Automated Detection of Spammers in Twitter

Sanjeev Dhawan[1], Kulvinder Singh[2]
Faculty of computer [1,2], Department of Computer Science & Engineering, University Institute of Engineering and Technology, Kurukshetra University, Kurukshetra, Haryana
Email: rsdhawan@rediffmail.com , kshanda@rediffmail.com
Isha[3]
M.Tech. (Computer Engineering)[3], Department of Computer Science & Applications, Kurukshetra University, Kurukshetra, Haryana
Email: ishasingla587@gmail.com

Abstract
In the world of digital applications, a new application called twitter made a major impact in online social networking and micro blogging. The communication between users is through text based post. The open structure and its increasing demand have attracted large number of programs known as automated program also called as bots. One side of this is genuine bots, generates a large volume of nonthreatening tweets, e.g. blog updates/news which compiles with twitters goal of becoming a news information web. Other side of this is malicious bots have been greatly misused by spammers to spread spam. Spammers are the users who send unsolicited messages to a large audience with the intention of advertising some product or to lure victims to click on malicious links or infecting user's system just for the purpose of making money. A lot of research has been done to detect spam profiles in online social networking sites (OSNs). Features for the detection of spammers could be user based or content based or both. In this paper, an attempt has been made to review and analyzed the existing techniques for detecting spam users and their profiles in Twitter. Current study provides an overview of the methods, features used, detection rate and their limitations for detecting spam profiles mainly in Twitter.

Keywords –Human, Online Social Networking Sites (OSNs), Spammers, Twitter, Legitimate users

1. Introduction

Twitter is the red hot tool for micro blogging and social networking these days. Started in the late march of 2006 and twitter's off-the-wall the features makes twitter stand tall in this cyber world. As it is era of blogging, micro blogging and people connecting through social sites hence one cannot overlook online blogging and social networking site named Twitter which differs from traditional blogging and has vital add inns. It is a web application which gives users features like Direct Messaging, Following People & Trending Topics, Links, Photos, Videos message, image, or video links to share with their peers/colleagues and with followers such as personal online diaries or news on particular subject also one important aspect to notice is the small message refers to only 140 characters. These short messages are called tweets. These tweets are public by default and visible to all those who are following the twitter. Hash tags are those which starts with special characters # and which is meant to group similar micro blog topics such as #economics and #amazing. With larger user databases in OSNs, twitter is becoming a more interesting target for spammers/malicious users. Spam can take different forms on social web sites and it is not easy to be detected. Spam (www.spamhaus.org) is defined as the way of sending unwanted bulk messages via electronic mail system. With the rise of OSNs, it has become a platform for spreading spam. Spammers intend to post advertisements of products to unrelated users. As per twitter policy (http://help.twitter.com) indicators of spam profiles are the metrics such as following a large number of users in a short period of time or 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 there is no clear identification 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. Some spammers post URLs as phishing websites which are used to steal user's sensitive data. Our paper aims to provide a review of the academic research and work done in this field by various researchers. This paper is structured as follows: Section 2 briefed security issues in OSNs; Section 3 presents definition of spammers and their motives; section 4 describes motivation behind the survey paper; Section 5 reviews the work done by various researchers; Section 6 gives proposed work; and finally section concludes the review.

2. Security Issues in OSNs

Online Social Networking sites (OSNs) are vulnerable to security and privacy issues because of the amount of user information being processed by these sites each day. Users of social networking sites are exposed to various attacks:

Viruses: Spammers use the social networks as a platform to spread malicious data in the system of users.

Phishing attacks: In this approach user's sensitive information is acquired by impersonating a trustworthy third party.

Spammers: Send spam messages to the users of social networks.

Sybil (fake) attack: Attacker gathers numerous fake identities and acts as genuine so that to destroy the reputation of original users in the network.

Social bots: A collection of fake profiles which are created to gather user's personal data.

Clone and identity theft attacks: Where attackers create a profile of already existing user in the same network or across different networks in order to fool the cloned user's friends. If victims accept the friend requests sent by these cloned identities, then attackers will be able to access their information. These attacks consume extra resources from users and systems.

3. Types of Spammers

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. The main motives of spammers are to Spread viruses, phishing attacks, disseminate pornography and compromise system reputation.

Spammers belong to one of the following categories:

Phishers: The users who behave like a normal user to acquire personal data of other genuine users.

Fake users: The users who impersonate the profiles of genuine users to spend spam content to the friends of that user or other users in the network.

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

4. Motivation Behind Review

Because of the ease of sharing information and to be in sync with ongoing topics, Social networks have become a target for spammers. Detecting such malicious users in OSNs is difficult as spammers are very well aware of the techniques available to detect them, OSNs provide a perfect platform for spammers to disguise as a genuine user and try to get malicious posts clicked by normal users for sake of making money. So detecting such users in order to make network secure and keep the private information of users confidential is the most important topic being delved into by various researchers

5. Related Work

Twitter is a social networking site just like Facebook and MySpace except that it only provides a micro blogging service where users can send short

messages (referred to as tweets) that appear on their friend's pages. Twitter user is only identified by a username and optionally by a real name. The success of social networks has attracted the attention of security researchers. Since social networks are strongly based on the notion of a network of trust, the exploitation of this trust might lead to significant consequences. Identification of anomalous user types in Twitter data is an important precursor to detailed analyses of Twitter behaviors as they could incorrectly skew the results obtained in terms of topics prevalent in the population. Identification of specific types of users as different from the rest of the population is, in essence, a form of creating a profile of the user's interaction with the platform. Significant work has been done by Alex Hai Wang [1] in the year2010 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), Naïve 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. In year 2010, Lee *et al.*[2]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 behavior, 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 used for 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. Similarly Benevenuto *et al.* [3] 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 have 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, in year 2010, Grace and Hakson [4] 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. The CSV files were trained using Naïve Bayes algorithm with 27% error rate than SVM algorithm has been used with error rate of 10%. Spam profiles detection accuracy is 89.3%. Limitation of this approach was that not very technical features had been used for detection and precision was 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%.

IJSER

A forward step in the same field was taken by McCord *et al.*[5] using 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), Naïve Bayesian and K-Nearest Neighbor have been used to identify spam profiles in Twitter. Method has been validated on 1000 users with 95.7% precision and 97.5% accuracy using the Random Forest classifier and this classifier gives the best results followed by SMO, Naïve Bayesian and K-NN classifiers. Limitation of this approach was that for considered dataset reputation feature had 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. Then onwards in 2013, Lin *et al.*[6] detected long-surviving spam accounts in Twitter on the basis of two different features that are URL rate and interaction rate. Most of the papers have used lot many features for detection of spam accounts like no of followers, no of following, followers/following ratio, tweet content, no of hashtags, URLs links etc. But as per this paper all these features are not so effective features like URL rate and interaction rate have been used for detection purpose. URL rate is the number of tweets with URL / total number of tweets and interaction rate is the number of tweets interacting / total number of tweets. 26,758 accounts have been crawled using Twitter API and 816 long surviving accounts have been analyzed J48 classifier with 86% precision. Limitation of the approach is that only two features have been used for spam profile detection and if spammers keep low URL rate and low interaction rate then this technique will not work as intended. According to Amleshwaram *et al.* [7] there are two types of spammer detection techniques – users centric which are based on the features related to user like followers/following ratio and another is URL centric which depends on detecting malicious URLs. Approach mentioned in this paper is hybrid which considers above mentioned both types of features. 15 new features have been proposed to detect spammers, along with an alert system to detect spam tweets. Tweet campaigns and techniques used by spammers have also been studied. Two datasets from Twitter have been used one with 500K users and another with 110,789 users. New features that have been used are: Bait oriented features which identify the techniques used by spammers to lure victims to click on malicious links like no of mentions, mentions to non followers hijacking trends, intersection with famous trends. Behavioral features include variance in tweet interval, variance in no of tweets per unit time, ratio of variance in tweet interval to variance in no of tweets per unit time, and tweeting sources. URL features include duplicate URLs. Duplicate domain names, IP/domain ratio. Content entropy features include dissimilarity of tweet content, similarity between tweets, URLs and tweet similarity. Profile features include follower/following ratio, profile's description language dissimilarity. Thereafter all these features have been collected from malicious users as well as benign users which were then given to four supervised learning algorithms like Decision Tree, Random Forest, Bayes Network and Decorate using Weka tool. 93.6% spammers with false positive rate of 1.8% have been detected with Decorate classifier giving best results. This technique has been shown to outperform giving best results. This technique has been shown to outperform Twitter's spammer detection policy. But this technique has been tested on only 31,808 users whereas Twitter is considering millions of users. Similar in 2011, Chakraborty *et al.*[8] 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 considering for differentiating abusive and non-abusive users are- profile based, content based and timing based.

IJSER

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%. In 2014, Miller *et al.* [9] attempted to treat the identification of spammers as an anomaly detection and not classification problem where outliers are flagged as spammers. They utilize a combination of user metrics and one gram text features. They then test two algorithms: DBSCAN which uses a density based similarity metric and K-Means which uses an Euclidean distance based metric. These approaches achieved an 82% and 71% F1 score respectively with high accuracy but low precision. After that in 2015, M.A Fernandes *et al.* [10] compared classification and clustering approaches to separate human from not human users in Twitter. An initial feature set of 70 variables was reduced to the most relevant for classification, thereby decreasing complexity and improving generalization performance.

6. Proposed work

In order to achieve the target results with better accuracy, an efficient approach will be designed by modifying sequential K- Means clustering algorithm to detect spam in twitter. The accuracies can be achieved by reducing the size of feature space using stepwise feature selection and category balancing from manual inspection of classification results.

7. Conclusion

During survey it became quite apparent that a lot of work has been done for detecting spam profiles on different OSNs. Still improvements can be made to get better detection rate by using a different technique and covering more and robust features as deciding parameter. So following are the few conclusions drawn from survey:

- Since Twitter have 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 need to increase the testing dataset to see the performance of any approach.
- Need to develop a multivariate model.
- Need to develop a method that can detect all kinds of spammers.
- Need to test the approaches on different combinations of spammers and non-spammers.

Many methods have been developed and used by various researchers to find out spammers in different social networks. From the papers reviewed it can be concluded that most of the work has been done using classification approaches like SVM, Decision Tree, Naïve Bayesian, and Random Forest. Detection has been done on the basis of user based features or content based features or a combination of both. Few authors also introduced new features for detection. All the approaches have been validated on very small dataset and have not been even tested with different combinations of spammers and non-spammers. Combination of features for detection of spammers has shown better performance in terms of accuracy, precision, recall etc. as compared to using only user based or content based features.

References

[1] 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.

[2] 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).

[3] 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.

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

IJSER

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

[6] Po-Ching Lin, Po-Min Huang, A Study of Effective Features for Detecting Long-surviving Twitter Spam Accounts, Advanced Communication Technology (ICACT), 15th International Conference on 27-30 Jan. 2013, IEEE.

[7]Amit A. Amleshwaram, Narasimha Reddy, Sandeep Yadav, Guofei Gu, Chao Yang, CATS: Characterizing Automation of Twitter Spammers, Texas A&M University, 2013, IEEE.

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

[9] Miller, B. Dickinson, W. Deitrick, W. Hu, and A. H. Wang. Twitter spammer detection using data stream clustering Technical report, Department of Computer Science, Houghton College, Houghton, NY, USA, 2014.

[10] M.A Fernandes, P.Patel, and T. Marwala, Automated detection of human users in Twitter, Department of Electrical and Electronic Engineering University of Johannesburg, Johannesburg, Gauteng, south Africa, volume 53, 2015, pages 224-231, 2015 INNS Conference on Big Data

IJSER