A Feature Based Approach to Detect Fake Profiles in Twitter

Jyoti Kaubiyal
National Institute of Technology, Kurukshetra
Thanesar, Haryana, 136119, India
+91 7895390896
jyotikaubiyal@gmail.com

Ankit Kumar Jain
National Institute of Technology, Kurukshetra
Thanesar, Haryana, 136119, India
+91 9455313000
ankit.jain2407@gmail.com

ABSTRACT
Social networking platforms, particularly sites like Twitter and Facebook have grown tremendously in the past decade and has solicited the interest of millions of users. They have become a preferred means of communication, due to which it has also attracted the interest of various malicious entities such as spammers. The growing number of users on social media has also created the problem of fake accounts. These false and fake identities are intensively involved in malicious activities such as spreading abuse, misinformation, spamming and artificially inflating the number of users in an application to promote and sway public opinion. Detecting these fake identities, thus becomes important to protect genuine users from malicious intents. To address this issue, we aim to use a feature-based approach to identify these fake profiles on social media platforms. We have used twenty-four features to identify fake accounts efficiently. To verify the classification results three classification algorithms are used. Experimental results show that our model was able to reach 97.9% accuracy using the Random Forest algorithm. Hence, the proposed approach is efficient in detecting fake profiles.

CCS Concepts
• Security and privacy→Social network security and privacy

Keywords
Fake Profiles; Social Media Platforms, Security, Machine Learning

1. INTRODUCTION
Social Media Platforms (SMPs) such as Twitter, Facebook, LinkedIn, Reddit, etc. provide space, for people around the globe, that share common personal, career interest to communicate and build social interests. They provide a place to share their ideas, interests, photos, and videos with other people. Over the years it has become a preferred means of communication and has attracted the interest of millions of users. As the user community grew, SMPs further became a space for industries to advertise their products, a way to spread news and information [1]. Just like everything can be used for good and evil purposes, there is another side to it [2]. Now SMPs are also being used by people with malicious intents to spread hatred, fake news, and misinformation [3].

BDIOT 2019, August 22–24, 2019, Melbourne, VIC, Australia

DOI: https://doi.org/10.1145/3361758.3361784

To target a greater audience, automated bots called Social Bots are being used [4]. Social bots were initially developed to automatically collect information from a variety of services and to provide automated responses. Although they were developed to provide useful service, they are today being used to spread misinformation and sway public opinion. A recent study done by USC and Indiana University found that between 29M and 48M accounts on Twitter are fake (or bots) [5]. Detecting these fake identities, thus becomes important to protect genuine users from malicious intents. The academia is intensively researching this domain to protect real users from these fake identities. Most of the approaches have been featured based on machine learning approaches. However, as the detection techniques have improved over the years, so does the spammers. They have started using the techniques which make their behavior resemble legitimate users.

The percentage of counterfeit accounts and automated bots on social media platforms has increased significantly as these social media has expanded. The effect of these bot accounts was seen during US presidential elections, where various reports suggested that Twitter bots shared Trump-related content in ratio 7:1 as compared to that of Hillary [5]. Various machine learning models have been proposed to detect bot accounts However, till date, there is no satisfactory solution found for this problem. In this paper, we intend to provide a user-end solution which may help users to protect themselves against these fake account and automated bots.

To deal with the issue, we intend to develop our machine learning model that not only uses classical feature-based approach but further incorporates mechanisms to detect these evasion techniques. Further, we also intend to do some analysis on users' tweet to further improve the accuracy of our model.

2. BACKGROUND
Whatever platform users may have been using, one thing is certain, plenty of people take social media safety less seriously than they should. They post, share, and retweet without any concern for their privacy. This is leading to new, potentially devastating attack vectors.
Spamming is one of the most common methods used among attackers. Fake accounts play a major role in spreading spam. To gain trust, the masquerader behind the fake accounts pursues the path of befriending and following a legitimate account [6]. When a good amount of legitimate accounts is in the list of friends of the fake accounts, it legitimizes the account and paves the way for

spammers to spread malicious content. Following methods are used by the fake accounts to spread spams on the social networking sites:

- Through bulk messaging: In bulk messaging, messages with a similar kind of text is sent to a bunch of people that can increase the visibility of a topic and make it a trending topic. It is used by companies for advertisements, but spammers can leverage this as an opportunity to spread malicious content.
- Through malicious links: Malicious links can direct a user to a fake website to steal his personal information which is named as phishing or to install the malware in the user's system. In social networking sites, these malicious links can be spread through tweets, personal messages, and comments.
- Through fraudulent reviews: This method is used by the companies who want to increase the sale of their product illegitimately. Fake profiles are made to write reviews for the products on e-commerce websites.
- Through likejacking: Likejacking is Facebook specific variation in which the attacker presents a webpage that consists of two layers. The layer on the back is designed with a Facebook "Like" button which is designed to follow the user's mouse cursor. The topical layer reveals the content with which the user can be lured. When the user clicks on that content, he actually clicks the like button. The more people like the post, the more it spreads.

3. RELATED WORK

Fake profiles are polluting Social networking sites and making them unreliable to use. They are massively being used to get information illegitimately, steal identity, defame someone, spread misinformation and malware, and boost follower counts for popularity. Therefore, Social networking brands are themselves taking a stern step against fake profiles. Researchers are also continuously trying to upgrade the techniques to find fake profiles through different approaches. For example, D. Ramalingam and V. Chinnaiah have given a model to detect fake profiles. They took a set of profiles from LinkedIn and firstly processed those profiles to extract features. After processing profiles using principal component analysis, a training module is developed using Resilient Back Propagation algorithm in a neural network. Support Vector Machines (SVMs) is used for classification of profiles [7].

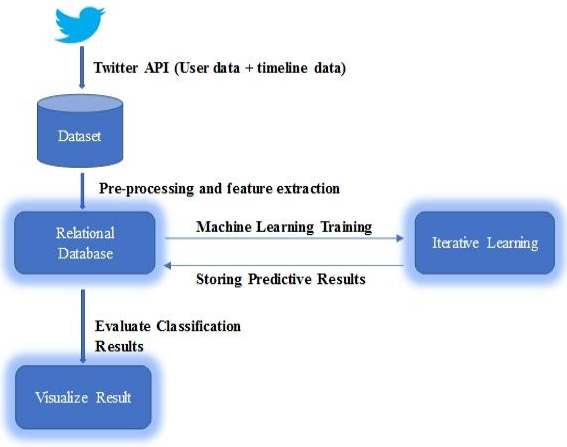

Figure 1. Architecture of proposed approach

Gao et al. [8] developed a framework named SybilFrame. It can detect a Sybil attack on Facebook and Twitter using multi-stage level classification. At stage1 the entire dataset is explored to extract information about the nodes and edges which is needed prior to calculation. At stage2 subsequent information is employed that correlates the nodes using Markov random field and loopy belief propagation. Independent sybils and the sybils that collaborate with identified sybils can be detected using Friend recommendation schemes [9].

Yang et al. [10] used a graph-based technique to detect fake accounts on OSNs. They leveraged user-level activities to classify benign accounts and fake accounts. Trust is calculated based on the votes for a friend request for being accepted or rejected on the nodes. The entire network through which the trust has been propagated is used as a basic criterion to detect sybils.

Amato et al. [11] Proposed a two-step detection method to identify human behavior in social networking sites. In step-1, the Markov chains is trained on the models of usual human behavior from the data available on social networks. In step-2, an activity detection framework is used to identify unusual activities by comparing them with usual behavior models.

Gong et al. [12] developed a supervised machine learning based model named DeepScan for detecting malicious account in location-based social networks. It leverages a deep learning algorithm to analyze the dynamic behavior of users. They also introduced the long short-term memory (LSTM) neural network to analyze time series from user activities.

Wang et al. [13] designed a framework, SybilBlind, to detect sybils. It is a structure-based framework that uses randomly assigned labels rather than relying on the manually labeled training set. To randomly sample a noisy dataset, they define a sampling trial and an aggregator is designed to aggregate the results in multiple sampling trials.

Yang et al. [14]reviewed different techniques that use artificial intelligence to combat bots in social media. They used a case study of Botometer, a bot detection tool, to illustrate algorithmic and interpretability improvements of bot scores, how people interact with AI countermeasures.

Kudugunta et al. [15]proposed a system to detect bot in social media accounts. An architecture based on a deep neural network that uses contextual long short-term memory is introduced in their model. They also proposed a technique that leverages synthetic minority oversampling to efficiently train deep networks using a small amount of labeled dataset.

Zhang et al. [16] designed a system, COLOR+, that can make spam account detection in mobile social networks easy. The technique also leverages fog computing. It makes the work of storing and calculating a local graph on a mobile device much easier. They have defined a threshold suspicion degree which classifies suspicious account from genuine accounts.

4. PROPOSED METHODOLOGY

To classify accounts as 'Bots' and 'Humans' supervised machine learning algorithms have been used in our model. Real twitter data is collected from different profiles using the Twitter API. After pre-processing the data, the supervised machine learning algorithm is applied for the classification of accounts. Three different machine learning algorithms, namely, Logistic regression, SVM, and Random Forest are applied to the same data and then compared. Figure 1 depicts the architecture of the proposed methodology.

4.1 Pre-processing and Feature Selection

The dataset obtained had raw data which required pre-processing to obtain useful information. To remove irrelevant and redundant data from the raw data feature selection is done. It is the most important process in machine learning since it improves learning accuracy, reduces computation time, and makes the data more relevant for learning model. A total of 24 features were identified and are shown in Table 2.

The details of the features are described below:

- Features based on the account: We can find malicious behavior of an account by analyzing account-based features. Fake accounts generally have a large number to followees than followers. They also tweet the same post after a definite interval of time in different profiles. So, the number of retweets in a fake profile is larger than a genuine user profile. Malicious users create a new account to commit attacks. Hence, the age of the account will be less for any malicious profile.
- Features based on tweets: In order to increase the visibility in twitter platform, adversaries use the tag in their tweets to attract their audience. They use currently trending topics in their tweets to attract more people. Twitter provides two types of tag to help users to address their views to their followers, hashtag (#) and mention-tags (@). A hashtag allows users to index topics on Twitter. It enhances the visibility of tweets and helps to categorize them. The mention-tag is utilized to highlight a user in tweets or in replies to messages. Thus, phishers use trending topics with their malicious tweets to increase the visibility of their tweets. The number of mentions is also large in malicious tweets as compared to legitimate tweets.
- Features based on ownership details: Ownership details helps to enhance the power of the model to detect malicious tweets containing malicious URL such as dates of domain creation that tells when the domain was created, registrar's name which tells the name of the domain provider, the period of URL i.e. for what period that URL is bought for since generally malicious URLs are procured for a short span of time. Adversaries generally create a domain just before they tweet.
- Features based on URLs present in the tweet: Sharing videos, pictures, and advertisements through URL on social networking sites is trending nowadays. It is also the most effective way to spread malicious content since users trust their friends and content they share. URLs of malicious websites have some distinctive features. For example, the number of dots in malicious URLs are more in number than legitimate URLs, subdomains are also more in number in malicious URLs than in legitimate URLs.

Table 1. Description of Features

	Feature	Description
Account-based features	Followers Count	The number of followers this account currently has.
	Friends Count	The number of users this account is following.
	Favourites Count	The number of Tweets this user has liked in the account's lifetime.
	Account-age	Time since the inception of the Twitter account
	Geo-Enabled	Indicates that the user has enabled the possibility of geotagging their Tweets.
	Friends-to-Followers ratio	Friends count / Followers count
Tweet-based features	Count of Retweets per Tweet	The total number of times a tweet has been retweeted by other users.
	Count of Favourites per Tweet	The number of times a particular tweet has been liked by other users.
	Count of Hashtags per Tweet	Count of hashtags in tweet sample "#topic"
	Count of Mentions per Tweet	Count of user mentions in tweet sample "@screenname"
	Count of URLs per Tweet	Count of URLs in tweet sample "http://www.url/"
	User Activity	Number of Tweets user posts per day since the account is created.
	Tweet Similarity	Percentage similarity between the tweets of all the users
Ownership-detail based features	Name of registering domain	Provider of the domain
	Period of ownership	Age of the domain
	The gap to create a twitter account	The amount of time elapsed since the initiation of the domain and the Twitter account
URL-based features	URL-length	Length of phishing URL is generally long
	Dots-count	No. of dots (.) are more in phishing URL
	Count of sub-domains	No. of subdomains in the phishing URL (marked by /) is more than one
	Dash count	Number of dashes (-) used in URL
	The header of the email	The path of the mail can be traced through the header of the mail. Fake emails take additional hops.
	SSL certificates	The age of certificate should be more than a year and it should be from a trusted issuer
Others	Presence of Iframe	Presence of Iframe in a website indicates that it is malicious
	Mouse hover	This feature checks if the address shown in the status bar is the same as the URL shown in the address bar

After pre-processing and feature selection the model is trained for the classification in which the dataset is separated into two distinct sets namely training set and test set. The training set is utilized to fit the parameters of the classifier. The test set is used to tune the architecture or hyperparameters of the classifier. In the testing phase, the trained classifier classifies the profile as fake and legitimate. Based on the classification result the model's accuracy is evaluated.

5. IMPLEMENTATION DETAILS

The main objective of our research is to find a technique that can detect fake profile in social media efficiently using machine learning algorithms. This section provides details regarding dataset used, pre-processing, feature selection, and the training models used in our experiment.

5.1 Dataset

We have used publicly available datasets [17] for the development of our classification model. It consists of real twitter user profile metadata and user timeline data that was collected using Twitter API [18]. Table 1 shows the statistics of the data that was used in the study. The dataset was divided into 80:20 ratios for training and testing purpose respectively. Train dataset was used to train the model and Test dataset was used to compute the accuracy of the model.

Table 2. Statistics of Dataset used in the study

Dataset	Description	Accounts	Tweets
Genuine Accounts	Verified accounts that are human operated	3,474	83,77,522
Social Spambots #1	Retweeter's of Italian political candidate	991	11,60,176
Social Spambots #2	Spammers for paid apps for mobile devices.	3,457	4,28,542
Social Spambots #3	Spammers of product on sale at Ecommerce sites	464	14,18,626
Fake Followers	Simple accounts that inflate the followers of other accounts	3,351	1,96,027

6. EVALUATION AND RESULTS

The following evaluation metrics were utilized to evaluate our models.

- Precision: It is the percentile of accounts that are fake and are correctly classified as fake. It gives the exact correctness of the model and is given by the equation

$$precision = \frac{\text{TP}}{\text{TP+FP}} \qquad (1)$$

- Recall: It is the rate of true positives that are correctly classified by the testing model. It is also called sensitivity and is given by the equation

$$recall = \frac{\text{TP}}{\text{TP+FN}} \qquad (2)$$

- F-Score: It is the harmonic average of recall and precision. It is given by:

$$F_1 = \frac{2*\text{precision}*\text{recall}}{\text{precision+recall}} \qquad (3)$$

- Weighted Accuracy (W-accuracy): It tells how often the test model is giving correct results and is given by:

$$w_{acc} = \frac{\lambda TN+TP}{\lambda\,(\text{TN+FP})+FN+TP} \qquad (4)$$

Where λ is the weight used to penalize the false positives since they are more costly than false-negative.

TP = True-positives; TN = True-negatives; FP = False-positives; FN = False-negatives

- ROC (Receiver Operating Characteristic) curve: It is the metrics used to evaluate the performance of a classification model. It tells the capability of a model to distinguish two classes.

The four metrics were selected after careful consideration of the essential parameters. Table 3 shows that our model is able to detect account as 'bot' and 'human' efficiently. Logistic regression and Random Forest were able to detect the fake accounts more efficiently with 95.3% and 97.9% accuracy respectively whereas SVM was having 80.8% accuracy. Precision and F-score of Logistic Regression and Random Forest was also better than the SVM with Logistic Regression having highest precision and F-score of 98%.

Table 3. Shows Results of Supervised Machine Learning Algorithm

Model	W-Accuracy	Precision	Recall	F-Score
Logistic Regression	95.7%	94	96	95
SVM	80.8 %	82	100	90
Random Forest	97.9%	98	98	98

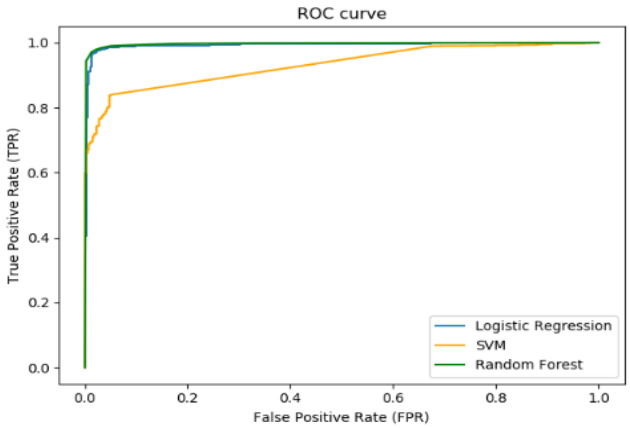

Figure 2. ROC Curve for different models

Figure 2 shows the comparative results between the three different models we used for our analysis. A curve is plotted between the True Positive rate and False Positive to see the ability of the model to classify the classes. Among all the three models Random Forest performed best and gave good results. It shows that Random Forest is able to classify the accounts more efficiently under the given circumstances and selected features.

The results clearly show that all the three models performed well in classifying accounts as "bots" and "human" with Logistic regression and Random Forest having better accuracy than SVM. After careful analysis of results one thing that was noticeable was that, out of all the wrong predictions, most of them were human operated fake accounts. While our model successfully detects

automated bot accounts, but human operated fake accounts were still hard to detect.

7. CONCLUSION AND FUTURE PLAN

Fake profiles have become a major concern in social networking sites and are difficult to detect. Hence, this paper provides a solution towards detecting fake profiles using features. The main objective of our model is to detect fake profiles using selected features efficiently. The classifier created during the training phase was able to detect fake profiles efficiently with 97.9% accuracy for the Random Forest algorithm.

In the future, we aim to further study the behavior of human-operated fake accounts in more depth and add more features to our existing model to get better performance in case of human-operated bot accounts. In addition to this, we also aim to incorporate tweet sentiment analysis to get better results. Finally, we aim to convert out windows application to a web browser-based extension that can perform real-time analysis of the twitter account and generates an alert when comes in touch with a bot account.

8. REFERENCES

[1] C. Shao, G. L. Ciampaglia, O. Varol, K.-C. Yang, A. Flammini, and F. Menczer, "The spread of low-credibility content by social bots," *Nat. Commun.*, vol. 9, no. 1, p. 4787, Dec. 2018.

[2] Z. Zhang and B. B. Gupta, "Social media security and trustworthiness: Overview and new direction," *Futur. Gener. Comput. Syst.*, vol. 86, pp. 914–925, 2018.

[3] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: Threats and solutions," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 4, pp. 2019–2036, 2014.

[4] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media," Sep. 2018.

[5] "Nearly 48 million Twitter accounts could be bots, says study."[Online].Available:https://www.cnbc.com/2017/03/10/nearly-48-million-twitter-accounts-could-be-bots-says-study.html. [Accessed: 02-Jun-2019].

[6] "5 Types of Social Spam (and How to Prevent Them)." [Online]. Available: https://thenextweb.com/future-of-communications/2015/04/06/5-types-of-social-spam-and-how-to-prevent-them/. [Accessed: 06-Jun-2019].

[7] D. Ramalingam and V. Chinnaiah, "Fake profile detection techniques in large-scale online social networks: A comprehensive review," *Comput. Electr. Eng.*, vol. 65, no. 3, pp. 165–177, 2018.

[8] P. Gao, N. Z. Gong, S. Kulkarni, K. Thomas, and P. Mittal, "SybilFrame: A Defense-in-Depth Framework for Structure-Based Sybil Detection," Mar. 2015.

[9] X. Ma, J. Ma, H. Li, Q. Jiang, and S. Gao, "ARMOR: A trust-based privacy-preserving framework for decentralized friend recommendation in online social networks," *Futur. Gener. Comput. Syst.*, vol. 79, pp. 82–94, 2018.

[10] Z. Yang, J. Xue, X. Yang, X. Wang, and Y. Dai, "VoteTrust: Leveraging Friend Invitation Graph to Defend against Social Network Sybils."

[11] F. Amato *et al.*, "Recognizing human behaviours in online social networks," *Comput. Secur.*, vol. 74, pp. 355–370, May 2018.

[12] Q. Gong *et al.*, "DeepScan: Exploiting Deep Learning for Malicious Account Detection in Location-Based Social Networks," *IEEE Commun. Mag.*, vol. 56, no. 11, pp. 21–27, 2018.

[13] B. Wang, L. Zhang, and N. Z. Gong, "SybilBlind: Detecting Fake Users in Online Social Networks Without Manual Labels," Springer, Cham, 2018, pp. 228–249.

[14] K. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots," *Hum. Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 48–61, Jan. 2019.

[15] S. Kudugunta and E. Ferrara, "Deep neural networks for bot detection," *Inf. Sci. (Ny).*, vol. 467, pp. 312–322, Oct. 2018.

[16] J. Zhang, Q. Li, X. Wang, B. Feng, and D. Guo, "Towards fast and lightweight spam account detection in mobile social networks through fog computing," *Peer-to-Peer Netw. Appl.*, vol. 11, no. 4, pp. 778–792, Jul. 2018.

[17] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The Paradigm-Shift of Social Spambots: Evidence, Theories, and Tools for the Arms Race."

[18] "UCI Machine Learning Repository." [Online]. Available: http://archive.ics.uci.edu/ml/index.php. [Accessed: 08-Apr-2019].