Popularity-Based Detection of Malicious Content in Facebook Using Machine Learning Approach

Somya Ranjan Sahoo and B. B. Gupta

Abstract In this world, people are encircled with various online social networks (OSNs) or media platform, various websites and applications. This brings media contents like texts, audio, and videos in daily basis. People share their current status and moments with their belongings to keep in touch by using these tools and software like Twitter, Facebook, and Instagram. The flow of information available in these social networks attract the cybercriminals who misuse this information to exploit vulnerabilities for their illicit benefits such as stealing personal information, advertising some product, attract victims, and infecting user personal system. In this paper, we proposed a popularity-based method which uses PSO-based feature selections and machine learning classifiers to analyze the characteristics of different features for spammer detection in Facebook. Our detection framework result shows higher rate of detection as compared to other techniques.

Keywords Online social networks · PSO · Facebook · Machine learning

1 Introduction

In day-to-day life of human being, online social network or media become an integral part for sharing of knowledge, thoughts, and personal communication among belongings and friends. Social networks like Facebook, Twitter, Instagram, and other media-related networks are used by teenagers frequently in their work schedule. The popularity of networks leads to get benefited by posting certain advertising, blogs, and posts. Leading industries realize the usefulness of OSN sites for brand management and directly communicating with users for their benefits. Therefore, online social networks are the origin of user's personal information and commercial content

S. R. Sahoo · B. B. Gupta (✉)
Department of Computer Engineering, National Institute of Technology, Kurukshetra,
Kurukshetra, Haryana, India
e-mail: gupta.brij@gmail.com

S. R. Sahoo
e-mail: somyaranjan.sahoo@gmail.com

A. K. Luhach et al. (eds.), *First International Conference on Sustainable Technologies for Computational Intelligence*, Advances in Intelligent Systems and Computing 1045,
https://doi.org/10.1007/978-981-15-0029-9_13

that can be used in many ways. Due to the flooding on content like user information, photos, blogs, and other relevant information, there is the chance for cyberattack and it degrades the performance and user experience and provides negative impact at server-side activities like mining of the data and behavioral analysis of the user. To provide the security and managing the quality of social interaction is an increasing challenge.

At present, several malicious contents spread by malicious users and different threats have been reported [1–4]. One of the most vernacular problems in online social network is social spam bots and spammers. Spammers spread malicious content in the form of link, hashtag, and fraudulent content among legitimate users. Spam are propagated by social spammers in the form of advertising, viruses in the form of link, phishing contents, and form of fake accounts in many pages to attract people. The spam contents are spread in autonomous fashion and act like machine to steal personal information of the user. Most common category of pages that attract the user to visit and spread the content by like and share are Government portal, athletic sites, public figures, TV and shows, artist contents, and various company-related matters. Due to the above concern, it becomes extremely essential to figure a method/framework that identifying the different activities of the user in social platform to detect the spammers and the characteristics. This paper presents an approach which can detect spammers in social network sites like Facebook and the behavior of account based on the popularity-based machine learning model.

The rest of the paper contains the following main contributions. In Sect. 2, we describe the various state-of-the-art techniques for prevention and detection of spammer contents in Facebook. In Sect. 3, we describe the proposed model that detects the spammer in Facebook by analyzing various posts by the user based on collected dataset by our crawler. We studied various features' analysis based on machine learning classification techniques and comparative analysis describes in Sect. 4. In Sect. 5, we conclude our work with future research direction.

2 State-of-the-Art Techniques for Detection and Prevention

Spammer detection in online social network becomes a trending topic in various fields like industries and academicians. Many methods/frameworks are developed to identify and trace spammers on Facebook accounts in current trends, including machine learning-based feature analysis, optimization-based method, and graph-based analysis. Sometimes various researchers and antivirus development companies built various algorithms and software to detect spammers in OSN. The different service providers also inbuilt certain detection technique to control spamming activity accounts in their end. Facebook reports several behavior of the content that defines spamming activity in various forms.

- Sending harmful messages in the form of links including malware and phishing sites.

- For attention toward specific accounts, use follower or following.
- Sending unwanted messages with @ and # to the users.
- Creating fake account to gain the credentials of the user accounts.
- To grab the attention of the user sending repeated messages.

Sami et al. [5] designed a framework to detect the fake accounts in Facebook by applying machine learning technique. The framework trained Bayesian classifier to identify fake accounts by identifying key traits in the user account. Sohrabi et al. [6] proposed one framework that selects the different features of the profile and analyzed the content through supervised and unsupervised learning method to detect spam content in the Facebook platform. The selection of the features are based on the PSO algorithm to analyze the content in the Facebook and generate the specific content. Campos et al. [7] proposed an algorithm to classify the account as a human-made, legitimate robot or a malicious account generated by bot in social network. The algorithm based on the concept of discrete wavelength transforms to identify the writing content in the post. They use two different datasets for their analysis to detect malicious content accounts. Gurumurty et al. [8] proposed a FrAppE tool using SVM machine learning classifier to detect the malicious content present in the account. The tool detects the malicious content that is generated from a source but the source is not the legitimate one. Zhou et al. [9] analyze the spam content in the social network platform based on the different perspectives like viability of the account, different sequence of the transaction, and the correlation between the accounts. The detection rate of spammer in this technique is quite impressive as comparison to the other techniques. Talukder et al. [10] introduced one the technique base on the user questionnaires called ‘AbuSniff’ to detect and defense Facebook accounts against spammers and abusive friends. They impose certain questions based on the profile activity to detect the account as abuse or not by imposing supervised learning algorithms. Wang et al. [11] introduced Katz similarity-based unsupervised method, semi-supervised method and graph embedding method to detect multiple accounts in OSN. The detection of Facebook accounts content is analyzed in the form of groups. The malicious activity contents that are present in the account can be detected with the cluster form approach. Dewan et al. [12] proposed a technique called Facebook Inspector (FBI) to detect malicious account online. This technique works on empirical finding upon pre-trained model that captures the difference between malicious and benign post. Such techniques are efficient to detect the malicious content and the account which they have seen in the past.

However, none of the techniques classify the different category of post shared by the user in their time line that content malicious link or post. For efficient detection of malicious content in Facebook, our method utilizes various features that are associated with the content. This approach eliminates the dependencies in the post-similarity. The accuracy of detection of malicious content is more as comparison to other detection system.

3 Proposed Method/Model

The overall procedure of our spam detection model based on the popularity of content and PSO-based selection of features described in Fig. 1. We first collected different datasets based on the profile content that shared by the users in Facebook. Then PSO-based extracted features are processed with the data contents. After that, based on the selected features, we process the content with the help of supervised learning to detect the spammer activity. In the following, we introduced our dataset and the extracted features concept and the targeted result what we are looking for.

3.1 Data Collection

For our experiment, we collected data shown in Table 1 from various profiles based on the popularity of the contents like job alert, athletes post, public figure, shows related to TV or stage performances, politician blogs, artistic related contents, various product advertising, and fresh news-related contents based on the popularity. We

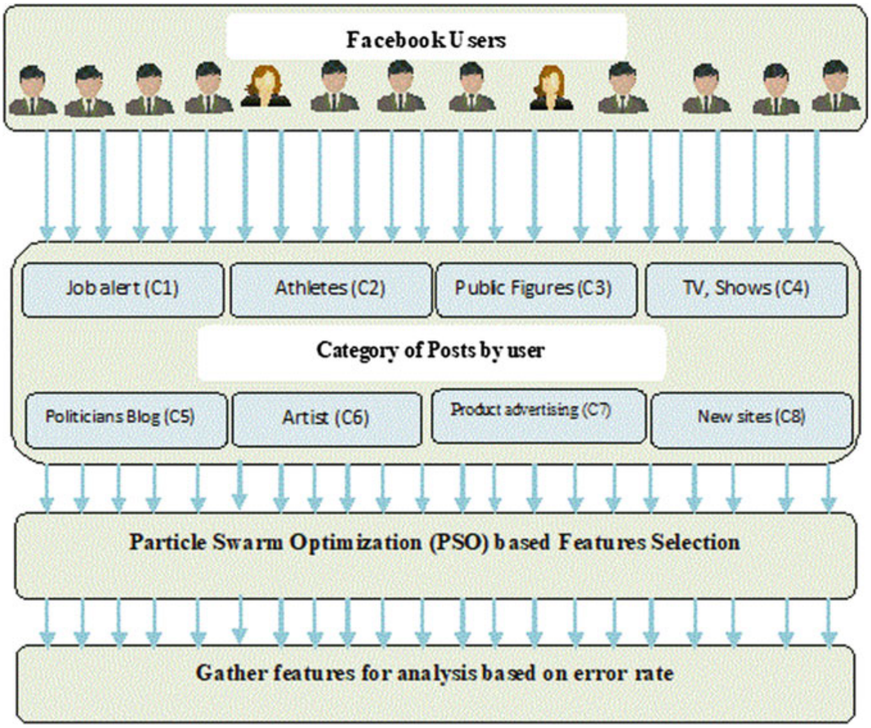


Fig. 1 Proposed model for data collection and feature extraction

Table 1 Collected dataset based on the category

Category of pages	Total number of profile pages	Mutual likes between page contents	Average mutual likes per profile
Politicians blogs	5908	41,729	7
Job alert	14,113	52,310	4
Artist	50,515	819,306	16
New sites	27,917	206,259	7
Product advertising	7057	89,445	12
TV, Shows	3892	17,256	4
Public figures	11,565	67,114	5
Athletes	1386	86,858	62

calculated the popularity index based on Eq. (1).

$$\text{Popularity Score (PS)} = \frac{\text{Number of Likes } (N^{\text{Likes}})}{\text{Number of Followers } \ (N^{\text{Followers}})} \tag{1}$$

3.2 Preprocessing and Collected Features

To get the resultant more exact, preprocessing is the steps to modeling the data is an inevitable part of data analysis. This process helps to find missing place values and replace those places with some values used in different features and normalized the content by removing useless attributes in the dataset. After preprocess the generated content, we selected certain profile based and content based features for observation and these features are the input parameter of the PSO algorithm. The details of collected features and their uses describes in Table 2.

We gather other profile related information's like profile name, profile image, cover photo, date of join and other relevant information by our crawler. The various content generated based on the features are provided as input parameter to the PSO algorithm for better decision making toward selecting best suitable feature for our operation to detect spammer accounts based on the profile content shown in Table 2. We also calculated fraction of message with URLs, replies, message with spam words etc.

3.3 Feature Selection Based on PSO Algorithm

The feature selection of the proposed system based on the particle swarm optimization (PSO) algorithm is described in algorithm 1. PSO is a self-adaptive stochastic

Table 2 Collected dataset based on the category

Features	Description	% as compared to features	
		Spammer	Legitimate
Friends	The peoples those who are connected with each other directly and their presence is in the list of each other consider a friend in the online social network	Less friends	More friends
Following	It is a feature by which the user can see the contents of a profile without being a friend	More	Less
Followers	If the user allows other users to view their contents that means the user is your follower. The user can see the blogs and post shared by you in their timeline	Less	More
Like	People put their positive opinion regarding a post such as image, video, or certain comments. By this phase, the owner of the account notices the number of friend visible the content in their wall post	Less	More
Comment	It is similar to the feature likes with opinion as an explanation regarding the content. It is visible only to the friends of their account not to others in the network	More	Medium
Reply to comment	Spammer uses this feature to reply a lot to attract the user toward their content further	More	Less
#tag	Facilitate toward identifying a specific subject #tag is used. Spammer uses many number of hashtag in a single post or comments $F_{\text{fbhashtag}} : \left\{\sum_{i=1}^{N} \text{msg}_{\text{hashtag}_i}\right\}$ (1) (No. of message with one hashtag) $F_{\text{hashtagfb}} : \left\{\sum_{i=1}^{N} \text{hashtag}_{\text{msg}_i}\right\}$ (2) (No. of hashtag in each message)	More	Less
URLs	These are the hyperlinks toward a specific page in the network by redirecting the users. Spammer uses the URL feature to spread malicious content in the network. A message with URL is the high probability of spreading malicious content $F_{\text{fbURL}} : \left\{\sum_{i=1}^{N} \text{msg}_{\text{URLs}_i}\right\}$ (3) (No. of message with one URLs) $F_{\text{URLfb}} : \left\{\sum_{i=1}^{N} \text{URL}_{\text{msg}_i}\right\}$ (4) (No. of URL in each message.)	More	Less

(continued)

Table 2 (continued)

Features	Description	% as compared to features	
		Spammer	Legitimate
Share	For sharing the content like photograph, video, and other advertising content in the network user uses the share feature in the Facebook to attract the people toward their accounts	More	Less
Messages with spam words	The content share in the social network for communication and content sharing in the form of text, image, or video is called message. The spammers always share certain content to attract the user toward their account. The spammer always used certain hashtag and URL associated with that content to redirect the user to other pages $F_{\text{fbspamword}} : \left\{ \sum_{i=1}^{N} \text{msg}_{\text{spamword}_i} \right\}$ (5) (No. of spam word in a single message)	More	Less
Stories	It is the visual information in the form of news feed	Less	More
No. of Group joined	For content gathering and sharing, people join different groups on their interest. Spammers joined more group as compared with legitimate one for attracting users toward their accounts	More	Less

optimization technique based on the idea of simulation modeling for searching of foods of a swarm of birds. In this process, each particle in the swarm "*i*" moved toward the optional point by using addition with a velocity factor with its own value. The velocity is influenced with various factors like different post, URLs, media content, number of hashtag, feedback to the comments, etc. The component called inertial simulates the activity of the features to move toward the same direction used previously. The features are moved around multidimensional search until find the solutions. By using the above discussion, the PSO for feature selection is as follows.

$$V_{ij}^{k} = w * C^{1} * r^{1} * \left((P_{\text{best}})_{ij}^{k-1} - X_{ij}^{k-1} \right) + C^{2} * r^{2} * \left((G_{\text{best}})_{i}^{k-1} - X_{ij}^{k-1} \right), \quad (2)$$

where

$i = 1, 2, 3, \ldots N_{\text{Attribute}}$
$j = 1, 2, 3, \ldots N_{\text{Particle}}$

The equation for the position update is

$$X_{ij}^{k} = X_{ij}^{k-1} + V_{ij}^{k} \quad (3)$$

where

$i = 1, 2, 3, \ldots N_{\text{Attribute}}$
$j = 1, 2, 3, \ldots N_{\text{Particle}}$

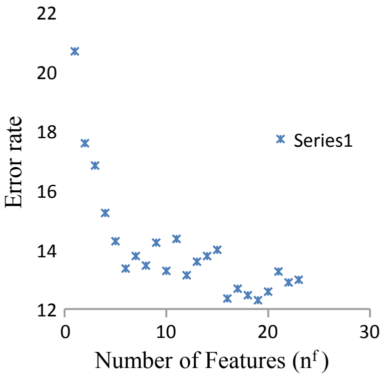
And,

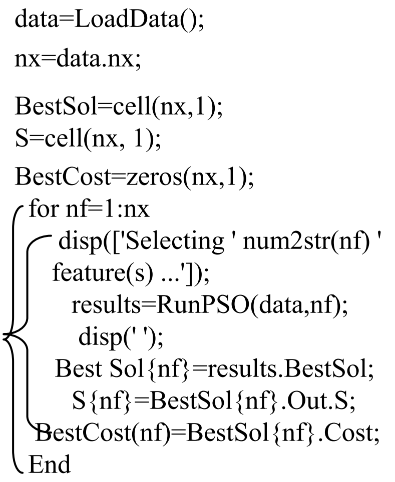
K Number of iteration count
V_{ij}^{k} (kth) iteration of attribute (i) of the velocity of particle (j)
X_{ij}^{k} (kth) iteration of attribute (i) of the position of particle (j)
W Weight of inertia
C^1, C^2 Acceleration co efficient
$(P_{\text{best}})_{ij}^{k}$ Until iteration (k), attribute (i) of the own best position of particle (j)
$(G_{\text{best}})_{i}^{k}$ Until iteration (k), attribute (i) of the best particle
$N_{\text{Attribute}}$ Attribute (i) of the best particle
N_{Particle} Number of particle present in the swarm
r^1, r^2 Uniformly generated distributed random number from the range 0 to 1

Based on the error rate, we selected certain features out of all features for our experimental approach shown in Fig. 2. The selection of the more features leads to a better result in terms of accuracy and error rate but it leads to decrease the time and memory in terms of efficiency. Permutation approach is used after selecting the best suitable features for analysis. The implementation phase uses a random key and a meta-heuristic approach for selecting features based on the permutation. Selecting the features having lower error rate leads to better result with less number of features. We selected twelve numbers of features for our observation to detect malicious accounts based on the pages collected.

Algorithm 1 The multi-objective PSO-based proposed algorithm for feature selection [13].

Fig. 2 Feature selection based on error rate





```
data=LoadData();
nx=data.nx;
BestSol=cell(nx,1);
S=cell(nx, 1);
BestCost=zeros(nx,1);
for nf=1:nx
    disp(['Selecting ' num2str(nf) '
   feature(s) ...']);
     results=RunPSO(data,nf);
      disp(' ');
   Best Sol{nf}=results.BestSol;
     S{nf}=BestSol{nf}.Out.S;
  BestCost(nf)=BestSol{nf}.Cost;
End
```

4 Experimental Analysis

The proposed framework detects spammer accounts in Facebook by using our dataset shown in Table 1. We selected randomly 200 posts from each categories contains malicious and benign accounts as a total of 1600 profile posts. Out of all the posts, more than 700 contents are spammer contents as per our analysis by using different features selected based on the PSO. We did the experiment using Python for selecting suitable features from all selected features and pass the contents through Weka tool [14], a machine learning platform for detecting spammer content by using some classification techniques. For all types of classification, we use tenfold cross-validation technique to get more suitable result. The result in terms of accuracy is the number of instances correctly predicted from all instances selected for experiment. The selected values for predicted result based the error rate and the different parameter that supported for successful execution.

4.1 Performance of Experiments

The performance reports attained by our framework for detecting spammer contents on our test set are based on the averages on tenfold cross-validation. It shows the content generated in the form of advertising and job alerts having higher percentage of spammer content as comparison to other form of posts. The detection rate gives the higher accuracy as 99.5% by using JRip classifier. The error detection rate and the time to build the model for our dataset also described in Tables 3 and 4. Our feature

Table 3 Different measures of our analysis based on dataset

Measures		TP rate	FP rate	Precision	Recall	F-measure	MCC	ROC area	PRC area	Correctly classified instances (Accuracy)
Classifiers	Random forest	0.9936	0.0032	0.9972	0.9935	0.9956	0.989	1.0000	1.0000	99.4771
	Random tree	0.9961	0.0074	0.9943	0.9966	0.9955	0.989	0.9955	0.9924	99.4771
	Bagging	0.9900	0.0066	0.9955	0.9907	0.9936	0.984	0.9984	0.9973	99.1955
	JRip	0.9951	0.0041	0.9961	0.9954	0.9967	0.990	0.9951	0.9952	99.5173
	J48	0.9933	0.0062	0.9955	0.9935	0.994	0.986	0.9955	0.9945	99.3162
	AdaBoost	0.9863	0.0269	0.9786	0.9865	0.982	0.960	0.9982	0.9961	98.029

Table 4 Error rate and model timing of different classification

Statistics	Classifier	Random forest	Radom tree	Bagging	JRip	J48	AdaBoost
Kappa statistics		0.9895	0.9895	0.9838	0.9903	0.9862	0.9606
Mean absolute error		0.0095	0.0052	0.0148	0.0072	0.0089	0.0248
Root mean square error		0.0652	0.0723	0.0791	0.0699	0.0082	0.1269
Relative absolute error		1.9055	1.0519	2.9704	1.4512	1.7842	4.9952
Root relative square error		13.08	14.50	15.84	14.02	16.515	25.4524
Model building time in second		1.08	0.8	0.37	0.33	0.19	0.28

selection model based on the PSO selected only 9 features for the experimental analysis based on the error rate and popularity of content spread by the user. We observed that, malicious contents are spread in the network by advertising content mostly with 20.33%. The percentages of malicious content based on the different categories are shown in Fig. 3.

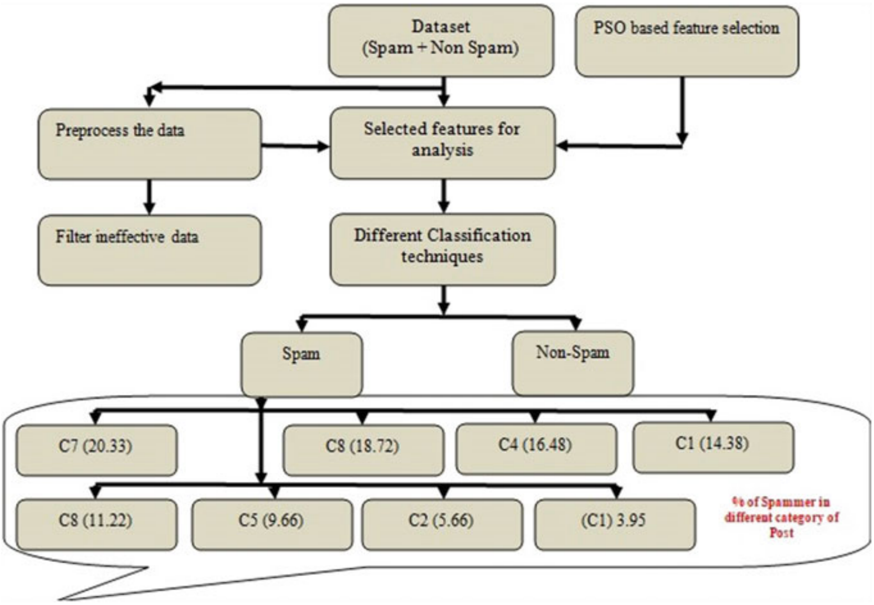


Fig. 3 Machine learning-based detection system

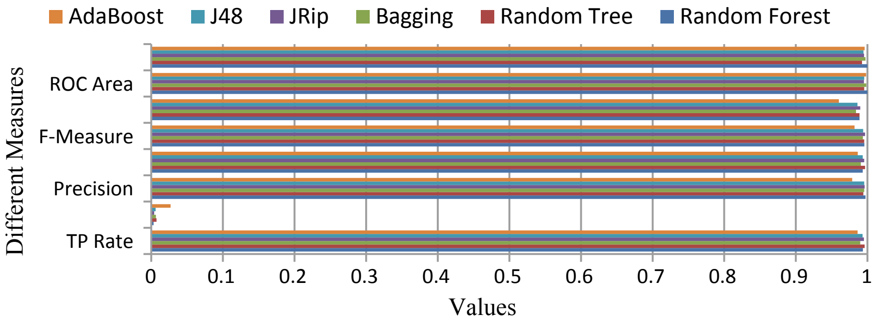


Fig. 4 Different measures based on classification

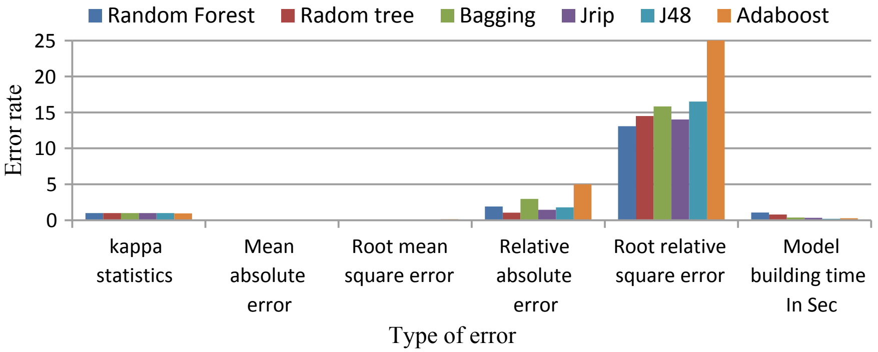


Fig. 5 Error rate based on classifications

As per experimental analysis, our proposed model gives better result with more accuracy for classifying the content as spam and non-spam compared to other methods. The graphical analysis of various classifications' result is depicted in Figs. 4 and 5.

4.2 Comparative Analysis

To detect spammers in various OSNs, our proposed model gathers and analyzes different posts associated with Facebook content. Our detection model achieves higher accuracy as compared to other states-of-the-art techniques described in Table 5.

Table 5 Comparative analysis with existing approaches

S. No	Research article	Dataset used	Accuracy
1.	Automatic detection of cybersecurity-related accounts on online social networks [6]	424 users information	97.17
2.	A feature selection approach to detect spam in the Facebook social network [15]	2 Lakh wall post	91.20
3.	Our proposed model	Different categories of wall posts (More than 5 Lakh)	99.6

5 Conclusion and Future Work

Nowadays, spammer's attacks on OSNs are the serious concerns for the users to protect their content. The specific features of social networking sites and the behavior of cybercriminals makes attract to classify malicious users. This paper introduced a spammer detection mechanism in Facebook by using machine learning-based model. Our research based on feature selection by using PSO algorithm from set of features and some supervised learning methods. The selected features process through different classifiers that give better accuracy in terms of ROC and TP rate. Our detection system also classifies the percentage of malicious content in each category of post. The best advantage of our overall research comprises in the step-down of various features and selected some specific features based on the error rate for detecting spammer activity. Our experimental approach for various model designs takes less time as compared to other researches. The result shows by using tenfold cross-validation, the popularity class prediction is more than 99% and the detection rate also. To achieve the detection rate more precision, PSO algorithm-based feature selection attain a very good selection rate.

Acknowledgements This research work is funded by Ministry of Electronics & Information Technology, Government of India under YFRF scheme and under the project Visvesvaraya Ph.D. Scheme which is implemented by Digital India Corporation.

References

1. Adewole, K.S., Anuar, N.B., Kamsin, A., Varathan, K.D., Razak, S.A.: Malicious accounts: dark of the social networks. J. Netw. Comput. Appl. **79**, 41–67 (2017)
2. Gupta, B.B. (ed.): Computer and Cyber Security: Principles, Algorithm, Applications, and Perspectives, pp. 666. CRC Press, Taylor & Francis, (2018)
3. Zhang, Z., Gupta, B.B.: Social media security and trustworthiness: overview and new direction. Future Gener. Comput. Syst. **86**, 914–925 (2018)
4. Gupta, B., Agrawal, D.P., Yamaguchi, S. (eds.): Handbook of Research on Modern Cryptographic Solutions for Computer and Cyber Security, p. 589. IGI Global, USA (2016)

5. Sami, M., Memon, S., Baloch, J., Bhatti, S.: An automated framework for finding fake accounts on Facebook. Int. J. Adv. Stud. Comput. Sci. Eng. **7**(2), 8–16 (2018)
6. Sohrabi, M.K., Karimi, F.: A feature selection approach to detect spam in the Facebook social network. Arab. J. Sci. Eng. **43**(2), 949–958 (2018)
7. Campos, G.F., Tavares, G.M., Igawa, R.A., Guido, R.C.: Detection of human, legitimate bot, and malicious bot in online social networks based on wavelets. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) **14**(1), 26 (2018)
8. Gurumurthy, S., Sushama, C., Ramu, M., Nikhitha, K.S.: Design and implementation of intelligent system to detect malicious Facebook posts using support vector machine (SVM). In: Soft Computing and Medical Bioinformatics (pp. 17–24). Springer, Singapore (2019)
9. Zhou, Y., Wang, X., Zhang, J., Zhang, P., Liu, L., Jin, H., Jin, H.: Analyzing and detecting money-laundering accounts in online social networks. IEEE Netw. **32**(3), 115–121 (2018)
10. Talukder, S., Carbunar, B.: AbuSniff: automatic detection and defenses against abusive Facebook friends (2018). arXiv preprint arXiv:1804.10159
11. Wang, X., Lai, C.M., Hong, Y., Hsieh, C. J., Wu, S.F.: Multiple accounts detection on Facebook using semi-supervised learning on graphs (2018). arXiv preprint arXiv:1801.09838
12. Dewan, P., Kumaraguru, P.: Facebook Inspector (FbI): Towards automatic real-time detection of malicious content on Facebook. Soc. Netw. Analy. Min. **7**(1), 15 (2017)
13. Sohrabi, M.K., Karimi, F.: A clustering based feature selection approach to detect spam in social networks. Int. J. Inf. Commun. Technol. Res. **7**(4), 27–33 (2015)
14. WEKA tool, http://www.cs.waikato.ac.nz/ml/weka
15. Aslan, Ç.B., Sağlam, R.B., Li, S. (2018). Automatic detection of cyber security related accounts on online social networks: Twitter as an example