Fake Social Media Profile Detection

Umita Deepak Joshi[1], Vanshika[2], Tushar Rajesh Pahuja[3], Ajay Pratap Singh[4],
Dr. Gaurav Singal[4], and Smita Naval[5]

[1] College of Engineering, Pune, India
[2] Maharaja Surajmal Institute of Technology, Delhi, India
[3] Thadomal Shahani Engineering College, Mumbai, India
[4] Galgotias university, Greater Noida, India
[5] Assistant Professor, Bennett University, Greater Noida, India
[6] Assistant Professor, MNIT, Jaipur, India

Abstract. Social media like Twitter, Facebook, Instagram, LinkedIn, etc. are an integral part of our lives. People all over the world are actively engaged in it. But at the same time, it faces the problem of fake profiles. Fake profiles are generally human-generated or bot-generated or cyborgs, created for spreading rumors, phishing, data breaching, and identity theft. Therefore, in this article, we discuss a detection model, which differentiates between fake profiles and genuine profiles on Twitter based on visible features like followers count, friends count, status count and more by using various machine learning methods. We used the dataset of Twitter profiles, TFP and E13 for genuine and INT, TWT and FSF for fake accounts. Here we talk about Neural Networks, Random Forest, XG Boost, and LSTM. The significant features are selected for determining the authenticity of a social media profile. Further, the architecture and hyperparameters are discussed. Finally, the models are trained, and results are obtained. As a result we get output as 0 for real profiles and 1 for fake profiles. After a profile is detected fake it can blocked/deleted and cybersecurity threats can be avoided. The language used for implementation is Python3 along with all the required libraries like Numpy, Sklearn, and Pandas.

Keywords: Neural Network, Random Forest, XG Boost, Social Media, Fake profile

1 Introduction

Social media has become a vital part of our lives. From sharing attractive extravagant photographs to follow celebrities to chat with close and far away friends, everyone is active on social media. It is a great platform to share information and interact with people. But everything has a downside. As social media is footing a firm spot in our lives, there are instances where it has turned out to be a problem.

There are 330 million monthly active users and 145 million daily active users on Twitter. Facebook also adds about 500,000 new users every day and six new users every second. Loads of information are shared over twitter every single day. From hot trending topics to the latest hashtags and news to one's most recent trip, you get everything on Twitter. People react, like, comment, share their views, raise their opinions all through the 280 character limit. There are genuine issues that are discussed, yet sometimes there are rumors. These rumors lead to conflicts between different sections of society. The concern of privacy, misuse, cyberbullying [4], false information has come into light in the recent past. All these tasks are performed by fake profiles.

Fake accounts can be human-generated or computer-generated or cyborgs [2]. Cyborgs are accounts initially created by humans but later operated by computers.

Fake profiles usually get created in pseudo names and misleading and abusive posts and pictures are circulated by these profiles to manipulate the society, or to push anti-vaccine conspiracy theories, etc. Every social media platform is facing the problem of fake profiles these days.

The goal behind creating fake profiles is mainly spamming, phishing, and obtaining more followers. The malicious accounts have full potential to commit cyber crimes. The counterfeit accounts propose a major threat like identity theft and data breaching.These fake accounts send various URLs to people which when visited, send all the user's data to faraway servers that could be utilized against an individual. Also the fake profiles, created seemingly on behalf of organizations or people, can damage their reputations and decrease their numbers of likes and followers.

Along with all these, social media manipulation is also an obstacle. The fake accounts lead to the spread of misleading and inappropriate information which in turn give rise to conflicts.

These hoax accounts get created to obtain more followers too. Who doesn't want to be voguish on social media? To achieve a high figure of followers, people tend to find fake followers. Over all,It has been observed that fake profiles cause more harm than another cyber crime. Thus it is important to detect a fake profile even before the user is notified.

In this very context here, we talk about detecting fake profiles on Twitter. We deploy various machine learning models. The dataset of twitter profiles E13 and TFP for genuine, and INT, TWT, FSF for fake is taken into use. To combat the creation of fake profiles, common defenses are:

1. Methods such as user verification must be incorporated while creating accounts on social media.
2. To detect abnormal activities, user behavior analysis must be employed. Bot detection solution consisting of analyzation based on real-time AI will be beneficial.
3. An automated bot protection tool must be used.

As a technical contribution, we designed a multi-layer neural network model, a random forest model, an XG boost model, and an LSTM model. The mentioned models are supervised machine learning models.

Also, the LSTM classifies based on tweets; the result can be combined with a

convolution neural network in the near future [6].
The paper is organized into sections. The past researches, data pre-processing, methodology, experimental results, the accuracy of models, conclusion, and future work are described in order.

2 Related Work

Social media: A Boon or Bane, this question has always subsisted. And all companies have aimed at providing a platform with the least errors and better experience. Hence, every day new developments and updates are done. Seeing that not enough is done so far for the detection of fake human identities on social media platforms like Twitter, we looked toward past research addressing similar problems.
Some methods classified profiles based on the activity of the account, the number of requests responded, messages sent, and more. The models use a graph-based system. Some methods also aimed at identifying between bots and cyborgs.Some past researches are mentioned below.
If certain words appear in a message, then the message is considered spam. This concept has been used to detect fake profiles on social media. For the detection of such words on social media, pattern matching techniques were used. But the significant drawback of this rule is that with time there is the continuous development and use of new words. Also, the use of abbreviations like lol, gbu, and gn is becoming popular on Twitter.
Sybil Guard [13] developed in 2008 aimed at limiting the corrupting influence of Sybil attacks via social media. It had constrained random walk by every node and was based on the occurrence of random-walk interactions. The dataset used was Kleinberg's synthetic social network.
Along with Sybil guard, another approach called the Sybil limit was also developed around the same time. Like Sybil guard, it also worked on the assumption that the non-Sybil region is fast mixing. It worked on the approach of multiple random walks by every node. And ranking was based on the occurrence of tails of walk intersection.
Sybil-infer was developed in 2009. It made use of methods like greedy algorithm, Bayesian inference technique, and Monte Carlo sampling with the assumptions like the non-Sybil region is fast mixing, and random walks are fast-mixing. The selection technique is threshold based on probability.
Mislove's algorithm,2010 worked on the Facebook dataset using greedy search and selected profiles based on metric normalized conductance.
In 2011, came a new model named facebook immune system that used random forest, SVM, and boosting techniques. It also used the Facebook dataset, and the feature loop was the selection technique.
An algorithm is used by Facebook to detect bots based on the number of friends which could be either related to tagging or relationship history. The rules stated above can identify bot accounts but are not successful to identify fake accounts created by humans. Unsupervised ML was used for detecting bots. In this tech-

nique instead of labeling, information was assembled based on closeness. The bots were recognized by grouping functions so admirably because of co-attributes.
Sybil rank [1][13] designed in 2012, is based on a graph-based system. The profiles were ranked based on interactions, tags, wall posts. The profiles with a high rank are labeled as real profiles and the ones with lower as fake. But this method was unreliable as there were instances where a real profile was ranked low.
Next, there was another model developed called the Sybil frame. It used a multistage level classification. It worked in two steps, firstly on a content-based approach and secondly on a structure-based method.
Filtering is also among one of the past approaches. A new threat or malicious activity is detected, and the account is added to the blacklist. But as far as human-fake accounts are concerned they tend to adapt and yet somehow avoid the blacklist.
Researches were also done to detect fake accounts based on factors like engagement rate and artificial activity. An engagement rate is the percentage of the interaction of the audience with a post. The engagement rate is calculated as (Total number of interactions/ Total number of followers) x100. These interactions could be in the form of likes, shares, or comments.
Artificial activity is based on the number of shares, likes, and comments made by a particular account. Insufficient information and the status of verification of email are also considered as an artificial activity.
In our model, we used a multi-layered neural network, random forest [9] approach, and XG Boost that work on the visible features of a profile. These extracted features are stored in a comma-separated file(CSV) that is easy to read by the model. Finally, after all, training, testing, and evaluating the model can label a profile as legitimate or not. We trained our models on Google Colab because Google provides the use of free GPU. The Google Colab 12GB NVIDIA Tesla K80 GPU that can be used up to 12 hours continuously. All the models were coded down in Python3.

3 Methodology

For the detection of fake Twitter profiles, we incorporated various supervised methods, all with the same goal yet different accuracy. Each model detects a fake profile based upon visible features only.
All these supervised models are fed the same dataset, and corresponding accuracy and loss graphs are plotted. Also, a comparison graph of the accuracy of different models is indicated. The models are trained using appropriate optimizers, loss functions, and activation functions.
The models used, mentioned below.

3.1 Pre-processing

Before proceeding for the models, we append one more stride i.e pre-processing. The data set is pre-processed before it is fed to a model. Our model aims at detecting a profile as a hoax or legitimate based on the visible characteristics. Henceforth, all the precise aspects are determined. Only the numerical data has been selected and the categorical features are discarded. The following traits are picked [10]:

friends	***followers***	***status count***	***listed count***	***fav count***	***geo enabled***	***lang num***

Then the data set of fake and genuine users are merged into one with an additional label for each profile i.e "isFake" that is a Boolean variable. It is then stored in the Y variable that is the response concerning a profile X. Finally the blank entries or NAN are substituted with zeros.

3.2 Artificial Neural Network

Neural networks [8] are the deep learning models that work similarly to the neuron network of a human brain. The neural network has layers, and each layer has neurons (nodes). We used the sequential from Keras. The model design with an input layer, three hidden layers, and an output layer has activation function ReLU for all but the output layer. Sigmoid, used as an activation function for the output layer. The model compiled using optimizer: Adam, loss function: binary cross-entropy. In our model, ANN of the stated above architecture is used. Sigmoid function finally provides the output between 0 and 1 and based on the prediction of a particular profile, labeled as fake or genuine.

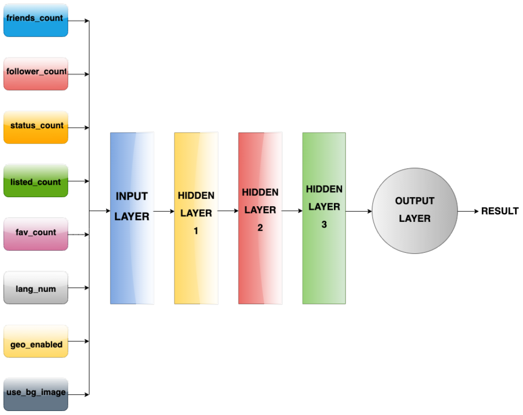


Fig. 1. ANN Architecture

Hyper parameters

- Rectified Linear Units (ReLU): Rectified Linear activation function is a piecewise linear function. ReLU (Fig. 2) is the default activation function for many neural networks as it is easier to train and produces better results .

$$R(z) = max(0, z) \tag{1}$$

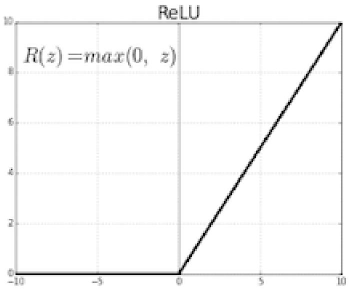


Fig. 2. Rectified Liner Units.

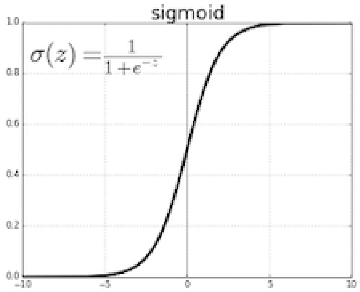


Fig. 3. Sigmoid Function.

- Sigmoid Function: This is also known as the logistic function. When values between 0.0 and 1.0 are required sigmoid function(Fig. 3) is used. It is a non-linear activation function and is differentiable and hence slope can be found at any two points.

$$\sigma(z) = \frac{1}{1 + e^{\text{-z}}} \tag{2}$$

  - If z is very large then e^{z} is close to zero and

$$\sigma(z) = \frac{1}{1 + 0} \approx 1 \tag{3}$$

  - If z is very small then e^{z} is large and

$$\sigma(z) = \frac{1}{1 + largenumber} \approx 0 \tag{4}$$

3.3 Random Forest

Random-forest also known as random-decision-forest is one of the methods that correspond to the category ensemble learning methods. This method is used in machine learning due to its simplicity in solving regression problems as well as classification.

Random-forest, unlike the decision tree method, generates multiple decision trees, and the final output is collectively the result of all the decision trees formed.

Similarly, we deployed the random forest [9] method for profile detection. The data is fed to the model and corresponding outputs are obtained. While training, the bootstrap aggregating algorithm is applied for the given set of X = x_1, x_2.......x_n and Y = y_1, y_2.......y_n responses, repeatedly (B times) random sample is selected and fits the trees(f_b) to the sample. After training the predictions for a given sample(x') is calculated by the formula specified below:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} f_{\mathrm{b}}(x') \tag{5}$$

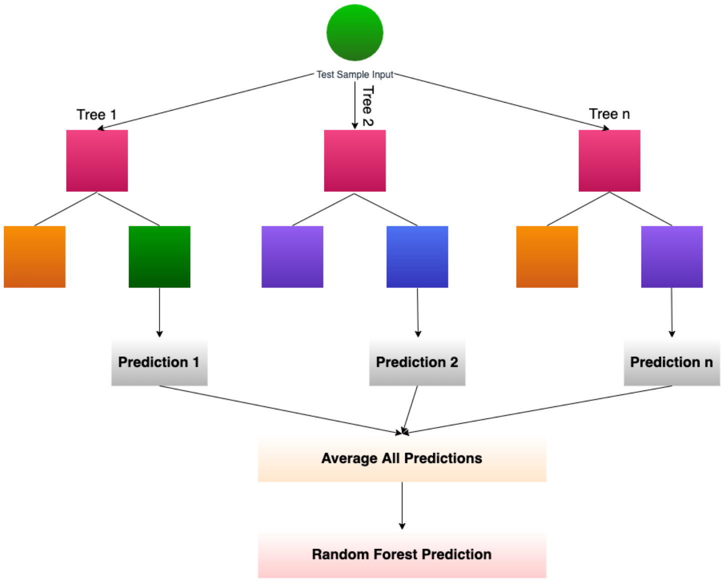


Fig. 4. Random Forest Architecture

3.4 Extreme Gradient Boost

XG Boost is another ensemble learning method used for regression. this implements the stochastic gradient boosting algorithm.

Random forest has a drawback, it is efficient only when all inputs are available i.e there is no missing value. To overcome this we use a gradient boosting algorithm.

As per the boosting algorithm, firstly, F_0(x) is initialized.

$$F_0(x) = argmin_\gamma \textstyle\sum_{i=1}^{n} \; L(y_\mathrm{i}, \gamma) \qquad (6)$$

Then iterative calculation of gradient of loss function takes place

$$r_\mathrm{im} = -\alpha[\frac{\partial L(y_\mathrm{i}, F(x_\mathrm{i})}{(F(x_\mathrm{i})}] \tag{7}$$

Finally the boosted model F_m(x) is defined

$$F_\mathrm{m}(x) = F_\text{m-1}(x) + \gamma_\mathrm{m} h_\mathrm{m}(x) \tag{8}$$

α is the learning rate
γ_m is the multiplicative factor

3.5 Long short-term Memory

LSTM is a recurrent neural network architecture. This architecture is capable of learning long-term dependencies.
In our project, we have developed a model using LSTM that classifies the profile as fake or genuine based on tweets. Before training the LSTM on the tweets, we pre-processed the data by forming a string of tokens from each tweet.
1. We have converted all tokens into lower case.
2. We have removed the stop words from tweets.
Then, we have transformed these tokenized tweets into an embedding layer to create word vectors for incoming words. The resulting sequence of vectors is then fed to the LSTM that outputs a single 32- dimension vector that is then fed forward through sigmoid activated layers to give the output.

4 Experimental Results

4.1 Dataset

We used the dataset available on MIB [17]. The data set consisted of 3474 genuine profiles and 3351 fake profiles. The data set selected was TFP and E13 for genuine and INT, TWT and FSF for fake accounts. The data is stored in CSV file format for easy reading by the machine.
All the labels on the x-axis depict the features used for the detection of the fake profile. These got selected during the pre-processing. The y-axis depicts the number of entries corresponding to each feature available in the dataset.

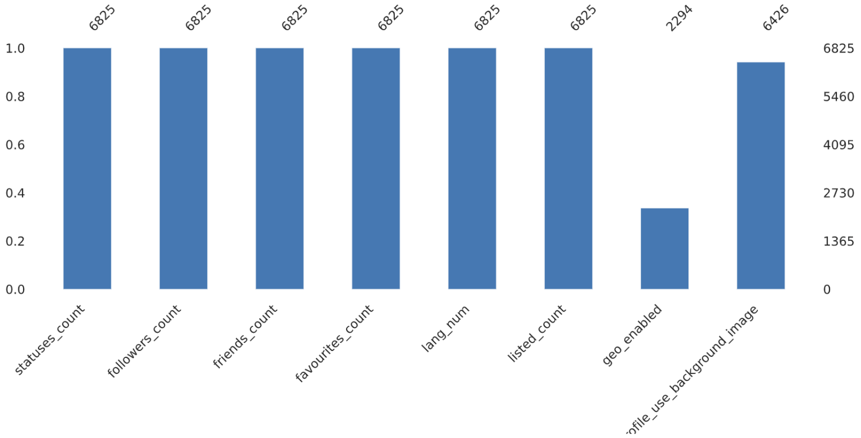
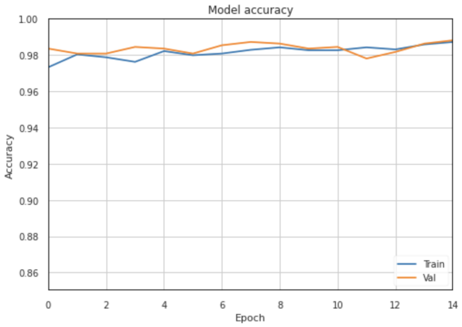


Fig. 5. Data set

4.2 Graphs and Charts

After training and testing all the models, the following results were obtained. The model accuracy, model loss vs the epochs graphs are plotted for neural network LSTM, and model accuracy comparison, and ROC curve for random forest, XG boost and other methods.

Neural Network: The model accuracy graph and model loss graph for the trained neural network are as follows:



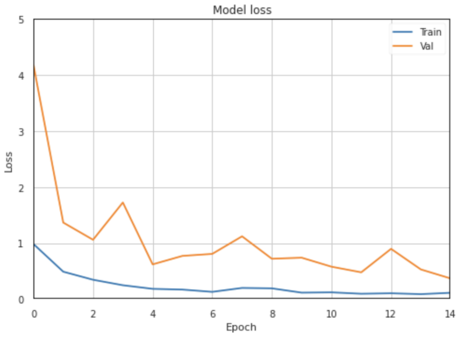


Fig. 6. Model Accuracy

Fig. 7. Model loss

After running for 15 epochs the above accuracy and loss graphs are obtained. Initially, starting from 0.97 the accuracy varies along the path and finally reaches it's maximum i.e 0.98. Similarly, the loss graph for testing data begins from 1, and for validation data begins from 4 and eventually reaches a minimum point, less than 0.5.
To calculate the loss Binary cross-entropy function is used. Initially, random weights get assigned to each feature and finally the machine defines a unique weigh to each feature.

Random Forest and other methods: In the comparison chart below we observe accuracy of different models namely random forest, xg boost, ada boost, and decision tree.
The maximum accuracy is achieved by XG boost that equals to 0.996. Further we have decision tree and random forest with approx similar accuracy of 0.99. At last we have ADA boost.
Histogram for accuracy comparison and the ROC curves are as follows::

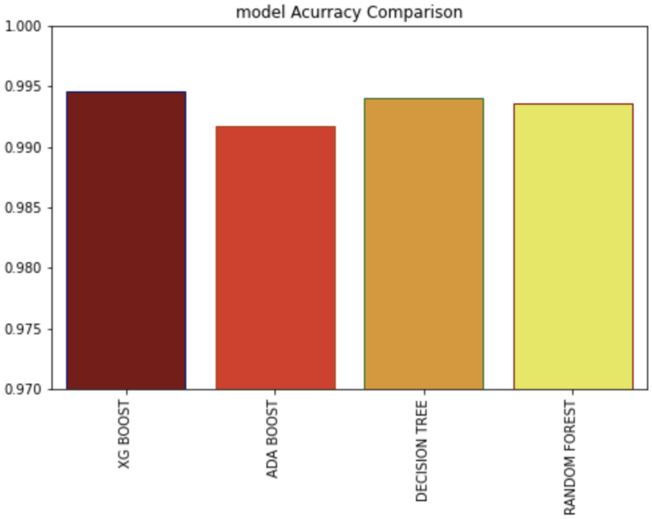


Fig. 8. Accuracy of Different Models

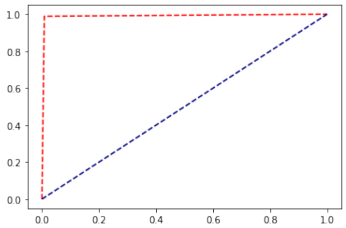


Fig. 9. ROC curve XG Boost

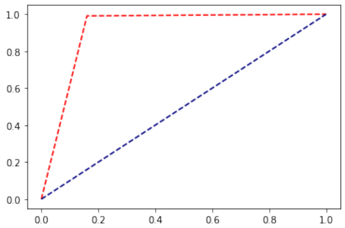


Fig. 10. ROC cruve Random Forest

5 Conclusion

In this design, we implemented the Neural Network, Random Forest, and XG Boost machine learning methods to train our system to detect fake Twitter profiles based on visible data. After training, validating, and testing our models on the MIB data set we finally arrive at an inference that the maximum accuracy achieved is 99.46% by XG Boost method followed by ANN and random forest. Further work can be done by combining images of profiles along with the categorical and numeric data and implement using a CNN. Also, including other parameters, combining different models, and assembling a real-time model may achieve better results.

Acknowledgment

We extend our sincere gratitude to Smita Naval and Dr. Gaurav Singal for their patient guidance, and useful critiques for this work. Their enthusiastic encouragement helped us to keep the progress on schedule.

References

1. Gergo Hajdu, Yaclaudes Minoso, Rafael Lopez, Miguel Acosta, Abdelrahman Elleithy: Use of Artificial Neural Networks to Identify Fake Profiles.
2. Estée Van Der Wal: Using Machine Learning to Detect Fake Identities: Bots vs Humans.
3. Aleksei Romanov, Alexander Semenov, Oleksiy Mazhelis and Jari Veijalainen: Detection of Fake Profiles in Social Media.
4. Yasyn ELYUSUFI, Zakaria ELYUSUFI, M'hamed Ait KBIR:Social Networks Fake Profiles Detection Based on Account Setting and Activity
5. Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, Maurizio Tesconi: Fame for sale: Efficient detection of fake Twitter followers.
6. Sneha Kudugunta, Emilio Ferrara:Deep Neural Networks for Bot Detection
7. Rohit Raturi: Machine Learning Implementation for Identifying Fake Accounts in Social Network August International Journal of Pure and Applied Mathematics (2018)

8. M. Likitha, K. Rahul, A. Prudhvi Sai, A.Mallikarjuna Reddy: Design and Development of Artificial Neural Networks to Identify Fake Profiles.
9. https://www.irjet.net/archives/V6/i12/IRJET-V6I12189.pdf
10. Yasyn Elyusufi, Zakaria Elyusufi, Aït Kbir M'hamed: SocialNetworksFakeProfilesDetectionUsingMachineLearningAlgorithms https://www.researchgate.net/publication/339012245/
11. Naman Singh, Tushar Sharma, Abha Thakral, Tanupriya Choudhury: Detection of Fake Profile in Online Social Networks Using Machine Learning in International Conference on Advances in Computing and Communication Engineering.(2018)
12. Jari Veijalainen, Aleksei Romanov, and Alexander Semenov:Revealing Fake Profiles in Social Networks by Longitudinal Data Analysis.
13. Devakunchari Ramalingam , Valliyammai Chinnaiah:Fake profile detection techniques in large-scale online social networks: A comprehensive review.
14. Kharaji MY, Rizi FS: An IAC approach for detecting profile cloning in online social networks. (2014)
15. Yu H, Gibbons PB, Kaminsky M, Xiao F. Sybillimit: A near-optimal social network defense against sybil attacks. In: IEEE symposium in security and privacy, (2008)
16. Fire M, Goldschmidt R, Elovici Y: Online social networks: threats and solutions. IEEE Commun Surv Tut9 (2014)
17. Sarah Khaled, Hoda M. O. Mokhtar, Neamat El-Tazi
18. Mulamba D, Ray I, Ray I. SybilRadar: A graph-structure based framework for sybil detection in on-line social networks. In: Proceedings of IFIP in- ternational information security and privacy conference.(2016)
19. Conti M, Poovendran R, Secchiero M.: Fakebook- Detecting fake profiles in on-line social networks. In: Proceedings of the international conference on advances in social networks analysis and mining, (2012)
20. Wang G, Jiang W, Wu J, Xiong Z: Fine-grained feature-based social influence evaluation in online social networks. IEEE Trans Parallel Distrib Syst (2014)
21. Silvia Mitter, Claudia Wagner, and Markus Strohmaier: A categorization scheme for socialbot attacks in online social networks. In Proc. of the 3rd ACM Web Science Conference (2013).
22. Norah Abokhodair, Daisy Yoo, and David W McDonald. Dissecting a social botnet: Growth, content and influence in Twitter. In Proc. of the 18th ACM Conf. on Computer Supported Cooperative Work Social Computing (2015).