DESIGN AND IMPLEMENTATION OF AUTOMATED SPAM DETECTION APPROACH IN TWITTER

[1]SHAIK NASREEN, [3]BABU PANDIPATI

[1]PG Scholar, Dept of CN, Quba College of Engineering and Technology, Nellore, AP, India.

[2]HOD, Dept of CSE, Quba College of Engineering and Technology, Nellore, AP, India.

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 users 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.

Index terms – Online Social Networking Sites (OSNs), Spammers, Twitter, Legitimate users.

I. 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.

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.

II. BACKGROUND 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 honey pots 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% no spammers and again with 10% non-spammers+90% spammers. Limitation of the approach is that less dataset has been used for validation. Similarly Benevenu to et al. [3] detected spammers on the basis of tweet content and user based features. Tweet content attributes used are – number of hash tags 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 rewetted.

Fraction of tweets containing URLs, fraction of tweets that contains spam words, and average number of words that are hash tags 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.

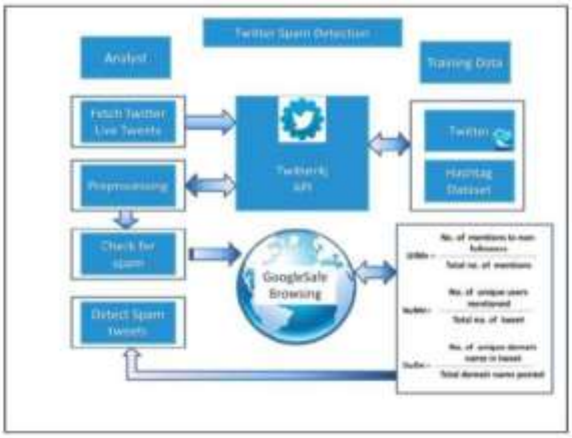


Fig. 1: System Architecture

III. PROPOSED WORK

Spammers often have similar behavioral patterns that can be easily detected. Spam feature analysis and useful features have been separated and described in [18] and [19]: Number of reviews per day: Number of reviews written in a single day by a user is displayed. Spammer Most spammers (75%) write comments more than 5 chapters per day, while 90% of non-spammers write comments no less than 3 times a day and 50% write reviews per day. Positive Positive Percentage: Positive comments mean reviews with a 4 or 5 star rating. Analyzing the information from the bad guys, spammers show that the percentage of positive reviews has been scattered. Consistent among users While about 85% of spammers have a positive opinion of 80% or more. Check length: Since spammers are paid according to the number of spam posted, they often write reviews. Short to maximize profits. The average length of 92% of users is over 200, while only 20% of spammers submit more than 135 reviews. Deviation of reviewers: Consider spammers often rated high or low. Their ratings are different from the average rating. [18] The author has calculated the absolute rating deviation from the review of other reviews of the same product. In the review, about 70% of non-spammers had a discrepancy of less than 0.6, while 80% of spammers had discrepancies. More than 2.5.

Early scoring deviation: When a product is published, the seller tries to promote the item from the beginning to earn attention. For this reason, spammers are the most likely to be eligible after the product has been published, with the average score calculated for the product, and two features: a review of the rating and weight of the rating, which states that the range Re search in the year 19 shows that it is possible. Use these features to check spam comments.

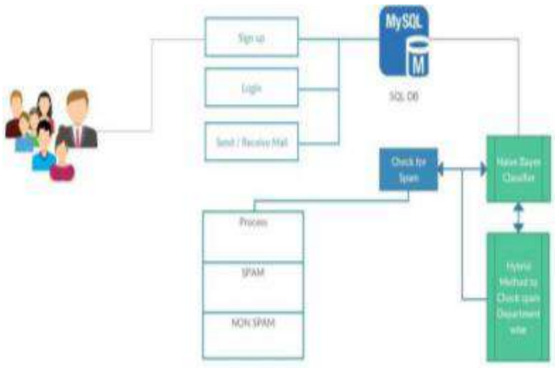


Fig. 2: Proposed System

IV. ALGORITHM AND TWEET ANALYSIS

HashTagging data set

To create the hash tagged data set, we first filter out duplicate tweets, non-English tweets, and tweets that do not contain #hashTags. From the remaining set (about 4 million), we investigate the distribution of #hashTags and identify what we hope will be the sets of frequent #hashTags that are indicative of positive, negative and neutral messages. These #hashTags are used to select the tweets that will be used for development and training

Pre-Processing

The first pre-processing technique is remove @ which means it scans the whole document of input dataset and after comparing it with @ it deletes @ from every available comment with @.The next step of pre-processing is remove URL where the whole input document gets scanned and compared with http:\\... and the comments having URL are deleted.

Further we move on to stop word removal being the next step in data pre-processing. Stop word removal exactly means that from the whole statement after scanning it removes the words like and, is, the, etc and only keeps noun and adjective. Tokenization and Normalization are carried out thereafter. Porter Stemmer Algorithm is used thereafter. The Porter stemming algorithm (or „Porter stemmer‟) is a process for removing the commoner morphological and in flexional endings from words in English. Its main use is as part of a term normalization process that is usually done when setting up Information Retrieval systems In the implementation part of Intelligent Twitter Spam Detection using a Hybrid Approach, we used Twitter 4J API, Google Safe Browsing Toolkit, A combination of classifiers including NB and SVM and unique feature sets that provide an intelligent spam detection solution. The system consists of 6 tabs which were designed in Net Beans using Java Swing for the front end part. The Twitter Spammer tab consists of buttons that Fetch Live Tweets and Recent Twitter Feed related to the trending hashtags. The tweets are then stored on a local dataset which are then loaded for further analysis. Thereafter pre-processing takes place and the tweets are sent to the next stage for implementation of classification algorithms using the Hybrid Approach. A decision is made based on the code that runs on the filtered tweets using the aforementioned feature vectors and the tweets are classified as SPAM. Furthermore there is a provision for showing Trends-Wise Analysis for a particular hashtag in the next stage after which the final stage shows the suspended twitter accounts which were labelled as SPAM by the system. The analysis for this research was conducted by mining 10,782 tweets comprising of 72 hashtags that were trending on twitter. The system was able to classify 2,466 tweets as SPAM while the others were found legitimate. After cross-checking with Google Safe Browsing it was found that 2,153 tweets did contain a malicious URL that re-directed the user to suspicious websites. This affirmed that the accuracy of the said system stands at 87.30% with this multi-tier approach which is greater than systems which use a single-classifier and non-hybrid approaches.

V. RESULT ANALYSIS

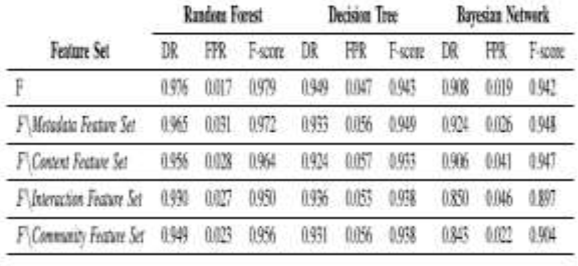
Following a detailed description of the features and feature extraction process in preceding sections, this section presents the experimental details and evaluation results of the proposed approach for detecting automated spammers in Twitter.

As stated in the beginning of this section, the performance of the proposed approach is evaluated using three classifiers, namely, random forest, decision tree, and Bayesian network, which are implemented in Weka.6 We have used ten-fold cross validation to ensure the participation of each instance in both training as well as testing procedure. The performance of the classifiers is evaluated using standard evaluation metrics, namely, DR, FPR, and F-Score.

The first row of this table presents the evaluation results of the classifiers considering all 19 features (F). It can be observed from the first row of this table that random forest performs best in terms of all three metrics DR, FPR, and F-Score. However decision tree is also good in terms of DR and F-Score with the values of 94.9% and 94.3% respectively. Bayesian network performs significantly good in terms of FPR and F-Score, but not as good in terms of DR.

To evaluate the discriminative power of features categories, we perform feature ablation test in which some features are removed from the feature set to observe their impact on classifiers' performance. Accordingly, the experiment mentioned above is repeated four times, excluding the features of a particular category in each repetition, using the set difference function, where F is the set of all 19 features and F1 is the set of features of a particular category. The second, third, fourth, and fifth rows of Table I present the evaluation results corresponding to the exclusion of feature categories. As presented in the table, overall, interaction-based features are efficient in terms of DR and F-Score. This feature category includes three new features and all the three are based on followers of user, which is one of the novelties of the proposed approach, which reflects the importance of followers for detecting spammers. Content based features also show moderate discriminating power for decision tree, although not good for the other two classifiers, that endorse the fact that bots still use content to trap users by using enticing contents in their posts and it does not depend on their sophistication level. As observed from the table that community-based features also show good discriminating power, and affect the classifiers efficiency. In addition, community-based features are the most discriminating features in terms of DR for Bayesian network. Metadata features show least impact on performance of the classifiers, which highlights the efficacy of random number generator algorithms, used by bots to achieve randomness in their behavior similar to those of human-beings.

Table 1. Performance Evaluation Of Classifiers Over The Dataset



	Random Forest			Decision Tree			Bayesian Network		
Feature Set	DR	FPR	F-score	DR	FPR	F-score	DR	FPR	F-score
F	0.976	0.017	0.979	0.949	0.047	0.943	0.908	0.019	0.942
F\Metadata Feature Set	0.965	0.031	0.972	0.933	0.056	0.949	0.924	0.026	0.948
F\Content Feature Set	0.956	0.028	0.964	0.924	0.057	0.933	0.906	0.041	0.947
F\Interaction Feature Set	0.930	0.027	0.950	0.936	0.053	0.938	0.850	0.046	0.897
F\Community Feature Set	0.949	0.023	0.956	0.931	0.056	0.938	0.845	0.022	0.904

VI. CONCLUSION

In this paper, we have proposed a hybrid approach exploiting community-based features with metadata-, content-, and interaction-based features for detecting automated spammers in Twitter. Spammers are generally planted in OSNs for varied purposes, but absence of real-life identity hinders them to join the trust network of benign users. Therefore, spammers randomly follow a number of users, but rarely followed back by them, which results in low edge density among their followers and followings. This type of spammer's interaction pattern can be exploited for the development of effective spammer's detection systems. Unlike existing approaches of characterizing spammers based on their own profiles, the novelty of the proposed approach lies in the characterization of a spammer based on its neighboring nodes (especially, the followers) and their interaction network.

REFERENCES

[1] Sharma K. and Jatana N. (2014)“Bayesian Spam Classification: Time Efficient Radix Encoded Fragmented Database Approach” IEEE 2014 pp. 939-942.

[2] Sharma A. and Anchal (2014), "SMS Spam Detection Using Neural Network Classifier",ISSN: 2277 128X Volume 4, Issue 6, June 2014, pp. 240-244.

[3] Ali M. et al (2014), , "Multiple Classifications for Detecting Spam Tweets by Novel Consultation Algorithm", CCECE 2014, IEEE 2014, pp. 1-5.

[4] Liu B. et al (2013) “Scalable Sentiment Classification for Big Data Analysis Using Na¨ıve Bayes Classifier” IEEE 2013 pp.99-104.

[5] Belkebir R. and Guessoum A. (2013), "A Hybrid BSO-Chi2-SVM Approach to Arabic Text Categorization", IEEE 2013, pp. 978-984.

[6] Blasch E. et al (2013), Kohler, “Information fusion in a cloud-enabled environment,” High Performance Semantic Cloud Auditing, Springer Publishing.

[7] Allias N. (2013) “A Hybrid Gini PSO-SVM Feature Selection: An Empirical Study of Population Sizes on Different Classifier” pp 107-110