International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 3, March - 2014

Detection of Spam links in Twitter

B. Mathiarasi
PG Scholar, Computer Science and Engineering,
KCG College of Technology,
Chennai, India

Dr. Emilin Shyni
Associate Professor, Computer Science and Engineering
KCG College of Technology
Chennai, India

***Abstract*—Online Social Networks have gained popularity in a short span of time. The increased popularity of these networks has led to spammers using these networks to spread spam content. Spam content is spread by sending spam URLs (Uniform Resource Locators) to friends in Twitter. The friend's account that sent the URL might have been compromised. This is one of the fastest ways to spread spam. It is not possible to identify a spamming link just by looking at it. Moreover, URL shorteners can easily hide the real link and make them pose as though they are harmless ones. The onset of URL shorteners have made it an easy possibility to spread spam and the identification of spam, very difficult. This work concentrates on identifying spam URLs that are present in Twitter profiles. Based on an exhaustive set of features, the links that are embedded in Twitter profile are classified as spam or non-spam. The classification accuracy has been verified using linear classification method.**

Keywords—Online social networking; Spam; Twitter;Classifier;

I. INTRODUCTION

The phenomenon of online social networking has increased tremendously in a short span of time. Of the various online social networking sites available, Twitter and Facebook are the most important and the popular ones. In Facebook, communication takes place by sending posts to friends. In the same way, in Twitter, information sharing is done through tweets. Friends in Online Social Networks otherwise called as OSN share information by sending links to their online contacts. People generally trust links that are sent by friends in an Online Social Network rather than the links that are sent through email by unknown people. The online contacts may not be known to the person in real life. This is so because, there are a lot of unwanted, unsolicited friend requests that come to a profile. In order to gain popularity in social networking sites, people, without analyzing, accept such anonymous friend requests. Thus, they are exposing their personal information to unknown people. In [2], it is shown that 45% of users of online social networking sites click on the links sent by their online contacts or friends/followers even if they do not know them personally. Spammers use this opportunity to spread spam links. When these spam links are clicked, it takes the user through a series of web pages which ask the user to enter private information. When the user enters private information, this information is gathered and used illegally. Such an attack is called phishing. It is a method by which certain sensitive information of the users such as username, password, bank account details, credit card information etc, is maliciously extracted electronically by posing as a trustworthy entity. Phishing is achieved by spreading links which direct the user to fake web sites requesting for sensitive user information. These web sites look very similar to the real ones. Hence the user is persuaded to enter his information. This information entered by the user is used for malicious purposes.

Both Facebook and Twitter are widely used nowadays. Recent statistics shows that, Facebook has over 1.23 billion active users which is a 16% increase over last year. Another most important social networking sites is Twitter. Twitter restricts the length of user message or tweets to 140 characters. This has led to the popularity of various URL shortening services. These services shorten the link submitted by the user. Using this facility, the spam URLs can be masked as a URL coming from a trusted site. When a user clicks on such a URL, it takes the user through malicious websites which try to extract the user's private information. The works related to profile based spam detection generally identify the spam accounts but they do not have a great utility in identifying those accounts that are compromised. Generally compromised accounts are dangerous and are used extensively for spreading spam. A Twitter user is more likely to click on the link sent by his follower. If the follower's account is compromised, then, there is a great possibility that the Twitter profile of the user clicking on the link will be compromised as well.

In order to address this issue, this work crawls URLs that are present in a Twitter profile to check if they direct to spam or not. A large number of tweets from different Twitter profiles have been taken into consideration. This work considers spam as phishing attempts that extract an account's private information, pharmaceutical scams and HTML pages that distribute malware. A large number of features are considered for the detection of spam links. These features operate on a large data set consisting of tens of thousands of URLs. In order to ensure accuracy, access to every link that is used to construct a landing page, HTTP headers and HTML content is needed. A linear classification method is used which achieves a great accuracy of 99.3% and a very small false positive rate of 0.7%. The salient features of this work include:

- A significant work for the detection of spam links present in tweets posted by a Twitter profile.
- A new classification architecture which uses a modified browser and a linear classifier which can scale up to a large number of features.

The main use of this work is that if it is known that a profile consists of malicious links that attempt to spread spam, then, the profile can be blacklisted. Thus this work can help in building up a database consisting of spam Twitter profiles which can later be blacklisted.

The paper is organized as follows: Section II talks about the related work done by others. This is followed by design and architecture of the proposed system which is dealt with in section III. Data collection methodology is discussed in detail in section IV. Following this, a detailed analysis of features is presented in section V. Section VI deals with the classification of URLs where the classifier used for the segregation is discussed in detail. The experimental setup and evaluation results are provided in Section VII. The paper is concluded in section VIII which also discusses the future work that is proposed to be done. This is followed by the references used by this work.

II. RELATED WORK

The enormous amount of information residing on online social networking sites has lured researchers to extract this information and study the problems that are faced by the social network community. Many works have been done for collecting and extracting information for various problems such as community detection, diffusion of information, and filtering spam content. Online social networking is gaining popularity because of which they attract a lot of spammers who misuse these networks in order to spread spam. Because of this reason, lots of works are involved in identifying spam on these networks. It is shown in [3] that about 10% of links that are posted in Facebook are spam and 97% of profiles which participate in spam campaigns are accounts that are compromised rather than fake accounts which are created for the specific purpose of spamming. According to [4], 8% of links posted in Tweets are spam which lead to phishing, malware and scams and 86% of the accounts that are a part of campaigns are compromised. This work also shows that blacklists are very slow in detecting new spamming threats. In order to overcome these attacks, several solutions have been proposed which classify spam profiles based on the frequency of posts, the number of links on profile posts and the ability of a profile to obtain friends [1],[5] and [8]. But, many of the features proposed to detect spammers are evasive in nature which can be evaded by spammers very easily.

This work is built on several significant works which propose spamming properties. One is the Lexicographical characteristics of the URL which is available in [6] and the other is the content of spam sites specified in [7]. In [9], spam links in general web services are studied including email and tweet spam. This approach is a more generic one where a system is proposed to filter the spam links in tweet streams as well as email. In contrast to this, our work concentrates only on Twitter. A combination of features proposed in many other works have been considered for analyzing the links in tweets and to find out if they are spam or non-spam.

III. DESIGN AND ARCHITECTURE

Classification of URLs as spam consists of several stages. These stages are portrayed in the architecture diagram of the system which is shown in Fig. 1. It consists of URL collection phase where the tweets from Twitter profiles which consist of links embedded in them are collected. This is followed by the feature collection and feature extraction phases. The final stage is the classification stage which gives the classification result and the detection rates and the false positive rates. The detection rate provides the accuracy of the classification mechanism. The design of this system includes a Twitter crawler that extracts all the URLs that are present in a profile. These are the links that are present in the tweets posted by the profile. All the URLs are collected and the system visits each and every URL to collect information such as page content, behavior of a page and related raw data. The data thus collected are converted into meaningful features. These results are then fed to the classifier for detecting spam. The classifier provides the detection rates and false positive rates which can be used to check the efficiency of the proposed URL spam detection method. The various stages in the spam detection in URLs are listed in the following subsections:

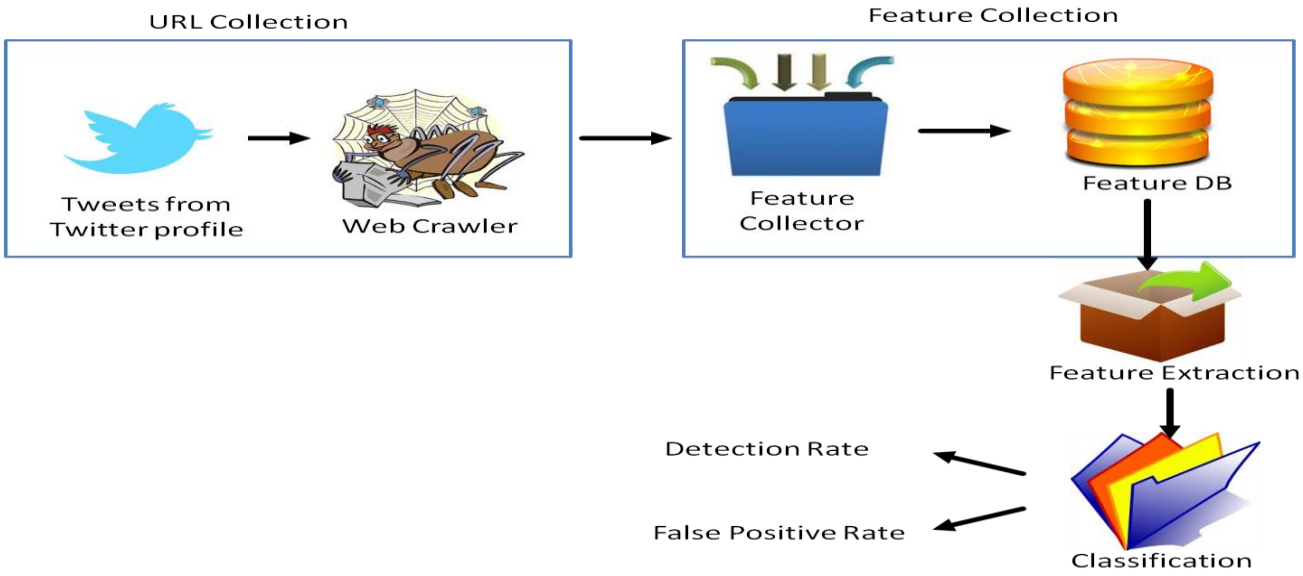

Fig. 1. Architecture of URL spam detection system

A. URL Collection

The system collects URLs from Twitter profiles. About 405 public Twitter profiles are considered for this purpose. A Twitter crawler crawls through these profiles in order to collect URLs that are present in the Twitter profiles. The output of this phase is the list of URLs that are present in tweets posted by a Twitter profile. A total of 41,890 URLs were collected from 405 Twitter profiles. Both the contextual information about the account as well as the tweets associated with the URL are collected as well.

B. Feature Collection

Every URL that is collected is visited using a modified web browser and details such as the contents of the page that includes HTML as well as page links are collected. Also, the page behavior including JavaScript activity and pop up windows are monitored. These raw data are collected which forms the content of the database. Feature collection deals with this type of raw data collection pertaining to the features that are discussed further in this section.

C. Feature Extraction

The previous phase collects all the details which are raw. The raw data which is collected cannot be processed as such by the classifier. This raw data has to be formatted to suit the requirements of classification mechanisms. The URLs are tokenized into binary features and the HTML content is converted into a set of words. To prepare for the next phase which is the classification stage, the features are modified into a meaningful feature vector as required by the classifier that we intend to use. The URL is split into various components that include domain, path and query parameters each of which is tokenized by splitting when non-alphanumeric characters are encountered. An equivalent process is applied to HTML and text strings such as HTTP headers. In these cases, the text is tokenized into a bag of words. The result of tokenization is converted into a binary feature vector. For each term that is present, a flag is set which indicates its presence.

D. Classification

This is the phase which classifies the URL as spam and non-spam. This is done based on the feature set that are fed into the classifier. The classification depends on the lexicographic properties of URL as well as the content of the page that is loaded. The final decision of the classifier depends on the individual values for the set of features that have been extracted from the given set of URLs. Each of these features plays an important role in the classification process. The robustness of the classification method depends on the list of features that are used to differentiate the spam URLs from the benign ones. These features have been carefully chosen after analyzing the past history as well as the means employed generally by spammers in order to lead the user to spam content. The features are exhaustive and been taken after referring to many previous works as well as adding features which we thought to be relevant ones which would help in distinguishing the normal URLs from the spam URLs.

The list of features that are used for classification is given in Table 1. Each of these features has been gathered after analyzing the importance of them in the identification of spam URLs. These features are analyzed in detail in the fifth section.

TABLE 1. List of features

Feature Name	Feature description
URL features	• Domain names • Path names • Query parameters • Number of subdomains • Length of the domain • Length of the path • Length of the URL
Redirect features	• URL features mentioned above for each redirect • Number of redirects • Type of redirect
HTML features	• Main HTML tokens • Content of the script
Pagelink features	• URL features for each of the links • Number of links • Ratio of Internal domain to external domains
JavaScript Event features	• Number of User prompts • Prompt tokens • Whether “onbeforeunload” event is present
Pop-up window features	• URL features for each pop up window URL • Number of pop up windows • Reason for the pop up of window
HTTP headers	• Field name and field value tokens

All these features play an important role in the identification of spam URLs. Spam URLs typically exhibit certain values for these features that help in detecting them. These values have been derived at by carefully studying spam nature of URLs as well as past instances of the presence of such features in spam URLs.

IV. DATA COLLECTION

The aim of the work is to identify spam URLs in the Twitter profiles. For this, we use a large dataset consisting of 405 profiles. These are public Twitter profiles. Only public Twitter data are collected which are available for public view. The profiles from which tweets are considered are chosen randomly. Sometimes, highly interactive profiles are chosen. Also, profiles that do minimal interactions are also chosen. This is to ensure that a wide range of tweets are considered for our evaluation. The profiles are also randomly chosen to prevent any bias against specific types of profiles and to ensure that there is a right mix up of normal profiles as well as spamming profiles. A typical example of spam tweets consisting of spam links taken from a compromised account is specified in Fig. 2.

Such spam tweets somehow make a user to click on to the link attached to the tweet. It is done by expressing surprise in finding some information about the user or by saying that there is a funny picture of the user in the net. These are some of the persuasive techniques that enable a user to click on the link without giving it a second thought. Such are the tricks that are usually employed by spammers. Other tweets say that some important information regarding updating of user accounts is needed.

@xxxxxxxx lol I had a weird feeling this was you http://t.co/osJZlQU544 Jan 21, 2014

@xxxxxxxx lol this was done by you? http://t.co/AWWKpbAJDi Jan 21, 2014

@xxxxxx I'm laughing so hard right now at this... http://t.co/6ChcX7InpL

@xxxxxx LOL you gotta read this, its epic http://t.co/OA6mUCo0qO Jan 21, 2014

@xxxxxx haha this blog by you is nuts http://t.co/RNLkY7NOTr Jan 21, 2014

Fig. 2. Sample spam tweets taken from a profile

All the links in each of the profiles are considered for analysis. A compromised account has normal tweets as well as spam tweets. So, all the tweets in a profile have been considered for our study. About 41,890 URLs were considered for spam detection. These URLs were present across all the 405 profiles that were collected.

V. ANALYSIS OF FEATURES

This section analyses the features that are collected by the feature collector. Each of the features is described in detail here. The values of these features help in distinguishing between a normal URL and a spam URL. These specific values are studied and analyzed from past experiences and blacklists that are available.

A. URL Features

The lexicographic features of a URL are typically indicative of whether a URL is spam or not. The length of a URL, number of sub domains and the words that appear in the URL can indicate whether the URL is spam. But, because of the presence of nested URLs and the availability of URL shortening services, it is impossible to determine the spamming nature of the URL. Hence for each URL, the initial URL as well as the final landing page of the URL is logged by the browser. The final landing page is arrived at after the execution of any redirects.

B. Redirect features

Apart from the initial and final landing pages, the chain of redirects also can provide details about the suspicious nature of the final landing page. Extremely long redirect chains, those that go through already known spam domains and the ones that are generated by JavaScripts and plug-ins, provide details of whether a final landing page is spam. In order to know these details, it is not only necessary to log the initial and final landing URLs but also all the redirects that lead to the final landing page. The monitoring also includes the reason for the redirect. This reason can be a HTTP response, a JavaScript event or a plugin.

C. HTML features

Sometimes the content of a page also indicates the presence of spam. The contents which denote spam include certain terms that appear on a page and layout similarity across spam pages. For this, the web browser saves the final landing page's HTML as well as that of all the sub frames on that page. The HTML features for image based spam and PDFs were not collected.

D. Pagelink features

These are links that appear on the final landing page. Only those links that automatically get loaded are considered. Embedded links such as HREFs are not considered. If a page contains a URL that is known to be spam, then, it would help in classifying the final landing page. An indication of spam is a page that comes loaded with innumerable URLs. In order to capture this, all the URLs on the final landing page are parsed. Every link is analyzed for redirects in the same way as mentioned before. Ratio of internal domains contacted versus external domains is also computed.

E. JavaScript Event features

JavaScript events such as pop up boxes, and the prompts that pop up before the user goes away from a page indicates the possibility of spam. Such pages that force the user to interact with them are strong indications of spam. So, all the messages that require a user action and onbeforeunload events are logged by the browser without reacting to it. In case a return value is expected, the browser provides a random value. The number of dialog boxes that are encountered, the text present in these dialogs and the presence of onbeforeunload events are all logged by the instrumented browser.

F. Pop-up window features

Pop-up windows also indicate the presence of spam. Upon the occurrence of pop up windows, the browser opens the window, and all the URL features that are collected for a new URL are collected and logged by the browser. The origin URL which created the pop up window is also logged along with the details of whether JavaScript or plug in that created the pop up window. The total number of pop up windows created as well as the features of each window is logged.

G. HTTP Headers

The HTTP headers that are created when a browser loads a landing page contain certain information such as the languages, versions of spam hosts, apart from cookie values and the other header fields. Those HTTP fields and values related to timestamps are ignored purposely so that the results are not biased.

These are the features present in the feature set used for the identification of spam URLs. These features are collected from the URLs given as input to the system. These features are then converted to feature vector in the feature extraction phase. The feature extraction phase is followed by the classification phase. It is in this phase that the URL is classified as benign or spam. In the next section, the classification method is discussed in detail.

VI. CLASSIFICATION OF URLS

The classification has to handle large data set and hence, we use linear classification method. In this classification method, decision is made on a linear combination of characteristics or features. The characteristics or features are presented to the classifier in what is called as a feature vector. If $\vec{x}$, the input feature vector to the classifier, is a real vector, then, y which is the output value is given by equation (1).

$$y = f(\vec{w} \cdot \vec{x}) = f\left(\sum_j w_j x_j\right), \quad (1)$$

Where $\vec{w}$ is the weight vector which is a linear mapping of $\vec{x}$ onto R. The linear classifier splits the high dimensional space with a hyper plane where one side of the hyper plane indicates one type of class say spam and the other side indicates the other class, say non-spam. A linear classifier is used especially when a great classification speed is required because this is the fastest classifier. When the number of dimensions $\vec{x}$ is very large, this classification is particularly suitable.

VII. EXPERIMENTAL SETUP AND EVALUATION RESULTS

The complete feature vector that is extracted is given to the linear classifier. A total of 41,890 URLs taken from about 405 Twitter profiles were fed into the system. The system classified the URLs as spam or non-spam based on the weights obtained for the feature vector collected. The accuracy obtained for the classifier is 99.3%. The False Positive rate was 0.7%. The False Negative Rate was 0.68%. Of the 19630 spam URLs, 19485 URLs were correctly identified as spam URLs. There were totally 22260 non-spam URLs. The actual spam and non-spam statistics are depicted in the Table 2.

TABLE 2. Spam URLs- Detected versus actual

	Detected	Actual
Number of spam URLs	19485	19630
Number of non-spam URLs	22108	22260

Fig. 3 depicts the results obtained for the linear classification graphically. The detection rate, false positive rate and false negative rate have been plotted for the linear classification method in this graph.

Detection rate provides details about how many of the spam URLs were classified correctly as spam and how many of the non-spam URLs were correctly classified as non-spam. The graph depicts only the detection rates for spam URLs. The detection rate for the non-spam URL is 99.3% again and it has not been plotted on the graph.

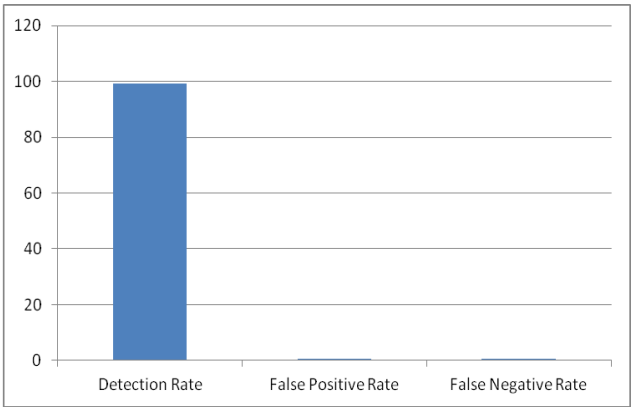

Fig. 3. Experimental results for classification

It can be clearly seen from the graph that the detection rate for this classification is very high which is at 99.3%. The false positive rate and the false negative rates are insignificant values. This proves that the classification method is a very effective one. Detection rates typically indicate whether the feature set is exhaustive and robust and whether the classification method is a suitable one for the detection mechanism. The result shows the robustness of the classifcation method as well as the features involved in the classification.

VIII. CONCLUSION AND FUTURE WORK

This work analyses the URLs from Twitter that is given as input and classifies the URLs as spam URLs or benign URLs. The linear classification provides a detection rate of 99.3% and a very minimal false positive rate of 0.7%. An exhaustive set of features has been considered for classification. This is to ensure that none of the spam URLs has been missed out. There are very few false positives and false negatives for this method. This shows the efficiency of the feature set used. We propose to extend this work to general web services and also to implement this system in real time scenarios where, as soon as a URL is encountered, the user can be instructed whether it is a harmless URL or a spam URL so that the user can take a run time decision of whether the URL can be clicked or not. Moreover, based on the classification results, we intend to generate blacklisted URLs which can be used for training the classifier to operate on live data. Also, we intend updating the blacklists continuously in the live system. This can be implemented using continuous learning scheme. This ensures that the black lists are up to date and any new spam URL will also be registered there.

REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter," in Proceedings of the Conference on Email and Anti-Spam (CEAS), 2010.

[2] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. Allyour contacts are belong to us: Automated identity theft attacks on social networks. In World Wide Web Conference, 2009.

[3] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Zhao, "Detecting and characterizing social spam campaigns," in Proceedings of the Internet Measurement Conference (IMC), 2010.

[4] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: the underground on 140 characters or less," in Proceedings of the ACM Conference on Computer and Communications Security (CCS), 2010.

[5] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots+ machine learning," in Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval, 2010.

[6] D. McGrath and M. Gupta, "Behind phishing: an examination of phisher modi operandi," in Proceedings of the 1st Usenix Workshop on Large- Scale Exploits and Emergent Threats, 2008.

[7] A. Ntoulas, M. Najork, M. Manasse, and D. Fetterly, "Detecting spam web pages through content analysis," in Proceedings of the 15th nternational Conference on World Wide Web, 2006.

[8] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting Spammers on Social Networks," in Proceedings of the Annual Computer Security Applications Conference (ACSAC), 2010.

[9] Thomas .K, Grier .C, Ma .J, Paxson .V, Song .D, Design and evaluation of a realtime url spam filtering service, in: IEEE Symposium on Security and Privacy, 2011.