Detecting Malicious Users in Twitter using Classifiers

Monika Singh
Department of Computer Science
PEC University of Technology
Chandigarh, India
monikaverma007@gmail.com

Divya Bansal
Department of Computer Science
PEC University of Technology
Chandigarh, India
divya@pec.ac.in

Sanjeev Sofat
Department of Computer Science
PEC University of Technology
Chandigarh, India
sanjeevsofat@pec.ac.in

ABSTRACT

The web has become a vital global platform that binds together almost all daily activities like communication, sharing, and collaboration. Impersonators, phishers, scammers and spammers crop up all the time in Online Social Networks (OSNs), and are even harder to identify. People in the public eyes like politicians, celebrities, sports persons, media persons and other public figures with huge followings are particularly vulnerable to this type of attacks. The main objective in this work is to identify those forged users who harm genuine ones, jeopardize the identity and hence the security and privacy of users. In this paper a framework for the detection of malicious users, non-malicious users and celebrities has been developed by using an attribute set for user classification based on user characteristics. For the purpose of detecting malicious users, non-malicious users and celebrities, a crawler has been developed for Twitter and data of around 22K users have been collected from publicly available information. Data of around 7,500 users have been used for training and testing purpose in Weka for classification of users. 5 classifiers have been used and compared on the basis of performance metrics like precision, recall, F-measure and accuracy. RandomForest outperforms all the classifiers with 99.8% accuracy.

Categories and Subject Descriptors

K.6.5 **[Management of Computing and Information Systems]:** Security and Protection - *Unauthorized access (e.g., hacking, phreaking)*

General Terms

Classifiers, Precision, Recall, F-measure, Accuracy, Followers, Followings, Tweets.

Keywords

Online Social Networks (OSNs), Twitter, Malicious users, Non-malicious users, Celebrities.

SIN '14, September 09 - 11 2014, Glasgow, Scotland Uk
Copyright 2014 ACM 978-1-4503-3033-6/14/09..$15.00.
http://dx.doi.org/10.1145/2659651.2659736

1. INTRODUCTION

Online social networks [1] allow users to create an online profile with updated personal and professional information. These networks started in 1997 with the launch of sixdegrees.com [1]. Social networks got popularity with the launch of Facebook in the year 2004 [1]. As per the statistics [2] from DMR (Digital Media Ramblings) till April 2014, Facebook has highest number of registered users (1.28 billion) and Twitter is at second position with 1 billion registered users followed by Google+ (343 million), LinkedIn (238 million) and Orkut (33 million). Increasing popularity of social networks has augmented cyber attacks. Various types of cyber attacks like identity theft attacks [3], social phishing attacks [4], spam attacks [5] and malware attacks [6]
Many incidents of such cyber attacks have been reported which compromised confidential and personal information of genuine users. In April 2013 [7], Twitter account belonging to the Associated Press was hacked and used to tweet that there were 2 explosions at the White House and President Barack Obama was injured. With the spread of this news there was a recordable downfall in the share market. And in Feb. 2013 [8] hackers had been able to gain access to around 2,50,000 accounts on Twitter including usernames, email addresses and passwords and in order to make compensation for that Twitter had to reset passwords of all 2,50,000 users.

As per the latest statistics from DMR (Digital Media Ramblings, a site that provides different statistics of social networking sites) [9], number of malicious users in Facebook and Twitter, two most popular social networks have been compared as shown in Table 1.

Table 1. Statistics of Facebook vs Twitter till April 2014

Statistics (in millions)	Facebook	Twitter
Total number of registered users	1280	1000
Fake users	140	20
Compromised accounts	0.6	-

Increasing reports of cyber attacks in OSNs is attracting security researchers to detect and mitigate threats to the users of OSNs. In this paper, we address the issue of detecting malicious and non-malicious users in Twitter. Twitter has been selected as the target

social networking site because it is the second most popular social networking site and the number of malicious users on it is increasing at a faster pace as depicted in table 1. For the detection purpose, we have crawled a large user data set of around 22k users from Twitter site. Then we created a labeled collection with users manually classified as malicious, non-malicious and celebrities. After that, we studied about the collected user behavior attributes to understand their relative discriminative power in distinguishing between malicious users and the two different types of non-malicious users as genuine ones and celebrities. These attributes have been used for calculating two parameters i.e. user score and tweet score which further classify users into three target categories. Then training and test dataset of around 7000 users has been used to check the feasibility of classification algorithms. Around 5 classification algorithms have been compared using Weka toolbox on the basis of evaluation metrics like precision, recall, accuracy and F-measure. It has been found that RandomForest classification approach is able to correctly identify the majority of the malicious users, misclassifying only a small percentage of legitimate users.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 describes crawling strategy and the test collection built from the crawled dataset. Section 4 describes and evaluates our strategy to detect three categories of users. Finally, section 5 offers conclusions and directions for future work.

2. RELATED WORK

This section is focusing on the existing work that has been done by various researchers for the detection of malicious users. Social network operators exploit authentication process [10] in order to protect their users and to make sure that the registered user is a factual live person. Social networking sites like Facebook offer user privacy settings that allow them to secure their personal data from other users in the network [11], [12]. Additional protection may include protection against hackers, spammers, social bots, identity cloning, phishing, and many other threats. Many commercial and open source products such as Checkpoint's SocialGuard [13], Websense's Defensio [14], UnitedParents [15], and RecalimPrivacy [16] offer tools for protection. In recent years, several published academic studies have proposed solutions for different social networking threats.

The domain of this literature review is the techniques that predict whether a given user is malicious or non-malicious. For the detection of spam users in various OSNs a summarized chart of papers [17-30] gone through has been given in table 2. Malicious profiles can be fake. And detection of fake profiles has been done in [31-37]. Fake profiles are the duplicate profiles created by malicious users by using the attributes of legitimate users. These users intend to send malicious messages to the friends of genuine user. Another category of malicious profiles could be compromised profiles. A compromised account [38] is an existing, legitimate account that has been taken over by an attacker by a phishing attack to obtain user's credentials.

After going through a rich set of studies at our disposal it has been found that researchers have done a significant work for the detection of spam, fake or compromised accounts in various OSNs like Twitter, Facebook, MySpace, LinkedIn etc. User based [33,20,34,37] and content based features [19,38,39] have been used. Any other feature like graphical distance, graph connectivity, Markov clustering method, URL rate, interaction rate, social relations, social activities, graph based features, neighbor based features, automation based features [28,30,22,40,41,23] have also been used.

Based on the literature, it has been found that there is a trade-off between achieved accuracy and dataset used. Few techniques have been able to achieve more than 90% accuracy but those techniques have been validated on small dataset of users with fixed number of malicious and non-malicious users. Technique which could identify all types of malicious users has not been proposed and results produced by the different techniques are identifying whether the user is malicious or not but no technique has been able to identify the extent of any user being malicious. Through this paper we have made efforts to identify three categories of users malicious, non-malicious and celebrities as explained in next sections.

3. USER DATA COLLECTION

For the purpose of evaluating proposed approach to detect malicious, non-malicious and celebrities in Twitter social network, we need a test collection of users. This collection needs to be pre-classified into three target categories. Since such collection is not publicly available for any of the social networking sites thus following steps have been taken in order to prepare user data collection.

Some terms need to be defined before presenting the steps taken to build test dataset. As per Twitter policy [42] malicious user is the one who follows a large number of users in a short period of time or if his post consists mainly of links or if popular hashtags (#) are used when posting unrelated information or repeatedly posting other user's tweets as your own. A non-malicious user is the one who is a genuine user and has almost equal number of followers and followings and who tweets moderately. There are users who have huge following count and tweet count, such users are famous personalities like celebrities / media persons or big organizations so we use the term celebrities for such type of users. Next, a subset of these users have been carefully selected and classified which is explained in section 3.2.

3.1 Crawling Twitter

In order to obtain a sample of malicious and non-malicious users, a crawler has been built using Twitter 4j which is an open source Java library for Twitter API [43]. Publicly available dataset has been gathered through Twitter REST API that works by making a request for a specific type of data. Thus details of users such as IDs, screen name, location, friend's details, follower's details etc. has been obtained encoded in JSON(Java Script Object Notation). There is a rate limit for calls to API which is limited to 350 requests per hour per host [44]. In order to avoid congestion Twitter has been crawled continuously for 5 weeks with rate limit of 300 requests per hour, gathering a total of 21,492 users with their 20 most recent tweets.

3.2 Building Dataset

In order to build test dataset, collected sample has been parsed to obtain four desired graph based features like number of followers, number of followings, number of tweets and account creation date. Using these four features an analytical model [45] with equations for two parameters user score and tweet score have been used for categorization of users as mentioned below:

$$\text{User score} = (10/\Phi)*(\log(\beta) + \log(\delta)) \quad (1)$$

$$\text{Tweet score} = (10/\Omega)*(\log(\alpha) + \log(\Upsilon) + \Psi) \quad (2)$$

where $\Omega = 6.1$, $\Psi = 3.4$, $\Phi = 8.4$

and α = number of tweets / number of months on Twitter

β = number of followers / number of followings

δ = number of followers + number of followings

Υ = tweet frequency per day

These two scores in the range of 0-10 have been calculated for collected sample of 21,492 users and stored in excel sheets.

In order to classify users as malicious, non-malicious and celebrities or big organizations four groups are defined. Group one consists of users with user score <1 and tweet score <1. This group includes users who follow large number of people than their followers and tweet less, thus forms malicious group. Group two consists of users with user score <1 and tweet score >1. This group includes people who follow large number of people and they also tweet more in order to gain attention of others. Such group also falls under malicious category. Group three consists of users with 1< user score <5 and 1<tweet score<5. Such people are non-malicious as they have almost equal number of followers and followings and tweet moderately. Last group four consists of people with user score>=5 and tweet score>=5. Such group follows less number of people than their followers and tweet more, so taken as celebrities or big organizations. After categorization of four user groups, a total of 7434 users' database has been obtained that is used for training and testing purpose.

4. CLASSIFICATION OF USERS

In this section, feasibility of applying different classification algorithms for the detection of malicious and non-malicious users has been investigated. Learning algorithms learn classification from previously classified data and then acquired knowledge is applied to classify unseen users into desired categories of malicious, non-malicious users and celebrities.

In section 4.1, metrics used to evaluate experimental results have been presented. Section 4.2 is about various classification algorithms used in Weka toolbox. In the classification process users from the test collection are directly classified into malicious, non-malicious and celebrities. Results from classification are presented in section 4.3.

4.1 Evaluation Metrics

The effectiveness of classification techniques has been assessed using the metrics like confusion matrix, recall, precision, F-measure and accuracy. Recall [46, 47] is defined as the ratio of correctly classified users to the number of users in a class. Precision [46, 47] is defined as the ratio of the number of user classified correctly to the total users predicted in a class. Accuracy [46, 47] is defined as the overall correctness. Confusion matrix [46, 47] has been used to explain these metrics. Precision P, recall R and accuracy A of the class malicious is computed as:

$P = a/(a+d+g)$,

$R = a/(a+b+c)$,

$A = a/(a+b+c+d+e+f+g+h+i)$,

F-measure is the harmonic mean between precision and recall and is given as

$F = 2PR/(P+R)$.

4.2 Experimental Setup

Weka toolbox has been used for classification purpose. Around 5 classification algorithms have been used and compared on the basis of evaluation metrics [46, 47] as discussed in section 4.1. From the collected sample of 21,492 users 7434 users have been obtained on the basis of desired parameters of user score and tweet score as mentioned in section 3.2. So out of 7434 users, 5203 (around 70%) have been used for training and remaining 2231 (30%) for testing. Results of running various classification algorithms have been given in section 4.3.

4.3 Results and Analysis

Table 3 shows the results of all 5 classifiers used with their precision, recall, F-measure and accuracy as discussed in section 4.1 also.

Table 3. Evaluation Metrics of 5 classification algorithms

SN	Classifier	Precision	Recall	F-measure	Accuracy
1	BayesNet	99.2	99.2	99.2	96.1
2	Naïve Bayes	95.5	88.7	90.7	95.8
3	SMO	85.6	86.8	85.1	81.6
4	J48	99.7	99.7	99.7	96.0
5	Random Forest	99.8	99.8	99.8	99.8

Table 3 clearly shows that RandomForest is giving highest accuracy of 99.8% which is not surprising as this classifier can work with imbalanced dataset and in our case training dataset contains unequal number of malicious, non-malicious and celebrities. And SMO classifier has given minimum accuracy of 81.6%. Plots of precision, recall, F-measure and accuracy have been shown in figure 1.

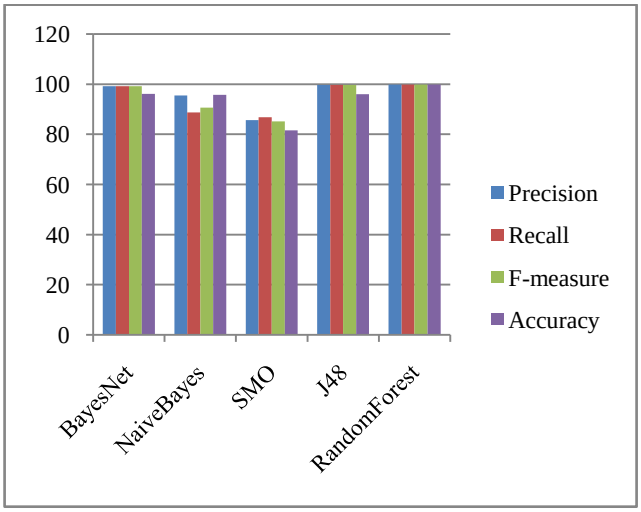

Figure1. Comparison chart of evaluation metrics of 5 classifiers

5. CONCLUSION AND FUTURE WORK

It can be clearly said that Social Networks have become a target for spammers. The information revealed on an OSN can be exploited by an attacker to embarrass, to blackmail, to impersonate or even to damage the image of profile holder. Cyber attacks cause a serious threat to the security and privacy of social networking users. In this paper the problem of identifying malicious and non-malicious users on one of the most popular social networking site, Twitter has been approached. Crawler has been designed for Twitter site to obtain around 22K user profiles, all their tweets and links of followers and followings. Few user based attributes like followers count, followings count, tweet count, date of creation of account useful to differentiate malicious, non-malicious and celebrities have been analyzed. These features are influenced by Twitter spam policy [42]. Our analysis study is leveraged towards a spammer detection mechanism. Then with an analytical model using all our four selected attributes, two parameters user score and tweet score have been calculated. On the basis of these two parameters a labeled collection of users classified as malicious, non-malicious and celebrities have been prepared. Celebrities are another category of non-malicious users. Thus a total of around 7500 users have been obtained out of which 70% have been used for training and rest 30% have been used as testing dataset. Using classification techniques available in Weka, 5 classification algorithms have been used and compared. Results show that RandomForest classifier giving highest accuracy of 99.8%.

The directions towards which this work can evolve are: we intend to increase and improve our labeled collection on the basis of more effective attributes. Periodical evaluation of the classification methods may be necessary in the future so that retraining mechanisms could be applied. This model is detecting malicious and two types of non-malicious users so there is a requirement to further identify different categories of malicious users like fake or compromised as well.

6. REFERENCES

[1] Boyd D.M. and Ellison N. B. 2007. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*. 13, 1 (17 Dec 2007), 210-230. DOI= 10.1111/j.1083-6101.2007.00393.x.

[2]Statistics of Social Networking Sites: http://expandedramblings.com/index.php/resource-how-many-people-use-the-top-social-media. Accessed: 2014-05-03.

[3] Jin L., Takabi H. and Joshi James B.D. 2011. Towards Active Detection of Identity Clone Attacks in Online Social Networks. In *Proceedings of the first ACM conference on Data and application security and privacy(CODASPY'11)*, ACM, New York, USA, 27-38.

[4] Jagatic T., Johnson N., Jakobsson M. and Menczer F. 2007. Social phishing, *Communications of the ACM Journal*. 50, 10 (Oct. 2007), ACM, New York, USA, 94–100. DOI= 10.1145/1290958.1290968.

[5]Spam Attacks Information:

http://www.whoswatchingcharlottesville.org/social.html. Accessed: 2014-04-23.

[6] Kuzma J. 2011. Account creation security of social network sites. *International Journal of Applied Science and Technology*, 1, 3 (June 2011), 8–13.

[7] Hacking of Twitter accounts: http://BBC News - AP Twitter account hacked in fake 'White House blasts' post.html. Accessed: 2014-04-23.

[8] Hacking of Twitter accounts: http://Twitter resets a quarter of a million accounts after hacker attack.html. Accessed: 2014-04-23.

[9] Statistics of Social Networking Sites: http://expandedramblings.com/index.php/march-2013-by-the-numbers-a-few-amazing-twitter-stats/3/#.U3xVv9KSyuE. Accessed: 2014-04-23.

[10] Liu Y., Gummadi K., Krishnamurthy B. and Mislove A. 2011. Analyzing facebook privacy settings: User expectations vs. reality. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, New York, NY, USA, 61–70. DOI= 10.1145/2068816.2068823.

[11] Mahmood S. and Desmedt Y. 2011. Poster: preliminary analysis of Google+'s privacy. In *Proceedings of the 18th ACM conference on Computer and communications security*. ACM, New York, NY, USA, 809-812. DOI= 10.1145/2046707.2093499.

[12] Facebook Policies: http:// developers.facebook.com/policy. Accessed: 2014-03-12.

[13] ZoneAlarm: http://www.zonealarm.com. Accessed:

2014-03-12.

[14] W. Defensio: http://www.defensio.com. Accessed:

2014-03-12.

[15] UnitedParents: http://www.unitedparents.com. Accessed:

2014-03-12.

[16] G.ReclaimPrivacy: http://www.reclaimprivacy.org. Accessed: 2014-03-12.

[17] Balasubramaniyan A. V.,Maheswaran A., Mahalingam V., Ahamad M. and Venkateswaran H. 2010. A Crow or a Blackbird? Using True Social Network and Tweeting Behavior to Detect Malicious Entities in Twitter, ACM.

[18] Lee K., Caverlee J. and Webb S. 2010. Uncovering Social Spammers: Social Honeypots + Machine Learning. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. New York, ACM, New York, NY, USA, 435-442. DOI=10.1145/1835449.1835522.

[19] Benevenuto F., Magno G., Rodrigues T. and Almeida V. 2010. Detecting Spammers on Twitter, In *Proceedings of Seventh annual Collaboration, Electronic messaging, Anti Abuse and Spam Conference (CEAS 2010)*. Washington, US.

[20] Gee G. and Teh H. 2010. Twitter Spammer Profile Detection, available online on: *cs229.stanford.edu/proj2010/GeeTeh-Twitter Spammer Profile Detection.pdf*.

[21] Wang A. 2010. Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach. *Data and Applications Security and Privacy XXIV, LNCS, Springer*

Berlin Heidelberg. 6166, 335-342.DOI= 10.1007/978-3-642-13739-6_25.

[22] Song J, Lee S. and Kim J. 2011. Spam Filtering in Twitter using Sender-Receiver Relationship. In *Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection (RAID'11),Springer-Verlag Berlin, Heidelberg.* 301-317. DOI= 10.1007/978-3-642-23644-0_16.

[23] Yang C., Harkreader R. and Guofei Gu 2011. Die Free or Live Hard? Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. In *Proceedings of the 14th International Conference on Recent Advances in Intrusion Detection (RAID'11), Springer-Verlag Berlin, Heidelberg*, 318-337. DOI= 10.1007/978-3-642-23644-0_17.

[24] McCord M. and Chuah M. 2011. Spam Detection on Twitter Using Traditional Classifiers, *Autonomic and Trusted Computing.* Lecture Notes in Computer Science, *Springer-Verlag Berlin, Heidelberg* , 6906, 175-186. DOI= 10.1007/978-3-642-23496-5_13.

[25] Ahmed F. and Abulaish M. 2012. An MCL-Based Approach for Spam Profile Detection in Online Social Networks, In *Proceedings of 11th International Conference on Trust, Security and Privacy in Computing and Communications* (Liverpool, United Kingdom United Kingdom, June 25-June 27), IEEE, 602-608. DOI= http://doi.ieeecomputersociety.org/10.1109/TrustCom.2012.83.

[26] Chakraborty A., Sundi J. and Satapathy S. 2012. SPAM: A Framework for Social Profile Abuse Monitoring, *available online at: www.cs.sunysb.edu/~aychakrabort/files/spam.pdf.*

[27] Amleshwaram A., Reddy N., Yadav S., Guofei Gu and Yang C. 2013. CATS: Characterizing Automation of Twitter Spammers. In P*roceedings of 5th International Conference Communication Systems and Networks (COMSNETS),* Bangalore, IEEE, 1-10. DOI= 10.1109/COMSNETS.2013.6465541.

[28] Lin P. and Huang P. 2013. A Study of Effective Features for Detecting Long-surviving Twitter Spam Accounts. In *Proceedings of 15th International Conference on Advanced Communication Technology (ICACT),* (PyeongChang, 27-30 Jan. 2013) IEEE, 841-846.

[29] Yang C., Harkreader R. and Guofei Gu 2013. Empirical Evaluation and New Design for Fighting Evolving Twitter Spammers. *IEEE Transactions on Information Forensics and Security*, 8, 8 (August 2013), 1280 – 1293. DOI= 10.1109/TIFS.2013.2267732 .

[30] Ahmed F. and Abulaish M. 2013. A generic statistical approach for spam detection in Online Social Networks. *Computer Communications Journal*, Elsevier, 36, 10-11(April 2013),1120-1129. DOI= 10.1016/j.comcom.2013.04.004.

[31] Kontaxis G., Polakis I., Ioannidis S. and Markatos E. 2011. Detecting Social Network Profile Cloning. In *Proceedings of International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops*)(21-25 March 2011, Seattle, WA) IEEE, 295-300.DOI= 10.1109/PERCOMW.2011.5766886.

[32] Yang Z., Wilson C., Wang X., Gao T., Zhao B. and Dai Y. 2011. Uncovering Social Network Sybils in the Wild. In *Proceedings of the ACM SIGCOMM conference on Internet measurement conference (IMC'11),* New York, USA, 259-268. DOI= 10.1145/2068816.2068841.

[33] Malhotra A., Totti L., Meira W., Kumaraguru P. and Almeida V. 2012. Studying User Footprints in Different Online Social Networks. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* (Istanbul Turkey, August 26-29, 2012), IEEE/ACM, 1065-1070. DOI= http://doi.ieeecomputersociety.org/10.1109/ASONAM.2012.184.

[34] Conti M., Poovendran R. and Secchiero M. 2012. FakeBook: Detecting Fake Profiles in On-line Social Networks. In *Proceedings of International Conference on Advances in Social Networks Analysis and Mining (ASONAM),* (Istanbul Turkey, August 26-29, 2012), IEEE/ACM, 1071-1078.DOI= http://doi.ieeecomputersociety.org/10.1109/ASONAM.2012.185.

[35] Fire M., Katz G. and Elovici Y. 2012. Strangers Intrusion Detection - Detecting Spammers and Fake Profiles in Social Networks Based on Topology Anomalies. *ASE Human Journal*. 26-39.

[36] Cao Q., Sirivianos M.,Yang X. and Pregueiro T. 2012. Aiding the Detection of Fake Accounts in Large Scale Social Online Services. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NSDI'12),* USENIX Association Berkeley, CA, USA.

[37] Flores M. and Kuzmanovic A. 2013. Searching for Spam: Detecting Fraudulent Accounts via Web Search. *Lecture Notes in Computer Science (LNCS),* Springer-Verlag Berlin Heidelberg, 7799, 208-217. DOI= 10.1007/978-3-642-36516-4_21.

[38] Egele M., Stringhini G., Kruegel C., Vigna G. 2013. COMPA:Detecting Compromised Accounts on Social Networks. In *Proceedings of Network and Distributed System Security Symposium (NDSS)*, CA, United States (23 Apr 2013).

[39] Stringhini G., Kruegel C. and Vigna G. 2010. Detecting Spammers on Social Networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, Austin, Texas USA, University of California, Santa Barbara, ACM, 1-9. DOI= 10.1145/1920261.1920263.

[40] Zhuy Y.,Wang X., Zhong E., Liuy N.N., Li H. and Yang Q. Discovering Spammers in Social Networks. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. DOI= 10.1.1.298.8002.

[41] Bilge L., Strufe T., Balzarotti D., and Kirda E. 2009. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In *Proceedings of International World Wide Web Conference Committee (IW3C2),* WWW 2009, (April 20–24, 2009, Madrid, Spain), ACM, 551-560. DOI= 10.1145/1526709.1526784.

[42] Twitter Policy of Spammers: http://help.twitter.com/forums/26257/entries/1831- The Twitter Rules. Accessed: 2014-03-20.

[43] Twitter API: https://dev.twitter.com/docs/rate-limiting/1.1. Accessed: 2014-03-23.

[44] Twitter rate limit for calls: https://dev.twitter.com/docs/api/streaming. Accessed: 2014-03-23.

[45] Sharma A., Ahluwalia A.,Deep S. and Bansal D. 2014. Friend or Foe: Twitter Users under Magnification, *Distributed Computing and Internet Technology*, Lecture Notes in Computer Science, 8337, 251-262. DOI= 10.1007/978-3-319-04483-5_26.

[46] Lu Z. 2004. Predicting Subcellular Localization of Proteins using Machine-Learned Classifiers. *Bioinformatics,* 20, 4, 547 – 556. DOI= 10.1.1.131.4463.

[47] Eisner R. 2005. Improving Protein Function Prediction using the Hierarchical Structure of the Gene Ontology. In *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology.* DOI= 10.1.1.131.7437.

Table 2. Summary of Work on Detection of Spam Profiles

SN	Year	Author	Metrics Used	Methodology	Dataset	Remarks
1	2010	Balasubramaniyan et al. [17]	Retweet, @mentions	PRTSN (Page Rank True Social Network)	Crawled 2,00,000 Twitter Users' data	In 31 days 181 accounts suspended by Twitter
2	2010	Lee et al. [18]	User based	Compared Decorate, SimpleLogistic, FT, LogiBoost, J48, RandomSubSpace, Bagging, LibSVM	Validated on 1000 Twitter users	-Decorate classifier giving highest accuracy-88.98% -Validated on 2 combinations of users -Validated on small dataset
3	2010	Benevenuto et al.[19]	User based & Content based	SVM	Validated on 1065 Twitter users	-Accuracy-87.6% (with user based and content based features) -Accuracy-84.5% (with only user based features)
4	2010	Gee et al. [20]	User based	Compared Naive Bayesian, SVM	Validated on 450 Twitter users with 200 recent tweets	-Accuracy-89.6% -Technical features not used -Deployment is possible if accuracy is 99%
5	2010	Wang [21]	User based and Content based	Compared Naive Bayesian, Neural Network, SVM & Decision Tree	Validated on 500 Twitter users with 20 recent tweets	-Naive Bayesian giving highest accuracy -93.5% -Validated on small dataset
6	2011	Song et al. [22]	Distance & connectivity in graph between friends and followers	If distance >4 then spam, and if no connection between friends and followers then	1,48,371 genuine and 308 spam data collected	- New users sending message to any user will be flagged as spam

				spam		
7	2011	Yang et al. [23]	18 features (8-existing & 10 new features introduced)	Compared Random Forest, Decision Tree, Decorate, Naive Bayesian	Validated on two datasets-5000 users and then 3500 users with 40 recent tweets	-Bayesian giving highest accuracy-88.6% -Crawled and validated on small dataset
8	2011	McCord et al. [24]	User based and content based	Compared Random Forest, SVM, Naive Bayesian, K-NN	Validated on 1000 Twitter users with 100 recent tweets	-Radom Forest classifier giving highest accuracy-95.7% -Unbalanced dataset used -Validated on small dataset
9	2012	Ahmed et al. [25]	Graph properties have been used	MCL (Markov Clustering)	Facebook data of 305 users collected	Good results
10	2012	Chakraborty et al. [26]	User based, Content based	Compared Random Forest, SVM, Naive Bayesian, Decision Tree	Trained on 5000 Twitter users with 200 recent tweets	SVM giving highest accuracy-89%
11	2013	Amit A. et al. [27]	Introduced 15 new features	Compared Random Forest, Decision Tree, Decorate, Naive Bayesian	Validated on 31,808 Twitter users	- Accuracy-93.6%
12	2013	Lin et al.[28]	URL rate, interaction rate	J48	Validated on 400 Twitter users	-Precision-86% -Only 2 features used for detection -Validated on small dataset
13	2013	Yang et al. [29]	18 features-10 new and 8 existing	Compared their approach with 4 existing approaches by Benevento, Wang, Stringhini, K.Lee using RF, DT, Decorate and Bayes Net	20,000 normal and 3060 spam users in Twitter	-Accuracy – 89% -Limitations : identified spammers are less, dataset used may be biased, -new features used are expensive to extract and calculate
14	2013	Ahmed et al.[30]	Wall posts/tweets, links, friends/followers, mentions, hashtags	Naïve Bayes, Jrip, J48	Validated with Facebook-320 profiles, Twitter-305 profiles	-Accuracy with Facebook – 96.4% with Naïve Bayes, -Accuracy with Twitter-98.7% with Jrip and Combined dataset 95.7% with J48. -Very very small dataset used for validation