2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO)
Amity University, Noida, India. Sep 3-4, 2021

2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) | 978-1-6654-1703-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICRITO51393.2021.9596322

Malicious Profile Detection on Social Media: A Survey Paper

Shruti Shinde
Computer Engineering & IT Department
College of Engineering, Pune
shrutiss19.comp@coep.ac.in

Dr. Sunil B. Mane
Computer Engineering & IT Department
College of Engineering, Pune
sunilbmane.comp@coep.ac.in

***Abstract*— Facebook, Twitter, Instagram, and LinkedIn are all popular online social media sites these days. Everyone uses different social media platforms, from children to adults. The use of these social media applications is increasing, which leads to a rise in social media crime. Here is where the word "fake profiles" comes into play; these fake profiles are responsible for distributing misleading details about a particular user or attempting fraudulent activities with bad intentions. Many researchers have been carried in response to the issue of fake profiles that exist in any social media application. There is research available in different directions to get a real problem solution. This paper summarizes all of the research on fake profile identification in social media platforms. This paper gives survey of work, indicating whether the user profile is fake or real.**

Keywords*— *Machine learning, Social media, Supervised, Unsupervised, Dataset.

I. INTRODUCTION

Social networking has grown into a massive medium for people to communicate with each other. People can connect with each other through social networking sites such as Twitter, Facebook, Instagram, LinkedIn, Tinder, and other matrimonial sites. Different social sites serve different purposes; for example, facebook and instagram are social platforms where people can make friends and communicate with one another through the internet. People use Twitter to share their opinions on a variety of topics. LinkedIn is a huge platform for students and employees to find jobs or hire people.

There are also disadvantages of all these benefits. These social media sites have malicious user accounts. They can be very dangerous and may distribute wrong content about particular users or distribute misleading information. It can harm the privacy of people. Most of the experimental work has been carried out in the area of the supervised and unsupervised machine learning algorithms.

The first step in detecting fake accounts is to pick the target profile for review in order to extract the profile's feature set, which includes the name, chat, user profile, history, location, background profile picture, friends list, followers, likes on the post, number of posts, comments on each post, and tagging. Then using supervised or unsupervised machine learning classifier, it determined if the target profile is fake or not. The study presents different supervised and unsupervised classification algorithms with giving increase in number of accuracy term. This paper provides a study of existing solutions for detecting malicious profiles on social media platforms. For this research, total 19 research papers are used and based on the research described in this paper, the accuracy rate ranges between 50 -97%.

The study of fake profile detection has been carried out with the use of different types of datasets. The Twitter and Facebook datsets are considered as the most common dataset among them. Selection of the attributes is the first and very crucial step in this process. The selection of features is depends on the following three types.

- User based features – This type gives information about the user profiles more briefly [19] with their account name, followers or following count, number of post posted on the profile etc.
- Content based features – In content-based type, the content or tweets posted by a user are considered. The description of specific user under their profile is also a example of such content based features. [19]
- Time-zone based features – This is the type where a person has posted posts in specific time [19] or with certain time gap.
- Graph based features - This type of feature mainly focused on the interactions of various users with others on particular social media sites [19]. For example, the users' friends and friends of friends present on Facebook social site. [19]

Rest of the paper is described as follow, Section II presents a study of the literature relevant to this study, Section III provides a general overview of the system for malicious user account detection, Section IV provides details on the general results provided by the models and Section V concludes the paper.

II. RELATED WORK

This section of the paper describes previous research on the identification of malicious accounts and presents available solution for it. In [1] with the use of Facebook as a social media application, this study proposes a solution with an analytical ranking system that compares both graph-based and feature-based approaches. This paper attempted to catch the fake users on Facebook as well as determine the relationship between fake users and actual user accounts. For identifying the respected results, they used a support vector machine with a suitable kernel function as a classification algorithm. The selection of the feature set for the regression model is the first step in the proposed model in this paper. The next step is to build the social graph, which is divided into two regions: Benign and Sybil. The third step is to use these social graphs to evaluate the probability of the

outcomes. As a result, normalizing the SVM output measures the possibility score approximately and provides a better guess for the Sybilwalk algorithm. This method is found to be effective for large datasets.

[2] This study proposed a method for detecting fake profiles on the Facebook social media site using support vector machines, random forests, JRip, and naive bayes algorithms. For the purpose of identifying fake users, this study considers the emotions behind their posts. They claim that a normal user can express a wide range of emotions through their posts, while a fake profile can express the same range of emotions with consideration for the motive. For this review, they used sentiment analysis. The dataset contains four columns: the user ID number, the user's content, the post ID number and a column that indicates whether the profile is fake. It is determined whether the respective profile is fake or not based on the various emotion categories. The random forest algorithm, which has the highest accuracy among the four algorithms, has 90 % accuracy.

For detecting fake profiles, the paper [3] implements a graph-based learning algorithm. The EGSLA algorithm is a supervised learning method. They compared the results of this algorithm with Support vector machine, k-nearest neighbour, and decision tree, which are all machine-learning algorithms. The study's dataset was gathered from Twitter using the scrappy python framework for scraping. Data collection, feature extraction, classification, and decision-making are the four phases proposed in this paper. The fraction of retweets a user has received, the user's normal retweet length, whether the user has used any URL in the tweet, and the user's average time between tweets are all factors taken into consideration. According to the paper's comparative review, the EGSLA is more accurate than the other three algorithms.

The researchers suggested a method for detecting duplicate accounts on social media apps [4]. Duplicate user accounts are those that have stolen account information from other users; these profiles are then used to promote misleading content or fake news with malicious purposes. This process is divided into four phases. The first is the detection phase, in which the victim suspects that another person is using his or her profile information. The profile matcher process follows, in which a query is operated to look up the name of the fake account on social media. The third step is the similarity measurement phase, which uses cosine similarity, n-gram similarity, and string matching similarity measures to determine how similar two accounts are. The similarity between two user accounts is measured using attribute and network similarities. The verification process is the final stage, in which it is determined whether the profile has been cloned. The data from Facebook and Twitter is used in this analysis.

With regard to the LinkedIn social media platform, paper [17] proposes a solution for cloned profile identification that provides a score based on the similarities between two profiles. The three components listed in the research are information distiller, profile hunter and profile verifier. The key concept behind this research is to identify the detail that uniquely identifies the particular individual. Using the details of a legitimate user, one may identify cloned accounts on social media that are identical to the same legitimate user.

The information we share on social media is extremely personal to each of us, and fake profiles can negatively impact this privacy by deploying various types of privacy attacks. For their study, the paper [5] addresses the problem of fraud users. They also described various types of attacks that can occur on social media platforms. They suggest a framework based on the social media graph's growth rate, as well as how users are linked to their friends and their relation graphs. In a social media network, the method provides empirical research, interaction, and statistics of users with other users. Facebook data is scraped from a user's wall using an API. They have considered the factors such as changes of number of friends over the time, the mean difference with respect to the total population over the social sites, average degree of datasets and considered all the nodes from the friends graph. According to the paper, they developed a sensing framework that can gather statistical data from a user's profile and helps to mitigate the danger of fake accounts on social media sites.

The paper [6] presents a method for detecting fake users using feature combination and data gathered from the Twitter social media platform. They also extracted 71 properties from the user's profile including timeline data. Supervised machine learning algorithms are used for classification tasks. They chose content-based and timeline-based features for predicting the results. Each variable feature set is chosen based on its significance. The model achieves 94% accuracy by using random forest for 19 features. The 10-cross fold validation method is used to determine accuracy. In addition to the random forest decision tree, naive bayes and neural network algorithms are being used. Various algorithms, such as the random forest algorithm, different algorithms require different numbers of features in their datasets to achieve accurate accuracy. They concluded that good feature quality and feature subset play an important role for accurately predicting outcomes.

LinkedIn is the world's largest professional social networking site. LinkedIn is a forum for job searching and recruitment. According to the paper [7], LinkedIn is used as a social media site for detecting fraud users. For detecting malicious accounts, this uses Neural Network Support Vector Machine, Principal Component Analysis, and Weighted Average data mining techniques. They created three datasets for review, each of which contains 74 user profiles. The results are optimised using a weighted average, and SVM and NN use the training testing approach for classification. This paper's suggested approach has an accuracy of 87% and a true negative rate of 94 %. They presented a study that compared SVM, NN, and the weighted average technique. In addition, the proposed approach completes the comparative analysis by using PCA for feature selection and presents a table of comparison that shows the result with all features and only selected features in the feature set.

Twitter is a well-known social media platform where millions of users share billions of tweets every day. However, with the widespread use of social media platforms comes the possibility of malicious users. These malicious accounts can pose a serious threat to the rest of Twitter's users. Twitter is used in the paper [8] to gather user data and detect fake accounts on Twitter. This paper provides an overview of the methods used to identify fake Twitter users. Machine learning-based strategies are discussed extensively, and they are divided into two categories: syntax analysis and

feature analysis. Syntax analysis approaches are based on the shorted URL or on user-posted tweets. The feature analysis can be performed in three ways: first, using the statistical information from the user's tweets, second, using the statistical information from the user's accounts, and third, using the social graphs of the user's friends. They also listed the various methods for detecting malicious users, which include clustering algorithms, classification algorithms, and hybrid models that combine two or more algorithms.

This [9] paper introduces a framework with a machine learning approach using Facebook and Twitter as social media sites. For detecting malicious users, they took into account the user's posts and status. The study introduced two architectures. The first architecture uses network identifiers with NLP to authenticate users. With the term bag of words, the second architecture employs a support vector machine as a classifier. The bag of words detects harmful words in a user's account and determines whether the user is genuine or not based on the content. Reinforcement learning is used in a paper [11] to detect bot accounts on social media sites. For more accurate results and improved model efficiency, the model needs a feedback mechanism.

Using Facebook user data, this paper [10] proposes a solution for fake accounts on the social media platform Facebook. Classic classification algorithms such as naive bayes, support vector machine, decision tree, and neural network are used to provide more accurate performance. These two algorithms, artificial bee colony and ant colony optimization, are used to select features. Weka tool is used here to predict fake users; it can also be used for data preprocessing, classification, clustering, and visualisation. Reference [12] introduces their work in this domain using NLP and machine learning algorithms. The model's accuracy is improved using support vector machine and naive bayes classification techniques. This study makes use of Facebook user profiles. Tokenization techniques such as stop word elimination, lemmatization, and stemming are used in the text pre-processing stage. The dataset is analysed using tokenization techniques.

Spam user profiles are extremely harmful, as shown by the fact that this problem exists on Twitter. [13] proposed an artificial neural network classification technique as a solution. The dataset for spam consumer identification comes from the H-spam 14 website. The artificial bee colony algorithm is used to improve the model. The main goal of this paper is to automatically recognize spam users and remove their accounts. For pre-processing the dataset, various text-processing techniques are applied to user tweets. They found that an artificial bee colony combined with artificial neural networks outperforms the naive bayes and support vector machine algorithms in terms of accuracy.

Spam accounts on social media can send spam messages, and spammers can use social media to convey malicious data. Spammers may use social media platforms to launch a variety of security attacks that are potentially dangerous to other users. They collected data from 82 twitter profiles in their paper [16], which provides an approach to spam profile detection that is available on Twitter. For feature selection, the information gain and relief techniques are used. They compared the results using the confusion matrix on four classification models: KNN, multilayer perceptron, decision tree, and naive bayes. The dataset is divided into two sections, each of which is labelled as either legitimate or spam profiles. Only information that is publicly available is included in the dataset.

Bot profiles generated by a computer with the aim of increasing the number of likes, comments, or followers on a particular post. Humans build fake profiles with the intent of providing false news or invading someone's privacy, among other things... Malicious users can be divided into two categories: bot profiles and human-created fake profiles. A research about bot and human-created false user profiles is presented in the reference [14]. The paper examines the different bot accounts and human-created fake users, as well as the methods for detecting these types of accounts. Another research with bot detection is seen in reference [15] it gives graph based approach where it represents the link of that network. They presented two phased framework with supervised as well as unsupervised algorithms. Data bootstrap, model training, and inference are the three main components of the model. The first step in model training is to use an unsupervised algorithm, and the second step is to use a supervised algorithm. The model generates a network graph of the device during the data bootstrap process. The final stage is inference, where the system determines whether the account is a bot account. [15]Gives a detailed representation of the model.

III. SYSTEM OVERVIEW

Figure 1 depicts the overall system architecture for detecting fake user accounts. There are four stages:

- Data Collection
- Data Pre-Processing
- Various Classification algorithms
- Model's final performance, which determines if the respective user account is fake or legitimate.

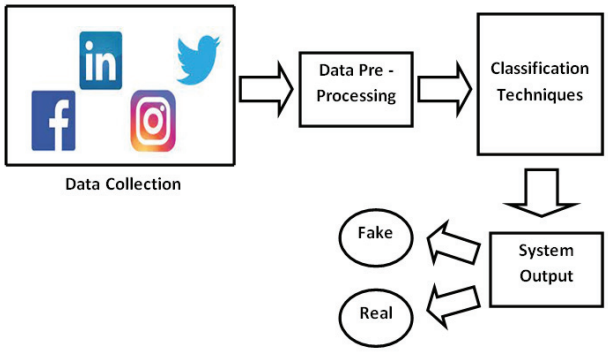

Fig. 1. System Overview

The first stage is to acquire data from users who are active on a particular social media platform. When it comes to research in this domain, the most widely used datasets are Twitter, Facebook, and LinkedIn user account info. Data is obtained from the Twitter API and the Facebook API in some research projects, although there are some limitations for collecting user data due to privacy policies. The data extracted from user profiles may not be available in good form or may be subject to restrictions that prevent us from using the data in the classification process. In this case, data pre-processing is used to prepare the extracted data for the classification stage. There are several data pre-processing steps that must be completed in order for the classification model to produce successful results.

For classifying the collected data as a fake profile or a genuine profile, various supervised and unsupervised classification algorithms are used. In the classification process, there is a first learning phase in which the model is trained on available data and then used to predict whether a user account is fake or genuine by considering data that is not present in the dataset. The prediction process, which is simply the model's testing phase. Support vector machine (SVM), Naive Bayes (NB), Random forest (RF), Neural networks (NN), Decision tree (DT), and K-nearest neighbors (KNN) are some of the general classification algorithms used. The flow chart in fig.2 represents the system's flow in predicting fake accounts on social media platforms.

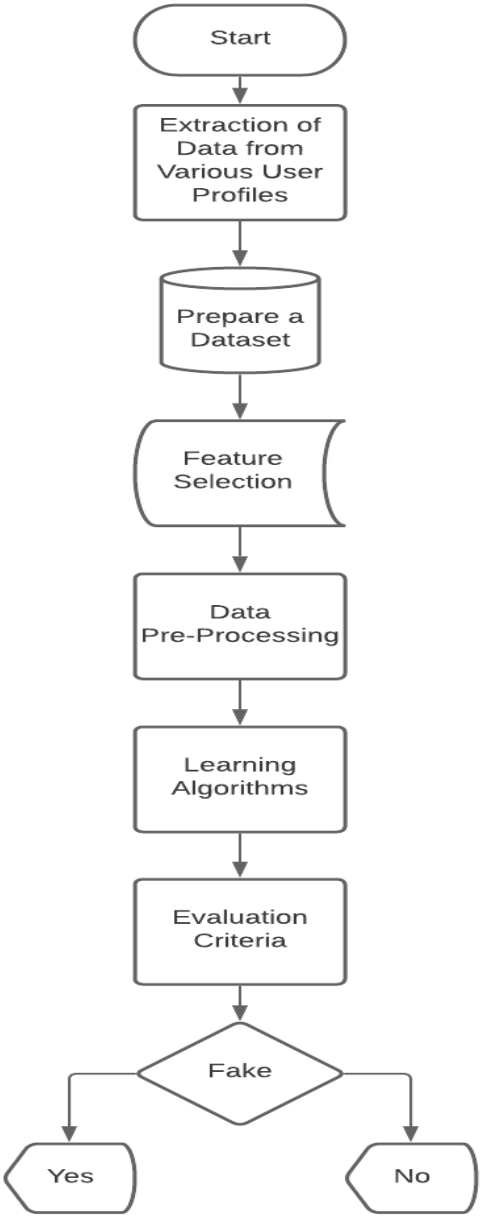

Fig. 2. General Flow of the System

The majority of the research is done with supervised and unsupervised algorithms on various datasets. Various feature selection methods are also used to extract important features from datasets in order to produce reliable results. Some researchers have proposed a hybrid method for improving accuracy by combining more than two algorithms. Mostly, confusion matrix is used as evaluation criteria for calculating the accuracy of the model.

IV. RESULTS

By using different machine learning supervised and unsupervised algorithms most of the malicious user accounts detected successfully. Malicious accounts include bot users, duplicate users, spam accounts, and false identities on social media platforms. The results of the computation model are calculated using following formula of accuracy with the help of confusion matrix. The confusion matrix, which is an evaluation criterion with parameters such as precision, F1-Score, and recall is used to measure the result of detecting the fake profile in most of the research papers. Table I gives comparative results from the research papers used in survey study.

TABLE I. ACCURACY COMPARISON

Dataset used	**Method**	**Malicious Account Type**	**Accuracy**
LinkedIn	Support vector machine	Account Fake	84%
Twitter	Naive Bayes	Fake Account	90.40%
Facebook	Random Walk Method	Fake Account	95%
Twitter	ANN	Spam Account	0.95 Precision rate
Twitter	Decision Tree and Naive Bayes	Spam Account	93%

V. CONCLUSIONS

Nowadays, social networking sites have evolved into a large medium for people to communicate with one another. In Advanced Persistent Threat situations, fake identities on social media are often used to spread malware or a connection to it. They're often used in other malicious activities including sending spam emails or falsely inflating the number of users in certain apps to promote them. We conclude that the paper presents some of the most important methods for detecting fake users on social media platforms. We can predict whether a user account is fake or real using a dataset that includes details about fake and genuine profiles. Patterns in the dataset are discovered using a variety of classification machine learning and deep learning algorithms. This paper also looks at some new terminology, such as hybrid approaches and random walk methods. The literature review is done with the user profile as clone, bot, spam or fake account identification taken into consideration.

REFERENCES

[1] N. C. Le, M. Dao, H. Nguyen, T. Nguyen and H. Vu, "An Application ˆ of Random Walk on Fake Account Detection Problem: A Hybrid Approach," 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), 2020, pp. 1-6, doi: 10.1109/RIVF48685.2020.9140749.

[2] M. A. wani, N. Agarwal, S. Jabin and S. Z. Hussain, "Analyzing Real and Fake users in Facebook Network based on Emotions," 2019 11th International Conference on Communication Systems Networks (COMSNETS), 2019, pp. 110-117, doi: 10.1109/COMSNETS.2019.8711124.

[3] Muthu, BalaAnand Natesapillai, Karthikeyan Subburathinam, Karthik Varatharajan, R. Manogaran, Gunasekaran Sivaparthipan, C B. (2019). An enhanced graph-based semi-supervised learning algorithm

to detect fake users on Twitter. The Journal of Supercomputing. 75. 10.1007/s11227-019-02948-w.

[4] P, Sowmya and Chatterjee, Madhumita, Detection of Fake and Cloned Profiles in Online Social Networks (March 9, 2019). Proceedings 2019: Conference on Technologies for Future Cities (CTFC), Available at SSRN:https://ssrn.com/abstract=3349673 or http://dx.doi.org/10.2139/ssrn.3349673.

[5] M. Conti, R. Poovendran and M. Secchiero, "FakeBook: Detecting Fake Profiles in On-Line Social Networks," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2012, pp. 1071-1078, doi: 10.1109/ASONAM.2012.185.

[6] I. David, O. S. Siordia and D. Moctezuma, "Features combination for the detection of malicious Twitter accounts," 2016 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), 2016, pp. 1-6, doi: 10.1109/ROPEC.2016.7830626.

[7] Adikari, S. Dutta, K.. (2014). Identifying fake profiles in linkedin. Proceedings - Pacific Asia Conference on Information Systems, PACIS 2014.

[8] S. Gheewala and R. Patel, "Machine Learning Based Twitter Spam Account Detection: A Review," 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), 2018, pp. 79-84, doi: 10.1109/ICCMC.2018.8487992.

[9] Raturi, Rohit. (2018). Machine Learning Implementation for Identifying Fake Accounts in Social Network.

[10] Wani, Suheel Yousuf Kirmani, Mudasir Ansarullah, Syed. (2016). Prediction of Fake Profiles on Facebook using Supervised Machine Learning Techniques-A Theoretical Model. International Journal of Computer Science and Information Technologies. 7 (4). 1735-1738.

[11] Venkatesan, Sridhar Albanese, Massimiliano Shah, Ankit Ganesan, Rajesh Jajodia, Sushil. (2017). Detecting Stealthy Botnets in a Resource-Constrained Environment using Reinforcement Learning. 75-85. 10.1145/3140549.3140552.

[12] 1P. Srinivas Rao, 2Dr. Jayadev Gyani, 3Dr.G.Narsimha, "Fake Profiles Identification in Online Social Networks Using Machine Learning and NLP", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13.

[13] Amit Pratap Singh, Maitreyee Dutta, "Spam Detection in Social Networking Sites using Artificial Intelligence Technique", International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-8S3.

[14] E. Van Der Walt and J. Eloff, "Using Machine Learning to Detect Fake Identities: Bots vs Humans," in IEEE Access, vol. 6, pp. 6540-6549, 2018, doi: 10.1109/ACCESS.2018.2796018.

[15] A. A. Daya, M. A. Salahuddin, N. Limam and R. Boutaba, "BotChase: Graph-Based Bot Detection Using Machine Learning," in IEEE Transactions on Network and Service Management, vol. 17, no. 1, pp. 15-29, March 2020, doi: 10.1109/TNSM.2020.2972405.

[16] A. M. Al-Zoubi, J. Alqatawna and H. Paris, "Spam profile detection in social networks based on public features," 2017 8th International Conference on Information and Communication Systems (ICICS), 2017, pp. 130-135, doi: 10.1109/IACS.2017.7921959.

[17] G. Kontaxis, I. Polakis, S. Ioannidis and E. P. Markatos, "Detecting social network profile cloning," 2011 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011, pp. 295-300, doi: 10.1109/PERCOMW.2011.5766886.

[18] Sahoo, Somya Ranjan Gupta, B B. (2019). Hybrid approach for detection of malicious profiles in twitter. Computers Electrical Engineering. 76. 65-81. 10.1016/j.compeleceng.2019.03.003.

[19] Joshi, Shruti Nagariya, Himanshi Dhanotiya, Neha Jain, Sarika. (2020). Identifying Fake Profile in Online Social Network: An Overview and Survey. 10.1007/978-981-15-6315-7-2