fake profile in online social networks using machine learning," in *Proc. IEEE Int. Conf. Adv. Comput. Commun. Eng.*, 2018, pp. 231–234.

[64] L. Caruccio, D. Desiato, and G. Polese, "Fake account identification in social networks," in *Proc. IEEE Int. Conf. Big Data*, 2018, pp. 5078–5085.

[65] M. Rahman, A. M. Likhon, A. Rahman, and M. H. Choudhury, "Detection of fake identities on twitter using supervised machine learning," Ph.D. dissertation, Brac Univ., Dhaka, Bangladesh, 2019.

[66] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on twitter," in *Proc. eCrime Researchers Summit*, 2012, pp. 1–12.

[67] A. Chakraborty, J. Sundi, S. Satapathy, "SPAM: A framework for social profile abuse monitoring," Stony Brook Univ., Stony Brook, NY, USA, CSE508 Report, 2012.

[68] S. Adikari and K. Dutta, "Identifying fake profiles in linkedin," in *Proc. Pacific Asia Conf. Inf. Syst.*, 2014, p. 278.

[69] R. Aswani, A. K. Kar, and P. V. Ilavarasan, "Detection of spammers in twitter marketing: A hybrid approach using social media analytics and bio inspired computing," *Inf. Syst. Front.*, vol. 20, no. 3, pp. 515–530, 2018.

[70] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake news detection on social media using geometric deep learning," 2019, *arXiv:1902.06673*.

[71] K. S. Adewole, N. B. Anuar, A. Kamsin, and A. K. Sangaiah, "SMSAD: A framework for spam message and spam account detection," *Multimedia Tools Appl.*, vol. 78, no. 4, pp. 3925–3960, 2019.

[72] M. Singh, D. Bansal, and S. Sofat, "Detecting malicious users in twitter using classifiers," in *Proc. 7th Int. Conf. Secur. Inf. Netw.*, 2014, pp. 247–253.

[73] Y. Boshmaf *et al.*, "Integro: Leveraging victim prediction for robust fake account detection in OSNs," in *Proc. Netw. Distributed Syst. Secur. Symp.*, 2015, pp. 8–11.

[74] I. David, O. S. Siordia, and D. Moctezuma, "Features combination for the detection of malicious twitter accounts," in *Proc. IEEE Int. Autumn Meeting Power, Electron. Comput.*, 2016, pp. 1–6.

[75] A. M. Al-Zoubi, J. Alqatawna, and H. Paris, "Spam profile detection in social networks based on public features," in *Proc. 8th Int. Conf. Inf. Commun. Syst.*, 2017, pp. 130–135.

[76] S. Khaled, N. El-Tazi, and H. M. Mokhtar, "Detecting fake accounts on social media," in *Proc. IEEE Int. Conf. Big Data.*, 2018, pp. 3672–3681.

[77] J. S. Alowibdi, U. A. Buy, S. Y. Philip, S. Ghani, and M. Mokbel, "Deception detection in twitter," *Social Netw. Anal. Mining*, vol. 5, no. 1, 2015, Art. no. 32.

[78] M. Conti, R. Poovendran, and M. Secchiero, "Fakebook: Detecting fake profiles in on-line social networks," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining.*, 2012, pp. 1071–1078.

[79] Z. Alom, B. Carminati, and E. Ferrari, "Detecting spam accounts on twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining.*, 2018, pp. 1191–1198.

[80] M. M. Swe and N. N. Myo, "Fake accounts detection on twitter using blacklist," in *Proc. IEEE/ACIS 17th Int. Conf. Comput. Inf. Sci.*, 2018, pp. 562–566.

[81] K. Shu, S. Wang, and H. Liu, "Understanding user profiles on social media for fake news detection," in *Proc. IEEE Conf. Multimedia Inf. Process. Retrieval*, 2018, pp. 430–435.

[82] W. Hua and Y. Zhang, "Threshold and associative based classification for social spam profile detection on twitter," in *Proc. 9th Int. Conf. Semantics, Knowl. Grids*, 2013, pp. 113–120.

[83] P. Galn-Garcia, J. G. d. l.Puerta, C. L. Gomez, I. Santos, and P. G. Bringas, "Supervised machine learning for the detection of troll profiles in twitter social network: Application to a real case of cyberbullying," *Logic J. IGPL*, vol. 24, no. 1, pp. 42–53, 2016.

[84] M. Alsaleh, A. Alarifi, A. M. Al-Salman, M. Alfayez, and A. Almuhaysin, "TSD: Detecting sybil accounts in twitter," in *Proc. 13th Int. Conf. Mach. Learn. Appl.*, 2014, pp. 463–469.

[85] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2013.

[86] M. U. S. Khan, M. Ali, A. Abbas, S. U. Khan, and A. Y. Zomaya, "Segregating spammers and unsolicited bloggers from genuine experts on twitter," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 551–560, Jul./Aug. 2018.

[87] C. Xiao, D. M. Freeman, and T. Hwa, "Detecting clusters of fake accounts in online social networks," in *Proc. 8th ACM Workshop Artif. Intell. Secur.*, 2015, pp. 91–101.

[88] E. M. Clark, J. R. Williams, C. A. Jones, R. A. Galbraith, C. M. Danforth, and P. S. Dodds, "Sifting robotic from organic text: A natural language approach for detecting automation on twitter," *J. Comput. Sci.*, vol. 16, pp. 1–7, 2016.

[89] S. M. Asadullah and S. Viraktamath, "Classification of twitter spam based on profile and message model using SVM," *Int. Res. J. Eng. Technol.*, vol. 4, no. 5, pp. 2862–2865, 2017.

[90] P. V. Phad and M. Chavan, "Detecting compromised high-profile accounts on social networks," in *Proc. IEEE 9th Int. Conf. Comput., Commun. Netw. Technol.*, 2018, pp. 1–4.

[91] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, 2019, Art. no. 150.

[92] F. Ahmed and M. Abulaish, "An MCL-based approach for spam profile detection in online social networks," in *Proc. IEEE 11th Int. Conf. Trust, Secur. Privacy Comput. Commun.*, 2012, pp. 602–608.

[93] M. BalaAnand, S. Sankari, R. Sowmipriya, and S. Sivaranjani, "Identifying fake user's in social networks using non verbal behavior," *Int. J. Technol. Eng. Syst.*, vol. 7, no. 2, pp. 157–161, 2015.

[94] A. M. Meligy, H. M. Ibrahim, and M. F. Torky, "Identity verification mechanism for detecting fake profiles in online social networks," *Int. J. Comput. Netw. Inf. Secur.*, vol. 9, no. 1, pp. 31–39, 2017.

[95] Y. Xuan, Y. Chen, H. Li, P. Hui, and L. Shi, "LBSNShield: Malicious account detection in location-based social networks," in *Proc. 19th ACM Conf. Comput. Supported Cooperative Work Social Comput. Companion*, 2016, pp. 437–440.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.

[96] J. Jia, B. Wang, and N. Z. Gong, "Random walk based fake account detection in online social networks," in *Proc. 47th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw.*, 2017, pp. 273–284.
[97] M. Fire, D. Kagan, A. Elyashar, and Y. Elovici, "Friend or foe? Fake profile identification in online social networks," *Social Netw. Anal. Mining*, vol. 4, no. 1, 2014, Art. no. 194.
[98] K. Zarei, R. Farahbakhsh, and N. Crespi, "Deep dive on politician impersonating accounts in social media," in *Proc. IEEE Symp. Comput. Commun.*, Jun. 2019, pp. 1–6.
[99] N. Kumar, P. Dabas, and Komal, "Detection and prevention of profile cloning in online social networks," in *Proc. IEEE 5th Int. Conf. Signal Process., Comput. Control.*, 2019, pp. 287–291.
[100] S. Revathi and M. Suriakala, "Profile similarity communication matching approaches for detection of duplicate profiles in online social network," in *Proc. IEEE 3rd Int. Conf. Comput. Syst. Inf. Technol. Sustain. Solutions*, 2018, pp. 174–182.
[101] M. Suriakala and S. Revathi, "Privacy protected system for vulnerable users and cloning profile detection using data mining approaches," in *Proc. IEEE 10th Int. Conf. Adv. Comput.*, 2018, pp. 124–132.
[102] C. Sandy, P. Rusconi, and S. Li, "Can humans detect the authenticity of social media accounts? on the impact of verbal and non-verbal cues on credibility judgements of twitter profiles," in *Proc. 3rd IEEE Int. Conf. Cybern.*, 2017, pp. 1–8.
[103] G. Kontaxis, I. Polakis, S. Ioannidis, and E. P. Markatos, "Detecting social network profile cloning," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops*, 2011, pp. 295–300.
[104] K. Zarei, R. Farahbakhsh, and N. Crespi, "Typification of impersonated accounts on instagram," in *Proc. IEEE 38th Int. Perform. Comput. Commun. Conf.*, 2019, pp. 1–6.
[105] P. Pourghomi, A. A. Halimeh, F. Safieddine, and W. Masri, "Right-click authenticate adoption: The impact of authenticating social media postings on information quality," in *Proc. Int. Conf. Inf. Digit. Technol.*, 2017, pp. 327–331.
[106] P. Pourghomi, M. Dordevic, and F. Safieddine, "Facebook fake profile identification: Technical and ethical considerations," *Int. J. Pervasive Comput. Commun.*, vol. 16, pp. 101–112, 2020.
[107] A. G. Karegowda, A. Manjunath, and M. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection," *Int. J. Inf. Technol. Knowl. Manage.*, vol. 2, no. 2, pp. 271–277, 2010.
[108] A. De Salve, P. Mori, and L. Ricci, "A survey on privacy in decentralized online social networks," *Comput. Sci. Rev.*, vol. 27, pp. 154–176, 2018.
[109] K. Van Royen, K. Poels, W. Daelemans, and H. Vandebosch, "Automatic monitoring of cyberbullying on social networking sites: From technological feasibility to desirability," *Telematics Inform.*, vol. 32, no. 1, pp. 89–97, 2015.
[110] R. Cohen-Almagor, "Social responsibility on the internet: Addressing the challenge of cyberbullying," *Aggression Violent Behav.*, vol. 39, pp. 42–52, 2018.
[111] I. Marin-Lopez, I. Zych, R. Ortega-Ruiz, S. C. Hunter, and V. J. Llorent, "Relations among online emotional content use, social and emotional competencies and cyberbullying," *Children Youth Serv. Rev.*, vol. 108, 2020, Art. no. 104647.
[112] V. Balakrishnan, S. Khan, and H. R. Arabnia, "Improving cyberbullying detection using twitter users' psychological features and machine learning," *Comput. Secur.*, vol. 90, 2020, Art. no. 101710.
[113] A. Makkar and N. Kumar, "Cognitive spammer: A framework for pagerank analysis with split by over-sampling and train by under-fitting," *Future Gener. Comput. Syst.*, vol. 90, pp. 381–404, 2019.
[114] M. A. Azad and R. Morla, "Rapid detection of spammers through collaborative information sharing across multiple service providers," *Future Gener. Comput. Syst.*, vol. 95, pp. 841–854, 2019.
[115] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "DeepFakes and beyond: A survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, 2020.
[116] P. K. Sharma, S. Rathore, and J. H. Park, "Multilevel learning based modeling for link prediction and users' consumption preference in online social networks," *Future Gener. Comput. Syst.*, vol. 93, pp. 952–961, 2019.
[117] M. Ahsan, M. Kumari, and T. Sharma, "Rumors detection, verification and controlling mechanisms in online social networks: A survey," *Online Social Netw. Media*, vol. 14, 2019, Art. no. 100050.
[118] A. Bodaghi and J. Oliveira, "The characteristics of rumor spreaders on twitter: A quantitative analysis on real data," *Comput. Commun.*, vol. 160, pp. 674–687, 2020.
[119] Z. Wu, D. Pi, J. Chen, M. Xie, and J. Cao, "Rumor detection based on propagation graph neural network with attention mechanism," *Expert Syst. Appl.*, vol. 158, 2020, Art. no. 113595.
[120] L. Chen, Z. Yan, W. Zhang, and R. Kantola, "TruSMS: A trustworthy SMS spam control system based on trust management," *Future Gener. Comput. Syst.*, vol. 49, pp. 77–93, 2015.
[121] A. Singh and S. Batra, "Ensemble based spam detection in social IoT using probabilistic data structures," *Future Gener. Comput. Syst.*, vol. 81, pp. 359–371, 2018.
[122] H.-S. Lee, H.-R. Lee, J.-U. Park, and Y.-S. Han, "An abusive text detection system based on enhanced abusive and non-abusive word lists," *Decis. Support Syst.*, vol. 113, pp. 22–31, 2018.
[123] M. O. Ibrohim and I. Budi, "A dataset and preliminaries study for abusive language detection in indonesian social media," *Procedia Comput. Sci.*, vol. 135, pp. 222–229, 2018.
[124] V. K. Jha, H. P, V. P N, V. Vijayan, and P. P, "DHOT-repository and classification of offensive tweets in the hindi language," *Procedia Comput. Sci.*, vol. 171, pp. 2324–2333, 2020.
[125] K. Asnani and J. D. Pawar, "Automatic aspect extraction using lexical semantic knowledge in code-mixed context," *Procedia Comput. Sci.*, vol. 112, pp. 693–702, 2017.
[126] K. Sreelakshmi, B. Premjith, and K. Soman, "Detection of hate speech text in hindi-english code-mixed data," *Procedia Comput. Sci.*, vol. 171, pp. 737–744, 2020.
[127] T. T. Sasidhar, B. Premjith, and K. P. Soman, "Emotion detection in hinglish (hindi english) code-mixed social media text," *Procedia Comput. Sci.*, vol. 171, pp. 1346–1352, 2020.
[128] H. Schwartz-Chassidim, O. Ayalon, T. Mendel, R. Hirschprung, and E. Toch, "Selectivity in posting on social networks: The role of privacy concerns, social capital, and technical literacy," *Heliyon*, vol. 6, no. 2, 2020, Art. no. e0 3298.
[129] L. Bahri, B. Carminati, and E. Ferrari, "Decentralized privacy preserving services for online social networks," *Online Social Netw. Media*, vol. 6, pp. 18–25, 2018.
[130] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," *Cogn. Syst. Res.*, vol. 61, pp. 32–44, 2020.

Authorized licensed use limited to: Thangal Kunju Musaliar College of Engineering. Downloaded on February 15,2022 at 05:45:39 UTC from IEEE Xplore. Restrictions apply.