iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 Issue: VI Month of publication: June 2015

DOI:

www.ijraset.com Volume 3 Issue VI, June 2015
IC Value: 13.98 ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Review: Efficient Spam Detection on Social Network

Girisha Khurana[1], Mr Marish Kumar[2]
[1]Student, [2]Assistant Professor, Department Of Computer Science GNI Mullana Kurushetra University

***Abstract* — With the rapid growth of social networking sites for communicating, sharing, storing and managing significant information, it is attracting cybercriminals who misuse the Web to exploit vulnerabilities for their illicit benefits. The rapid growth of Twitter has triggered a dramatic increase in spam volume and sophistication. The abuse of certain Twitter components such as ''hashtags'', ''mentions'', and enables shortened URLs spammers to operate efficiently .In this paper we have reviewed the existing techniques for detecting spam users in Twitter social network. Features for the detection of spammers could be user based or content based or both and spam classifier methods.**

Keyword: Social network, Twitter, Weibo , Classifier, Malware

I. INTRODUCTION

Within the past few years, online social network, such as Face-book, Twitter, Weibo, etc., has become one of the major way for internet users to keep communications with their friends. According to Statista report [1], the number of social network users has reached 1.61 billion until late 2013, and is estimated to be around 2.33 billion users globe, until the end of 2017.

However, along with great technical and commercial success, social network platform also provides a large amount of opportunities for broadcasting spammers, which spreads malicious mes-sages and behavior. According to Nexgate's report [2], during the first half of 2013, the growth of social spam has been 355%, much faster than the growth rate of accounts and messages on most branded social networks.

The impact of social spam is already significant. A social spam message is potentially seen by all the followers and recipients' friends. Even worse, it might cause misdirection and misunderstand-ing in public and trending topic discussions. For example, trending topics are always abused by spammers to publish comments with URLs, misdirecting all kinds of users to completely unrelated web-sites. Because most social networks provide shorten service on URLs inside messages it is difficult to identify the content without visiting the site.

A. Types of Spammers

1) Spammers: are the malicious users who contaminate the information presented by legitimate users and in turn pose a risk to the security and privacy of social networks. Spammers belong to one of the following categories [15]:

2) Phishers: are the users who behave like a normal user to acquire personal data of other genuine users.

3) Fake Users: are the users who impersonate the profiles of genuine users to send spam content to the friends' of that user or other users in the network.

4) Promoters: are the ones who send malicious links of advertisements or other promotional links to others so as to obtain their personal information.

B. Motives Of Spammers

1) Disseminate pornography
2) Spread viruses
3) Phishing attacks
4) Compromise system reputation

C. The Twitter Social Network

Twitter is a social network service launched in March 21, 2006 [16] and has 500 million active users [16] till date who share information. Twitter uses a chirping bird as its logo and hence the name Twitter. Users can access it to exchange frequent information called 'tweets' which are messages of up to 140 characters long that anyone can send or read. These tweets are public by default and visible to all those who are following the twitter. Users share these tweets which may contain news, opinions, photos, videos, links, and messages. Following is the standard terminology used in Twitter and relevant to our work:

1) *Tweets [17]:* A message on Twitter containing maximum length of 140 characters.
2) *Followers & Followings [17]:* Followers are the users who are following a particular user and followings are users whom user follows.
3) *Retweet [17]:* A tweet that has been reshared with all followers of a user.
4) *Hashtag [17]:* The # symbol is used to tag keywords or topics in a tweet to make it easily identifiable for search purposes
5) *Mention [17]:* Tweets can include replies and mentions of other users by preceding their usernames with @ sign.
6) *Lists [17]:* Twitter provides a mechanism to list users you follow into groups
7) *Direct Message [17]:* Also called a DM, this represents Twitter's direct messaging system for private Communication amongst user.

As per Twitter policy [18], indicators of spam profiles are the metrics such as following a large number of users in a short period of time1or if 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. There is a provision for users to report spam profiles to Twitter by posting a tweet to @spam. But in Twitter policy [18] there is no clear indication of whether there are automated processes that look for these conditions or whether the administrators rely on user reporting, although it is believed that a combination approach is used.

D. Threats On Twitter

1) Spammed tweets [19]: Twitter allows its users to post tweets of maximum 140 characters but regardless of the character limit, cybercriminals have found a way to actually use this limitation to their advantage by creating short but compelling tweets with links for promotions for free vouchers or job advertisement posts or other promotions.

2) Malware downloads [19]: Twitter has been used by cyber criminals to spread posts with links to malware download pages. FAKEAV and backdoor[19] applications are the examples of Twitter worm that sent direct messages, and even malware that affected both Windows and Mac operating systems. The most tarnished social media malware is KOOBFACE [19], which targeted both Twitter and Facebook

3) Twitter bots [17]: Cybercriminals tend to use Twitter to manage and control botnets. These botnets control the users’ accounts and pose a threat to their security and privacy.

II. LITERATURE SURVEY

In the past ten years, email spam detection and filtering mechan-isms have been widely implemented. The main work could be summarized into two categories: the content-based model and the identity-based model. In the first model, a series of machine learning approaches [3,4] are implemented for content parsing according to the keywords and patterns that are spam potential. In the identity-based model, the most commonly used approach is that each user maintains a whitelist and a blacklist of email addresses that should and should not be blocked by anti-spam mechanism [5,6]. More recent work is to leverage social network into email spam identifica-tion according to the Bayesian probability [7]. The concept is to use social relationship between sender and receiver to decide closeness and trust value, and then increase or decrease Bayesian probability according to these value.

With the rapid development of social networks, social spam has attracted a lot of attention from both industry and academia. In industry, Facebook proposes an EdgeRank algorithm [8] that assigns each post with a score generated from a few feature (e.g., number of likes, number of comments, number of reposts, etc.). Therefore, the higher EdgeRank score, the less possibility to be a spammer. The disadvantage of this approach is that spammers could join their networks and continuously like and comment each other in order to achieve a high EdgeRank score.

In academia, Yardi et al. [9] studies the behavior of a small part of spammers in Twitter, and find that the behavior of spammers is different from legitimate users in the field of posting tweets, followers, following friends and so on. Stringhini et al. [10] further inv-estigates spammer feature via creating a number of honey-profiles in three large social network sites (Facebook, Twitter and Myspace) and identifies five common features (followee-to-follower, URL ratio, message similarity, message sent, friend number, etc.) potential for spammer detection. However, although both of two approaches introduce convincible framework for spammer detection, they lack of detailed approaches specification and prototype evaluation.

Wang [11] proposes a naïve Bayesian based spammer classifica-tion algorithm to distinguish suspicious behavior from normal ones in Twitter, with the precision result (F-measure value) of 89%. Gao et al. [12] adopts a set of novel feature for effectively reconstructing spam messages into campaigns rather than examining them indivi-dually (with precision value over 80%). The disadvantage of these two approaches is that they are not precise enough.

Benevenuto et al. [13] collects a large dataset from Twitter and identify 62 feature related to tweet content and user social beh-avior. These characteristics are regarded as attributes in a machine learning process for classifying users as either spammers or non-spammers. Zhu et al. [14] proposes a matrix factorization based spam classification model to collaboratively induce a succinct set of latent feature (over 1000 items) learned through social relation-ship for each user in RenRen site (www.renren.com). However, these two approaches are based on a large amount of selected feature that might consume heavy computing capability and spend much time in model training

III. EXISTING METHOD FOR SPAM DETECTION

Different techniques have been used by researchers to find out the spam profiles in various OSNs. We are focussing only on the work that has been done to identify spammers in Twitter as it is not only a social communication media but in fact is used to share and spread information related to trending topics in real time. Table 1 is showing the summary of the papers reviewed regarding the detection of spammers in Twitter**.**

Table 1. Outline of techniques used for the detection of spammers

Author	Metrics Used	Methodology Used	Dataset Used	Result
Alex Hai Wang[20]	Graph based and Content based	Compared Naive Bayesion , Neural Network ,SVM & Decision tree	Validated on 500 Twitter with 20 recent tweets	Naive Bayesion giving highest accuracy 93.5%
Lee et al.[15]	User Based	Compared Decorate, Simple Logistic, FT, Logi Boost ,RandomsubSpace,Bagging,j48,LibSVM	Validated on 1000 Twitter Users	Decorate giving highest accuracy 88.98%
Benevenoto et.al[21]	User based and Content Based	SVM	Validated on 1065 Twitter Users	Accuracy 87.6% with User Based & Content Based features and Accuracy 84.5(With only user based features
Gee et.al[22]	User Based	Compared Naive Bayesion ,SVM	Validated on 450 Twitter Users with 200 recent tweets	Accuracy 89.6%
McCord et.al[23]	User Based and Content Based	Compared Random Forest,SVM, Naive Bayesion ,KNN	Validated on 1000 Twitter Users with 100 recent tweets	Random forest giving highest accuracy 95.7%
Chakraborty et.al[24]	User Based and Content Based	Compared Random Forest ,SVM, Naive Bayesion ,Decision Tree	Trained on 5000 Twitter Users with 200 recent tweets	SVM giving highest accuracy-89%
X. Zheng et al[25]	User Based and Content Based	SVM	Validated on 30,000 weibo users	SVM giving highest accuracy-99%

Significant work has been done by Alex Hai Wang [20] in the year 2010 which used user based as well as content based features for detection of spam profiles. A spam detection prototype system has been proposed to identify suspicious users in Twitter. A directed social graph model has been proposed to explore the “follower” and “friend” relationships. Based on Twitter’s spam policy, content-based features and user-based features have been used to facilitate spam detection with Bayesian classification algorithm. Classic evaluation metrics have been used to compare the performance of various traditional classification methods like Decision Tree, Support Vector Machine (SVM), Naive Bayesian, and Neural Networks and amongst all Bayesian classifier has been judged the best in terms of performance. Over the crawled dataset of 2,000 users and test dataset of 500 users, system achieved an accuracy of 93.5% and 89% precision. Limitation of this approach is that is has been tested on very less dataset of 500 users by considering their 20 recent tweets.

Lee et. al.[15] deployed social honeypots consisting of genuine profiles that detected suspicious users and its bot collected evidence of the spam by crawling the profile of the user sending the unwanted friend requests and hyperlinks in MySpace and Twitter. Features of profiles like their posting behaviour, content and friend information to develop a machine learning classifier have been used for identifying spammers. After analysis profiles of users who sent unsolicited friend requests to these social honeypots in MySpace and Twitter have been collected. LIBSVM classifier has been usedfor identification of spammers. One good point in the approach is that it has been validated on two different combinations of dataset – once with 10% spammers+90% non-spammers and again with 10% non-spammers+90% spammers. Limitation of the approach is that less dataset has been used for validation.

Benevenuto et. al. [21] detected spammers on the basis of tweet content and user based features. Tweet content attributes used are - number of hashtags per number of words in each tweet, number of URLs per word, number of words of each tweet, number of characters of each tweet, number of URLs in each tweet, number of hashtags in each tweet, number of numeric characters that appear in the text, number of users mentioned in each tweet, number of times the tweet has been retweeted. Fraction of tweets containing URLs, fraction of tweets that contains spam words, and average number of words that are hashtags on the tweets are the characteristics that differentiate spammers from non spammers. Dataset of 54 million users on Twitter has been crawled with 1065 users manually labelled as spammers and non-spammers. A supervised machine learning scheme i.e. SVM classifier has been used to distinguish between spammers and non spammers. Detection accuracy of the system is 87.6% with only 3.6% non-spammers misclassified.

Twitter facilitates its users to report spam users to them by sending a message to “@spam”. So Gee et. al. [22] utilized this feature and detected spam profiles using classification technique. Normal user profiles have been collected using Twitter API and spam profiles have been collected from“@spam” in Twitter. Collected data was represented in JSON then it was presented in matrix form using CSV format. Matrix has users as rows and features as columns. Then CSV files were trained using Naive Bayes algorithm with 27% error rate then SVM algorithm has been used with error rate of 10%. Spam profiles detection accuracy is 89.3%. Limitation of this approach is that not very technical features have been used for detection and precision is also less i.e. 89.3% so it has been suggested that aggressive deployment of any system should be done only if precision is more than 99%.

McCord et.al. [23] used user based features like number of friends, number of followers and content based features like number of URLs, replies/mentions, retweets, hashtags of collected database. Classifiers namely Random Forest, Support Vector Machine (SVM), Naive Bayesian and K-Nearest Neighbour have been used to identify spam profiles in Twitter. Method has been validated on 1000 users with 95.7% precision and 95.7% accuracy using the Random Forest classifier and this classifier gives the best results followed by the SMO, Naive Bayesian and K-NN classifiers. Limitation of this approach is that for considered dataset reputation feature has been showing wrong results i.e. it is not able to differentiate spammers and non-spammers, unbalanced dataset has been used so Random Forest is giving best results as this classifier is generally used in case of unbalanced dataset, and finally the approach has been validated on less dataset.

Chakraborty et. al. [24] have proposed a system to detect abusive users who post abusive contents, including harmful URLs, porn URLs, and phishing links and divert away regular users and harm the privacy of social networks. Two steps in the algorithm have been used- first is to check the profile of a user sending friend request to other user as for abusive content and second is to check the similarity of two profiles. After these two steps it is supposed to recommend whether the user should accept friend request or not. This has been tested on Twitter dataset of 5000 users which was collected with REST API. Features considered for differentiating abusive and non-abusive users are- profile based, content based and timing based. Classifiers like SVM, Decision Tree, Random Forest and Naïve Bayesian have been used. SVM outperforms all classifiers and model is performing with an accuracy of 89%.

X. Zheng et al. [25] in the year 2015 which used content and user based features listed in the following: the number of followees, the number of followers, the number of messages, the number of friends following each other, the number of favorites, the number of created days, fraction of followees per followers, fraction of original messages, number of messages per day, the average number of reposts, the average number of comments, average number of likes, the average number of URLs, the average number of pictures, the average number of hashtags, the average number of user mentioned, fraction of messages containing URLs, fraction of messages containing pictures.

In this paper, they have introduced a machine learning based spa-mmer detection solution for social networks. The solution considers the user's content and behavior feature, and apply them into SVM based algorithm for spammer classification. Through a multitude of analysis, experiment, evaluation and prototype implementation work, have shown that proposed solution is feasible and is capable to reach much better classification result than the other existing approaches.

IV. CONCLUSIONS

Due to the increasing popularity and heavy use of social networks like Twitter, the number of spammers is rapidly grow-ing. This has resulted in the development of several spam detection techniques .From the papers reviewed it can be concluded that most of the work has been done using classification approaches like SVM, Decision Tree, Naive Bayesian, and Random Forest,KNN. Detection has been done on the basis of user based features or content based features or a combination of both.Work done by the X. Zheng[25] is significant and giving highest accuracy than the other existing approaches.

FUTURE WORK

Twitter has millions of active users and this number is constantly increasing. And almost all the authors have used very small testing dataset to see the performance of their approach. So there is a need to increase the testing dataset to see the performance of any approach. .there is need to improve classifiers for optimizing the detection rate.

REFERENCES

[1] Statista, ⟨http://www.statista.com/⟩.

[2] Nexgate. 2013 State of Social Media Spam, ⟨http://nexgate.com/wp-content/ uploads/2013/09/Nexgate-2013-State-of-Social-Media-Spam-Research-Report.pdf⟩, 2013.

[3] M. Uemura, T. Tabata, Design and evaluation of a Bayesian-filter-based image spam filtering method, in: Proceedings of the International Conference on Information Security and Assurance (ISA), IEEE, 2008, pp. 46–51.

[4] B. Zhou, Y. Yao, J. Luo, Cost-sensitive three-way email spam filtering, J. Intell. Inf. Syst. 42 (1) (2013) 19–45.

[5] J. Jung, E. Sit, An empirical study of spam traffic and the use of DNS black Lists, in: Proceedings of the 4th ACM SIGCOMM Conference on Internet Measure-ment, ACM, 2004, pp. 370–375.

[6] M. Antonakakis, R. Perdisci, D. Dagon, W. Lee, N. Feamster, Building a dynamic reputation system for DNS, in: Proceedings of the Third USENIX Workshop on Large-scale Exploits and Emergent Threats (LEET), 2010.

[7] Trust evaluation based content filtering in social interactive data, in: Proceed-ings of the 2013 International Conference on Cloud Computing and Big Data (CloudCom-Asia), IEEE, 2013, pp. 538–542.

[8] J. Kincaird, Edgerank: the secret sauce that makes Facebook's news feed tick, TechCrunch, 2010, ⟨http://techcrunch.com/2010/04/22/facebook-edgeran⟩.

[9] S. Yardi, D. Romero, G. Schoenebeck, Detecting spam in a Twitter network, First Monday 15 (1) (2009).

[10] G. Stringhini, C. Kruegel, G. Vigna, Detecting spammers on social networks, in: Proceedings of the 26th Annual Computer Security Applications Conference, ACM, 2010, pp. 1–9.

[11] A.H. Wang, Don't follow me: spam detection in Twitter, Security and Cryptography (SECRYPT), in: Proceedings of the 2010 International Conference on. IEEE, 2010, pp. 1–10.

[12] H. Gao, Y. Chen, K. Lee, D. Palsetia, A. Choudhary, Towards online spam filtering in social networks, in: Proceedings of the Symposium on Network and Distributed System Security (NDSS), 2012.

[13] F. Benevenuto, G. Magno, T. Rodrigues, V. Almeida, Detecting spammers on Twitter, in: Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-abuse and Spam Conference (CEAS), 2010.

[14] Y. Zhu, X. Wang, E. Zhong, N.N. Liu, H. Li, Q. Yang, Discovering spammers in social networks, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI), 2012.

[15] Kyumin Lee, James Caverlee, Steve Webb, Uncovering Social Spammers: Social Honeypots + Machine Learning, Proceeding of the 33rd International ACM SIGIR conference on Research and development in information retrieval, 2010, Pages 435–442, ACM, New York (2010).

[16] http://en.wikipedia.org/wiki/Twitter-Information of Twitter.

[17] Anshu Malhotra, Luam Totti, Wagner Meira Jr., Ponnurangam Kumaraguru, Virgılio Almeida, Studying User Footprints in Different Online Social Networks ,International Conference on Advances in Social Networks Analysis and Mining, 2012, IEEE/ACM.

[18] http://help.twitter.com/forums/26257/entries/1831- The Twitter Rules.

[19] http://about-threats.trendmicro.com/us/webattack-Information regarding Twitter threats.

[20] Alex Hai Wang, Security and Cryptography (SECRYPT), Don’t Follow Me: Spam Detection in Twitter, Proceedings of the 2010 International Conference, Pages 1-10, 26-28 July 2010, IEEE.

[21] Fabricio Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, Detecting Spammers on Twitter, CEAS 2010 Seventh annual Collaboration, Electronic messaging, Anti Abuse and Spam Conference, July 2010, Washington, US.

[22] Grace gee, Hakson Teh, Twitter Spammer Profile Detection, 2010.

[23] M. McCord, M. Chuah, Spam Detection on Twitter Using Traditional Classifiers, ATC’11, Banff, Canada, Sept 2-4, 2011, IEEE

[24] Ayon Chakraborty, Jyotirmoy Sundi, Som Satapathy, SPAM: A Framework for Social Profile Abuse Monitoring

[25]Detecting spammers on social networksXianghan Zheng , Zhipeng Zeng , Zheyi Chen , Yuanlong Yu , Chunming www.elsevier.com/locate/neucomputing