Second International Congress on Technology, Communication and Knowledge (ICTCK 2015)
November, 11-12, 2015 - Mashhad Branch, Islamic Azad University, Mashhad, Iran

Spam Detection In Social Networks: A Review

Nasim eshraqi
Department of Software Engineering
Mashhad Branch, Islamic Azad University
Mashhad, Iran
nassim.eshraqi@mshdiau.ac.ir

Mehrdad Jalali
Department of Software Engineering
Mashhad Branch, Islamic Azad University
Mashhad, Iran
jalali@mshdiau.ac.ir

Mohammad Hossein Moattar
Department of Software Engineering
Mashhad Branch, Islamic Azad University
Mashhad, Iran
moattar@mshdiau.ac.ir

***Abstract*— The social networks provide a way for users to maintain contact with their friends. Increasing of the social network's popularity allows all of them to collect large amounts of personal information about their users. Unfortunately, this information wealth as well as its ease to access users' information could attract malicious group's attention. That's why these networks have been invaded by spammers while, there have been a lot of work to diagnose and fix them. With regard to this issue that spammers look for new ways to target these networks every day, there have being continuous actions to identify spammers and malicious email. The aim of this paper is to examine previous works in the field of spam detection in social networks.**

Keywords: Spam detection; Social Networks; Twitter; Facebook.

I. INTRODUCTION

In recent years, some of sites such as Facebook and Twitter have placed among 20 websites with the highest internet visitors. Most of the social networks have offered mobile operating system, which allow users to access services the sites via the cell phone so they can possibly access them everywhere [1].

With the advent of social plug-in such as Like on Facebook or tweet on twitter, integration between social media and the websites is difficult. Usage of social networks with the web voluntaries can induce spammers for malicious operation [2].

Social networks usually regulate that personal information is visible to whom. From the security perspective, the social networks have unique features. Users share usually a significant part of their personal information with their friends. This information may be private or public. If information is private, so accessing to it is regulated by a reliable network. In this condition, the user only allows friends to see the information. Unfortunately, sites on the social networks don't provide mechanisms of a strong identity confirmation, therefore the user pervades hidden in the account of another person. In addition, usually the users get the popularity accept any friend requests and disclose their personal information for the unknown people. Most users of the social network sites, click easily on sending links by friends who, even if they do not know that person in their own real life. This behavior may be abused by spammers who advertise a website [3].

The available filtering techniques, such as Transactional Analysis Filters can reduce spams considerably. Each of the social networks needs to build a dependence spam filter for itself and currently they support from one spam team to maintain a spam prevention technique [4].

In this paper, we will consider various articles about spam detection on social networks. For this purpose, we examine the features of spam user account detection in the second section. The third part is devoted to the features of spam post detection. Fourth section refers to articles about classification and clustering methods for detecting spam and the results have been examined and the results are analyzed and concluded about the research in section five.

II. SPAM PROFILE DETECTION FEATURES ON THE SOCIAL NETWORKS:

Various papers have done in the field of spam detection on the social networks.
Each of these studies has raised one or more features for spam detection. Some articles have written just for a social network and some have examined different networks. In addition, some have written about the spam user accounts detection and some were about spam wall post detection in the social networks.

We will study each of these cases individually in different parts.

In 2010, Alex Hai Wang [5] divided the features of the Twitter spammers into two groups: content -based and graph-based features and described the comment completely about spammer detection.

In content-based features part, it was mentioned there are four features to detect spam user account.

- **Repetitive Tweets**: If a user account sends repetitive tweets, it will be known as a spammer.
- **Links**: If most of the sent tweets from a user account contain the link, they will be known as spammer.

978-1-4673-9762-9/15/$31.00 ©2015 IEEE

- **Trending topics**: if a user account sends unrelated material to trending topics, it will be known as a spammer.
- **Replies and Mentions**: If most of the tweets sent from a user account containing the replies and mentions, known as a spammer.

In graph-based features, Twitter can be modeled as a social graph model. Spammers follow a large number of users, but those who follow them are limited. This ratio is called follow rate or credibility degree. Surely, if flowers are lower than followings, so credibility will be low and near to zero and it will be a spam account with a high possibility.

In 2010, Gianluca Stringhini et al [3] considered six features for spam detection on Facebook, Twitter and MySpace.

- **FF ratio (R)**: ratio of sending friend requests number to the number of users that have accepted the request is intended as a measure for spam detection.

  Because spam isn't a real person, so anyone knows it in real life and only a part of the user account accepts friend requests. Therefore, there is a clear difference between the number of sent friend requests by a friend and the number of those requests has been accepted.
- **URL ratio (R)**: the second feature for spam detection is presence of URL in wall posts. For the user's attraction to spam web pages, spammers send links to own wall posts. Therefore, link ratio is defined as the number of wall posts containing links of a user account to the total wall posts of the account, which is a criterion for spam detection. Since most messages in Facebook contain a link to other pages of Facebook, for computation of this feature, we should calculate just a URL that is related to a site.
- **Message similarity (S)**: the third feature is penetration of similarity between sending messages by a user. In terms of two features of message size and content, more spammers, send very similar messages. Parameter similarity S is defined to:

$$S = (\sum p \;\; P * c(p)) / (l_a \, l_p) \qquad (1)$$

  That P is a set of possible message-message compounds between each of two entered messages for one special account, p is a unique couple. P(c) is a function of words number calculation for two shared messages, in average la is length of messages sent by a user and lp is some of message compounds. The idea in the back of this formula is that at a profile which is sending similar messages, S has low value.
- **Selecting and searching by friends (F)**: the fourth feature, is related to this issue that, whether there are user accounts that have searched the determined account as their friend or not. This property is called F and defined to:

$$F = T_n / D_n \qquad (2)$$

  That T_n is the total number of names among the friends' user accounts and D_n is different name number. Legal user accounts have this property with the value near to 1 however, spammers may reach to 2 value or more.

- **Sent messages (M)**: observations show that spammers send hundreds of messages.
- **Number of friends (FN)**: finally, the number of friends is a searched profile.

In 2011, Saeed Abu Nimeh and his colleagues [6], were examined spammers by paying attention to the chronological and geographic properties. They reported, spammers check intervals by using the social networks to understand in what hours of the day, more users use these networks and in those hours of the day, they send remarkable volume of the spam contents.

Researches showed volume healthy of wall posts increases steadily every day between certain hours. Maybe it's why people often turn to social networks after work. Malicious wall posts are increased in the early evening and they are reduced in the early morning. The model also indicates irregularities.

If the user account sends a huge amount of tweets in a short period of time, it will be known as spammer.

Also they explained that, spam links are a little part of total volume of messages, but they distribute asymmetries at various countries and this shows that spammer's activities are widespread in term of geographic condition. Studies also show domain of spam posts mainly is widespread. It seems first hosts of spammers and malicious posts are some western countries with maximum technology, but a little number of small countries, such as Lithuania, is very low host to with malicious and illegal posts or they aren't host for such posts.

In 2011, De Weng et al [4], offered a framework for social spam detection that can be used by every social network for spam detection. This framework defines three models of object representation in the social networks, namely profile model, message model and web page model.

Profile model which is defined here has 74 features obtained by the Google API. Selected features can cover generally features are used by user account in all web sites such as Facebook, MySpace, Twitter and Flicker.

Defined message model has 15 features based on common features usable in messages, such as "To", "From", "Timestamp", "Subject" and "Content". Also, there are several features in message model that are shared with a social network and also receipt email messages, for example, receiver IP and other features of the header.

Web page model has features based on header content information HTTP .For example, "Connection", "Content-length", "Server" and "Status" are features of HTTP. All of these features are used to identify spam user accounts.

In 2012, Krishna Chaitanya and her colleagues [2], studied about the modern technique of spam detection. In this research, they discussed four modern popular techniques which are used by the invaders to make spam in the social networks sites, including clickjacking, malicious browser extensions via drive-by-downloads, URL shorteners and socially engineered script injection.

Spammers use several methods to commit theft users' information. For example, a spammer can create a counterfeit profile, then introduce itself as individual's friend. When a victim accepts the friendship request, spammer can steal personal information.

In another example, a spammer can send a destructive link and encourage the user to click on it. Clicks are directed to various goals. The method is called click theft. Similarly, a spammer may recommend a counterfeit video page to a user, and the user is forced to watch it by downloading a plug-in. When the user accepts it, in fact, he/she downloads a Browser Extension, that it injects destructive JavaScript and control victim's account. It is an easy way to download.

Spammers also apply certain engineered script injection of destructive JavaScript. They use shorteners link for spam attacks on the social networks. Since shorteners URL can make the real link complex, users can't understand who may send them.

In 2013, Faraz Ahmed & Muhammad Abulaish [7] suggested a statistical method for spam detection in the online social networks. They determined a set of 14 statistics features that are common in Facebook and Twitter for spam accounts detection.

Set of Facebook features include the following:

- **Related features to the wall posts in friends pages** (total number of user-submitted posts on friends 'walls, maximum user-submitted posts on friends' walls and the number of posts on friends 'walls separately)
- **Related features to links** (total number of shared links by a user, the number of links which are shared at least once by the user and the average repetition links)
- **Related features to tags** (total number of tags per post shared by a user, the number of users and pages that are tagged and the average tags in each post)

Twitter feature sets include the following:

- **Related features to the activity of Twitter** (follower's number of a user and Hashtag that a user is involved in it. Hashtags are similar to liked pages in Facebook.)
- **Related features to tweets in user's Hashtag** (total number of Hashtags, which are used by a user, the maximum amount of histogram repetition in Hashtags used by a user and the average number of histogram in each of Hashtag which is used by a user)
- **Related features to information related to Mentions** (total Mentions number made by a user in his/her tweets, the total number Mentions made by a user in his/her tweets and others and total Mentions made for each friend)
- **Related features to links** (the number of shared links by a user, the number of individual links of a user and the average number of repetition links)

In 2014, Zachary Miller and his colleagues [8], used the following features for spammers detection:

The number of followers, the number of friends, the number of interests, the number of lists, the number of tweets, the number of retweets, authenticated user, date, rate of followers, the number of links, the number of Mentions and answers, the number of Hashtag and 95 features are chosen from the full ASCII set for accessibility on a standard US keyboard

In 2013, Marcel Flores and her colleagues [9], offered a system for the detection of malware and spam accounts by using a web browser. Their system used web search to measure their online presence of users and user accounts that are underrepresented on the web, and it is considered as a spammer. In fact, the system itself is not used for spam detection and it is used in combination with learning techniques more expensive machine is used as the first stage for detection and in addition to reducing costs, leads to more effective spammer's detection and in fact, it is used as an auxiliary source to detect spammers.

Firstly, input data with search module are taken from Twitter. This module uses a web browser for the Username. Twitter doesn't need meaningful names for user accounts and these names can be a corporate name or titles. After the search, whole results enter into analysis phase and in this phase, a number of noise reduction techniques, are used to remove search results have been achieved from Twitter users without effect on being spam or health accounts detection. Finally, the analysis module checks the remaining results for each user account and if there isn't any results about searching username and name, the accounts is considered as a spammer otherwise it is considered as a normal account.

Also, a black list of usernames and displayed names is provided. The black list includes 10 domains which have been detected in search as a noise and have been removed in analyzes.

In 2012, Faraz Ahmed & Muhammad Abulaish [10], used an approach based on Markov Clustering (MCL) for spam profiles detection on the social networks and they worked on a real dataset network of Facebook. They modeled the social networks with a weighted graph, in which user accounts were considered as nodes and their interactions as the edges of the graph. The weight of an edge, have been defined as a connection of a pair of user profiles and it shared as a function of actual social interactions between active friends, liked page and shared URL. MCL is applied on the weighted graph and lead to produce different clusters such as various categories of user accounts, which include spam user accounts and normal user accounts.

III. SPAM POSTS DETECTION FEATURES ON THE SOCIAL NETWORKS

In 2010, Hongyu Gao and his colleagues [11], examined a way for diagnosis and posts description and campaigns of social spam.

Study on wall posts on Facebook contains two phases. In the first phase, all wall posts are analyzed and focused on posts that contain a link. In the second phase, features of destructive wall post are analyzed. To detect wall posts of spam, it is used semantic similarity criteria. Then, for differentiation identification in normal wall posts or spam posts behavioral signs are used. Each single account can send special number of wall posts, so spammers must apply more user accounts to great campaigns. Spam campaigns must be detected from the counterfeit accounts before identification there to obtain maximum productivity; as a result, in term of time, the posts of every campaign are sequenced.

In 2011, Kristofer Beck [12] analyzed Tweets for identification of malicious tweets. He argued that there are some special words and expressions in tweets that indicate tweets are spam. By identifying these words, we can assess the probability of tweets spam. So there are 5 steps that should be taken:

- Is tweet contains links or not? (X0)
- Is the message contains the word "Chat"? This word has various applications in the spam tweets. (X1)
- Is the message contains the word "With"? The word "Chat" is used often associated with "With". (X2)
- Is the word "Chat" in user's biography or not? (X3)
- The word "naughty" has direct relationship with biography. (X4)

Now, by using the following formula, possibility of being spam of tweet can be examined:

$$z = X_1 + X_2 + X_3 + X_4 + \dots \quad (3)$$

$$\beta = 1 / (1+ e^{-z}) \quad (4)$$

Where β shows risk of malicious messages and it is probability of being spam. According to the conducted survey, z has a minimum and a maximum that sequentially are one and six and if it is between 2 and 3, it will be considered as a spam. In fact, accurate, its average for the spam tweets is 2.33. So, by calculating the values, we can examine probability of being spam of the words.

In 2013, Juan Martinez-Romo and his colleagues [13], offered a way based on two new aspects for spam tweets detection. One of them is to detect spam tweets in isolation and without previous information and the other one is detection of spam tweets by using statistical analysis of language in trending topics.

In addition, they have developed a machine learning system with orthogonal features, which can be used in combination with other features for spam detection.

They have considered 12 content-based a series of statistical properties of language features for this purpose.

Content-based features include the following:

"The number of links in each word, the number of Hashtags in each word, the number of words, the number of characters, the number of links, the number of the Hashtags, the number of the numeral characters, Mentioned users, the number of words which are member of list of spam words, the number of words, the number of times have been answered to a tweet, the number of written tweets on a topic by the user and last time that user posted a tweet."

They also offered several new features based on a language model to identify spam on Twitter. Firstly, they found popular topics and related tweets for them and then divided tweets into two health tweets and suspicious tweets. Suspicious tweets have a link to a web page. Then all suspected tweets links and tweets were analyzed and tweets were classified into spam and nonspam. In terms of content, spam tweets are different from popular topics and the same divergence is used between the different language models for their detection. Since opening of each linked web page requires a lot of time and cost, this paper analyzed only the title of pages. It also examined divergence between previous tweets of any user and its relationship with different issues.

Text mining of suspicious tweet in trending issues is very important. Then, 10 tweets before and 10 tweets after will be examined to determine, target tweet has a semantic relationship with the issue or not. Also, other tweets are sent by a user on a page and its relationship with and popular issues are examined. Then it is modeled as linguistic characteristics of the model. The cost of the operation is very low, but has a considerable impact.

IV. SPAM DETECTION METHODS

Methods that have been used for spam detection generally can be classified in two groups: classification and clustering methods. In this section, we examine different ways of articles for spam detection. Table 1, showed the results of spam detection process for various parameters and algorithms in the social networks such as Twitter and Facebook.

A. Classification Methods

In most previous works in field of spam detection in the social networks, classification methods were used. In [5], it is compared various Classification algorithms such as decision tree, nerves networks, support vector machine and k-nearest neighbor. The Bayesian classification algorithm is used to detect suspicious behaviors from common behaviors, data analyses and examination of the spam detection system in Twitter. Results showed Bayesian classification algorithm has the best performance to measure F and detection ratio.

In [4], the Bayesian classification algorithm is used to detect spam in Twitter network. The values in table 1 for article [4] were obtained according to Table 11 from this article.

In [7], 3 Bayesian classification algorithm are used, J48 and JRIP to detect spammers in the networks such as

Table1-Algorithm' result

Refrences	Detection method	Netwok's name	Algorithms	Precision	Recall	F-measure	Accuracy	FPR
[6]	Classification	Twitter	Decision Tree	0.667	0.333	0.444	-	-
			Neural Networks	1	0.417	0.588		-
			Support Vector Machines	1	0.25	0.4	-	-
			Naïve Bayes	0.917	0.917	0.917	-	-
[4]	Classification	Twitter	Naïve Bayes	0.9187	0.799	0.8546	0.8642	0.07
[8]	Classification	Twitter	Naïve Bayes	-	0.976	-	-	0.075
			Jrip	-	0.987	-	-	0.014
			J48	-	0.983	-	-	0.017
		Facebook	Naïve Bayes	-	0.964	-	-	0.089
			Jrip	-	0.912	-	-	0.09
			J48	-	0.898	-	-	0.081
		Combined	Naïve Bayes	-	0.733	-	-	0.309
			Jrip	-	0.935	-	-	0.071
			J48	-	0.957	-	-	0.048
[10]	Classification	Twitter	Supplements Machine Learning	0.7396	0.7467	0.7431	0.7902	0.1067
[13]	Classification	Twitter	SVM	0.8732	0.893	0.883	0.922	0.063
[9]	Clustering	Twitter	DenStram	0.7272	1	0.8421	0.9711	0.31
			StreamKm++	0.5591	1	0.7172	0.9393	0.65
			Combine	0.7939	1	0.8851	0.98	0.21
[11]	Clustering	Facebook	MCL			0.88		

Facebook and Twitter. The values in table 1 for article [7] were obtained according to Table 3 from this article.

In [9], a subsidiary and complementary system is used for the machine learning classifier to improve spammer's detection performance on Twitter.

In [13], the machine learning classifier is used for spam detection tweets on Twitter.

B. Clustering methods

In [8], it is used ongoing data clustering method for spammer detection on Twitter. Algorithms like DenStream and StreamKM ++ algorithms are used in combination.

V. CONCLUSIONS

In general, it can be said, for spam detection on social networks, at first we must understand the network and its features. For this purpose, network should be considered from different angles. Then features of spammers and spam posts should be determined. After determining the features, a proper way for spam detection should be chosen. Clustering and classification methods have a lot of algorithms. Therefore, the most appropriate algorithm should identify and work on it.

REFERENCES

[1] Benevenuto, F., et al. *Detecting spammers on twitter*. in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. 2010.

[2] Krishna Chaitanya, T., et al. *Analysis and detection of modern spam techniques on social networking sites*. in *Services in Emerging Markets (ICSEM), 2012 Third International Conference on*. 2012. IEEE.

[3] Stringhini, G., C. Kruegel, and G. Vigna. *Detecting spammers on social networks*. in *Proceedings of the 26th Annual Computer Security Applications Conference*. 2010. ACM.

[4] Wang, D., D. Irani, and C. Pu. *A social-spam detection framework*. in *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*. 2011. ACM

[5] Wang, A.H. *Don't follow me: Spam detection in twitter*. in *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*. 2010. IEEE.

[6] Abu-Nimeh, S., T. Chen, and O. Alzubi, *Malicious and Spam Posts in Online Social Networks*. Computer, 2011. **44**(9): p. 23-28.

[7] Ahmed, F. and M. Abulaish, *A generic statistical approach for spam detection in Online Social Networks*. Computer Communications, 2013. **36**(10): p. 1120-1129.

[8] Miller, Z., et al., *Twitter spammer detection using data stream clustering*. Information Sciences, 2014. **260**: p. 64-73.

[9] Flores, M. and A. Kuzmanovic. *Searching for spam: detecting fraudulent accounts via web search*. in *Passive and Active Measurement*. 2013. Springer.

[10] Ahmed, F. and M. Abulaish. *An mcl-based approach for spam profile detection in online social networks*. in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2012 IEEE 11th International Conference on*. 2012. IEEE.

[11] Gao, H., et al. *Detecting and characterizing social spam campaigns*. in *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement*. 2010. ACM.

[12] Beck, K. *Analyzing tweets to identify malicious messages*. in *Electro/Information Technology (EIT), 2011 IEEE International Conference on*. 2011. IEEE.

[13] Martinez-Romo, J. and L. Araujo, *Detecting malicious tweets in trending topics using a statistical analysis of language*. Expert Systems with Applications, 2013. **40**(8): p. 2992-3000.