Machine Learning Approach for Detection of Malicious Urls and Spam in Social Network

Miss. Roshani K. Chaudhari [1], Prof. D. M. Dakhane [2]

[1] Student, Department of CSE, Sipna COET, Amravati, Maharashtra, India

[2] Professor, Department of CSE, Sipna COET, Amravati, Maharashtra, India.

Abstract - *Twitter is prostrate to malicious tweets having URLs for spam distribution. Conventional Twitter spam detection methods take advantage of account features such as the ratio of tweets containing URLs and the date of creating an account, or relation features in the Twitter graph. These detection methods are ineffective against feature fabrications or consume much time and resources. In this paper we have proposed a machine learning system to find Malicious URLs and spam and to identify whether a given tweet is spam of not in a Social Network such as Twitter. By collecting dataset and training the classifier we classified the input tweet. The Naive Bayes algorithm, a supervised learning model with associated learning algorithms which are used to analyze data used for classification and regression analysis. After classification the sensitivity of each tweet is calculated. After experimental results it is found that the trained classifier is shown to be accurate and has low false positives and negatives.*

Key Words: Classification, Naïve Bayes, Stemming, Suspicious URL.

1. INTRODUCTION

Online Social Network such as Twitter allows its users to, among other things, micro-blog their day to day activity and talk about their interests by posting short messages called tweets which are consist of 140 characters. Twitter is extremely popular with more than 100 million active users who post about 200 million tweets every day [1]. As the dissemination of information is very easy on Twitter, makes it a popular way to spread external content like articles, images and videos by embedding URLs in tweets. However, these URLs may link to low quality content such as malware, spam websites or phishing websites. Malware, short for malicious software, is software used to disrupt computer operation, collect sensitive and important information, or gain access to private computer systems.

Phishing is the act to attempt for acquiring information such as usernames, passwords, and credit card details and sometimes, indirectly money by masquerading as a reliable entity in an electronic communication. Spam is flooding the Internet with different copies of the same message, in an endeavor to force the message on user or people who would not otherwise choose to receive it. Most of the spam is commercial advertising. Recent statistics show that on an average, 8% tweets consist of spam and other malicious content. Twitter also provides a shortening service. Social networking sites have become one of the important ways for users to keep trail and communicate with their friends online. Sites such as Face book, MySpace, and Twitter are frequently encompassed by the top 20 most-viewed web sites of the Internet.

In all current Online Social Networks (OSNs) the client-server architecture is adopted. The OSN service provider acts as the controlling entity. All the content in the system are stored and managed by it. OSN is using online spam filtering is installed at the OSN service provider side. Once installed, it inspect sever message before reading the message to the intended recipients and makes urgent decision on whether or not the message under analysis should be dropped. If the message is illegal mean instantly dropped the message otherwise it is forwarded to the corresponding receiver.

Different Twitter spam detection schemes have been proposed, to cope with malicious tweets. These schemes can be divided into account feature-based and relation feature-based schemes. Account feature-based schemes use the differentiating features of spam accounts such as the ratio of tweets containing URLs, the date of account creation, and the number of followers and friends. However, malicious users can easily contrive these account features. The relation feature-based schemes depend on more robust features that malicious users cannot easily assemble such as the distance and connectivity apparent in the Twitter graph. Deriving these relation features from the Twitter graph, however, requires an important amount of time and resources, because the Twitter graph is terrific in size. Many suspicious URL detection schemes [2] have also been introduced. They use static or dynamic crawlers and may be executed in virtual machine honeypots, like Capture-HPC [3], HoneyMonkey, and Wepawet, to examine newly observed URLs. These schemes divide URLs according to several features comprising DNS information, lexical features of URLs, URL redirection, and the HTML content of the landing pages. However, malicious servers can bypass investigation by selectively providing benign pages to crawlers.

In this machine learning approach, a detection of malicious URLs or spam in Twitter is done using the collected dataset, rather than exploring the landing pages of individual URLs in each tweet, which may not be successfully fetched, we deal with correlated redirect chains of URLs included in a number of tweets. Because attackers' resources are finite and need to be reused, a part of their redirect chains must be shared. We found a different number of meaningful features of suspicious URLs derived from the correlated URL redirect chains and related tweet context information. We assembled a Dataset which contains large number of Malicious URLs tweets from the Stanford University and trained a statistical classifier with their features. From results it is found that the trained classifier has high accuracy and low false-positive and false-negative rates.

2. LITERATURE REVIEW

In the recent times a lot of research work has been carried out for the design a better detection mechanism.

G. Stringhini, G. Vigna and C. Kruegel in 2010 [4] used account features such as Friend-Follower ratio, URL ratio and message similarity to differentiate spam tweets. This paper resolves to which extent spam has entered social network and how spammers who points social networking sites operate. To assemble the data about spamming activity, a large and disparate set of "honey-profiles" are established on three large social networking sites and then analyzed the collected data and identified peculiar behavior of users who influenced honey-profiles. Features are developed based on the analysis of this behavior which is used for detection. A. Wang in 2010 [5] modeled Twitter as directed graph where user accounts are represented by vertices and the type of relationship between users, friend or follower is actuated by the direction of edge. In this paper, detection mechanism is based on graph based features like in-degree and out-degree of nodes and content based features like presence of Trending topics and HTTP links in tweets. This work applies machine learning methods to automatically discriminate spam accounts from normal ones. Based on the API methods provided by Twitter to excerpt public available data on Twitter website, a Web crawler is developed. Finally, a system is established to assess the detection method. J. Song, S. Lee, and J. Kim in 2011[6] viewed Twitter as an undirected graph and made use of Menger's theorem to evaluate the values of message features such as distance and connectivity between nodes in order to achieve detection. The relation features prototype system such as distance and connectivity are exclusive features of social networks and are difficult for spammers to forge or manipulate. This system analyses spammers in real-time, this implicates that when a message is being delivered, clients can classify the messages as spam or benign. C. Yang, R. Harkreader, and G. Gu (2011) [7] in their research used time based aspects such as tweet rate and following rate besides graph based aspects and content based aspects in order to perform detection. H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary [8] suggested a detection system based on message features such as interaction history between users, average number of tweets containing URL, average tweet rate, and unique URL number. In OSNs, multiple users are connecting and interacting via the message posting and viewing interface. The system analyses every message and calculates the feature values before rendering the message to the intended recipients and makes immediate determination on whether or not the message under investigation are dropped.

Some preceding works are based on URL detection schemes. Ma, L. K. Saul, S. Savage, and G. M. Voelker in 2009 [9] recommended a system which detect malicious websites by verifying lexical features and host based features of URL. This application is precisely applicable for online algorithms as the size of the training data is bigger than can be effectively processed in batch and because the distribution of features. Prior works relied on batch learning algorithms. But online techniques are far better for two reasons: (1) Online techniques can process huge numbers of examples far more efficiently than batch techniques. (2) Changes in malicious URLs and their features over time can simply be adapted. D. Canali, M. Cova, G. Vigna, and C. Kruegel in 2011 found that HTML aspects, JavaScript aspects and URL based aspects can be used for efficient detection of malicious websites [10].

H. Kwak, C. Lee, H. Park, and S. Moon proposed a work in 2010 [1]which mainly focuses on Twitter, a social networking service, more than 41 million users as of July 2009 and is growing rapidly. Twitter users tweet about any topic within the 140-character limit. Twitter offers an Application Programming Interface (API) which is easy to crawl and collect data. Most often mentioned words, phrases and hash tags are tracked by Twitter and posted them under the title of "trending topics" repeatedly. A hash tag is a representation through Twitter users for creating and following a thread of consideration by prefixing a word with a '#' character. In order to describe influential on Twitter.

Juan Chen and chuanxiongguo described online disclosure of phishing attacks and prevention of phishing attacks [11]. Prabhu, Dhanalakshmi and Chellapan had depicted how to identify the phishing websites and to assure secure transaction [12].

3. PROPOSED WORK

Proposed work is done for identifying malicious links and spam by using Naïve Bayes Algorithm. Proposed framework works in two stages as shown in Figure 3:

Stage 1: Training dataset,

Stage 2: Testing input tweet.

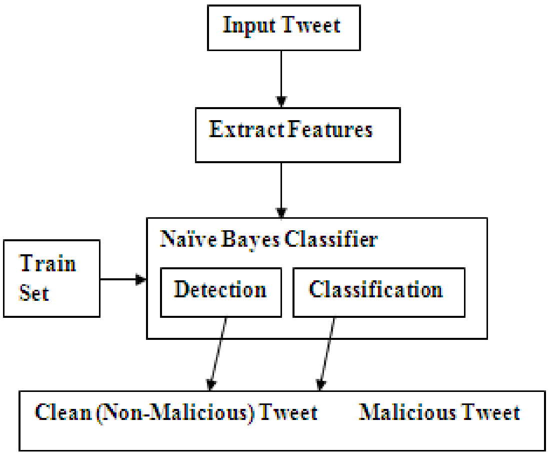

Fig. 3. The Framework of proposed method

These stages can operate consecutively as in batched learning, or in an interleaving manner: additional data is collected to incrementally train the classification model while the model is used in detection and identification.

3.1 Proposed Modules

Proposed work is executed in two main modules which are explained below.

3.1.1 Training Data Set

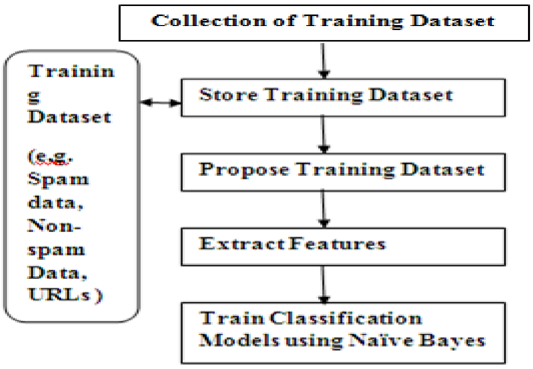
Collection of Training Dataset
Trainin g Dataset (e.g. Spam data, Non-spam Data, URLs)
Store Training Dataset
Propose Training Dataset
Extract Features
Train Classification Models using Naïve Bayes

Fig. 3.1.1. Training Dataset

The training modules includes following steps:

1) Data Gathering
2) Preprocessing
3) Feature Extraction
4) TF-score

In data gathering the standard dataset for spam, non-spam and URLs (malicious and non-malicious) is collected from Stanford University site. This collected data get preprocessed using stemming and stop word removal. After that the feature gets extracted from the preprocessed data. In feature extraction the token, keyword, and link get separated. Here the features for spam dataset as well as non-spam dataset are separately calculated. The Malicious URLs, esp. those for phishing attacks, usually have distinguishable patterns in their URL. Among these lexical options, the typical domain/path token length (delimited by '.', '/', '?', '=', '-', '') and that phishing URLs show completely different lexical patterns. After feature extraction, the term frequencies for each word and urls get calculated and maintained for further purpose.

3.1.2 Testing Module

In this Module as shown in Figure 3.1.2 an unknown tweet, is given to a system as an input. This input tweet gets preprocessed using stop word removal and stemming. Again Unknown input tweet which may contains URLs and spam related words given for testing is submitted to Extract Features associated with URL, and maps these features with extracted features from known train set. Mapping is based on Classification Model (Naïve Bayes) is applied to detect a Malicious URL and spam related words. After detecting and classifying the tweet, the sensitivity is calculated.

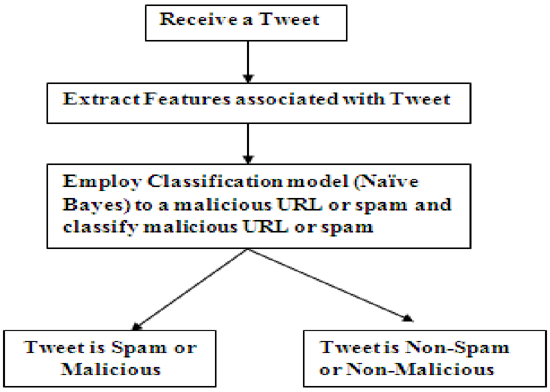
Receive a Tweet
Extract Features associated with Tweet
Employ Classification model (Naïve Bayes) to a malicious URL or spam and classify malicious URL or spam
Tweet is Spam or Malicious
Tweet is Non-Spam or Non-Malicious

Fig. 3.1.2. Testing Module

3.2 Mathematical Model

Naive Bayes Rule is the basis for many machine-learning and data mining methods. The rule (algorithm) is used to create models with predictive capabilities. It provides new ways of exploring and understanding data. It is used when data is high and we want efficient output compared to other methods.

The probability model for a classifier is a conditional model over a dependent class variable C.

$$p(C \mid F1,\ldots,Fn)$$

Using Bayes' theorem,

$$p(C \mid F1,\ldots,Fn)=(p(C)p(F1,..Fn/C))/(p(F1,..Fn))$$

•p(C |F1,…,Fn)= probability of instance F1,…,Fn being in class C.

•p(F1,..Fn/C) = probability of generating instance F1,…,Fn by given class C, One can imagine that being in class C, causes to have feature F1,…,Fn with some probability.

•p(C) = probability of occurrence of class C,

•p(F1,…,Fn) = probability of instance F1,…,Fn occurring.

In simple words the above equation can be written as

Posterior=(Prior*Likelihood)/Evidence

The denominator is independent of C and the values of the features Fi given, so that the denominator is effectively constant. The numerator is equivalent to the joint probability model p(C,F1..Fn)

Naïve Bayes is an classification approach mostly used for detection and categorization of text documents. By providing a set of classified training samples, an application can learn from these examples, so as to predict the class of unknown URL. With a small number of outcomes or classes, conditional on several feature variables F1 through Fn. The features (F1, F2, F3, F4) which are present in URL are independent from each other. Every feature Fi (1<=i<=4) text binary value showing whether the particular property comes in URL. The probability is calculated that the given URL belongs to a class m (m1: Non-spam and m2: Spam) as follows:

P (m1/F) = (P (m1)*P (F/mi))/P (F)

Where all of P(F) are constant meanwhile P (Fi|m1) and P(mi) can be easily calculated from training. The proportional to P (m1|F), P(m2|F) is calculated and the results are as follows:

P(m1|F)P(m2|F) > b (b>1), Benign link or non-malicious.

P(m2|F)P(m1|F) > b , Malicious link.

3.3 Sensitivity of Tweet

After classification of tweet using Naïve Bayes Classifier, the sensitivity of tweet is calculated. The sensitivity can be calculated using total numbers of spam words or malicious word found in input tweet and total number of preprocessed words.

4. RESULT AND ANALYSIS

In this implemented work, the tweets from user are specially taken as input. On these tweets various operations are applied. This system used in this project is evaluated and tested by taking different input tweets.

Again for evaluating different factors such as Precision, Recall, F-measure and Accuracy, some input tweets has taken and classified using the implemented system. The statistical measures are considered (TN, FP, TP and FN). It is found that the value of TN=4, TP=6, FP=2 and FN=1 which is shown in following graph.

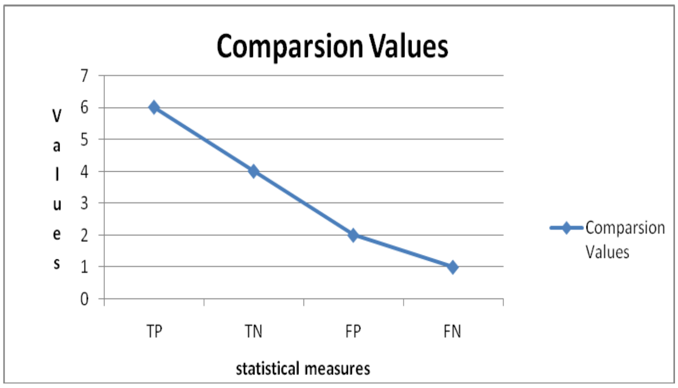

Chart-4.1 Comparisons values

Hence from these measures, the values for Precision, Recall, F-measure and Accuracy get calculated.

The calculated values are shown in graph

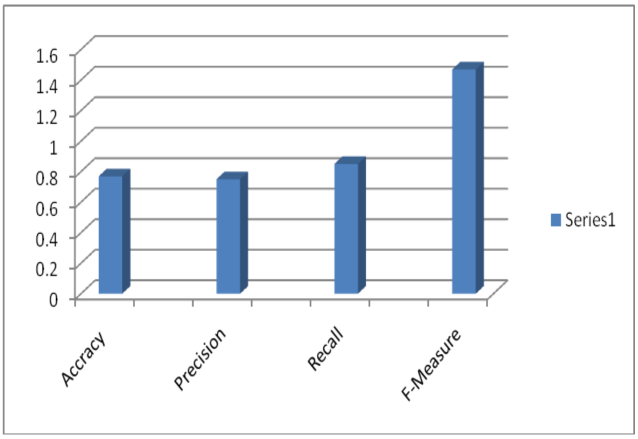

Chart. 4.2 Evaluation of Accuracy, Recall, Precision and F-Measure

It is observed that the values for Precision, Recall, F-measure and Accuracy are 0.75, 0.86, 1.47 and 0.77 respectively. Hence it found that the accuracy of implemented system is more.

5. CONCLUSIONS

In this paper, we have suggested a machine learning approach for the detection of malicious urls and spam. The technique used in this system is a naïve bayes classifier used to classify the input tweet whether it is malicious (spam) or not. The naïve Bayes classifier classifies the tweet on the basis of posterior probabilities of tweet. After classifying the tweet, the sensitivity is calculated. After calculating all the results it is found that the trained classifier is shown to be accurate and has low false positives and negatives. Also the sensitivity of each tweet is calculated successfully.

ACKNOWLEDGEMENT

I would like to thank to the almighty for giving me the knowledge and courage to complete the research work on Twitter successfully. Moreover I express my gratitude to my respected Professor Prof. D. M. Dakhane, for following me to do research work deliberately. Finally, I wish to express my gratitude to my family and my friends, for providing endless support in every aspect possible, to ease my path on this journey.

REFERENCES

[1] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In Int. World Wide WebConf. (WWW), 2010.

[2] THOMAS, K., GRIER, C., MA, J., PAXSON, V., AND SONG, D. Design and Evaluation of a Real-Time URL Spam Filtering Service. In *Proceedings of the IEEE Symposium on Security and Privacy* (May 2011).

[3] ANDERSON, D. S., FLEIZACH, C., SAVAGE, S., AND VOELKER, G. M. Spamscatter: characterizing internet scam hosting infrastructure. In *Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium* (Berkeley, CA, USA, 2007), USENIX Association, pp. 10:1–10:14.

[4] G. . Stringhini, C. Kruegel, and G. Vigna, "Detecting Spammers on Social Networks,"*Proc. 26th Ann. Computer Security Applications Conf. (ACSAC), 2010.*

[5] A. Wang, "Don't Follow Me: Spam Detecting in Twitter," Proc. Int'l Conf. Securityand Cryptography (SECRYPT), 2010.

[6] J. Song, S. Lee, and J. Kim, "Spam Filtering in Twitter Using Sender-ReceiverRelationship,"Proc.14th Int'l Symp. Recent Advances in Intrusion Detection.(RAID), 2011.

[7] C. Yang, R. Harkreader, and G. Gu, "Die Free or Live Hard? Empirical Evaluationand New Design for Fighting Evolving Twitter Spammers," Proc.14th Int'l Symp.Recent Advances in Intrusion Detection (RAID), 2011.

[8] H. Gao, Y. Chen, K. Lee, D. Palsetia, and A. Choudhary, "Towards Online SpamFiltering in Social Networks," Proc. 19th Network and Distributed System SecuritySymp. (NDSS), 2012.

[9] J. Ma, L.K. Saul, S. Savage, and G.M. Voelker, "Identifying Suspicious URLs: AnApplication of Large-Scale Online Learning,"Proc. 26th Int'l Conf. Machine Learning(ICML), 2009.

[10] D. Canali, M. Cova, G. Vigna, and C. Kruegel,"Prophiler: A Fast Filter for theLarge-Scale Detection of Malicious Web Pages,*" Proc. 20th Int'l World Wide Web Conf. (WWW), 2011.*

[11] Ollman, G.(2004) The phishing Guie-Understanding and Preventing , White paper , Next Generation Security software Ltd.

[12] Neil Chou, Robert Ledesma, Yuka Teraguchi, D anBoneh, and John C.Mitchell. Client-side defense against web-based identity theft.Proc. NDSS 2004,2004.