A MIXED TECHNIQUE TO DETECT AUTOMATED SPAM SENDERS IN A NETWORK

[1]Darla Vandana, [2]V. Sesha Bhargavi

[1]M.Tech Student, [2]Assistant Professor,

[1&2] Department of Information Technology,

[1&2] G. Narayanamma Institute of Technology and Science, Hyderabad, India.

Abstract : Twitter, a popular micro-blogging service, is traditionally used to share messages and updates with a maximum of 280 characters. It is very open in nature and has a large user base which is often exploited by spammers to commit cybercrimes, like phishing, cyber bullying, harassment and spreading rumors. The proposed approach is to discriminate users depending on their activities with their corresponding *followers* where the user can bypass the features related to his/her activities, but avoiding those which are based on the *followers* is challenging. In this paper, a technique to detect automated spammers by combining metadata based features, content based features and interaction based features is proposed. Nine different features are identified for learning the dataset that includes both the legitimate users and spammers. The distinction between the feature categories is analyzed; interaction-based and content-based features are decided to be more effective for the detection of spammers, while metadata-based features are less effective.

IndexTerms **- Social Network, Automated Spammers, Social Network Security, Spambot Detection.**

I. INTRODUCTION

Twitter, a popular blogging service, is used to share messages and news with a maximum of 280 characters. It is very open in nature and Twitter has a user base up to a great extent which is certainly a benefit for the spammers to commit cybercrimes. Approaches were proposed by scientists to resolve these issues, which depend on characterization of the user and interactions between the users.

In this paper, a technique for the identification of automated spammers by combining *content-based* features, *metadata-based* features and *interaction-based* features is proposed. The approach of the proposed system lies in the depiction of users, which are based on their interactions with their *followers*. The distinction between the feature categories is analyzed; interaction-based and content-based features are decided to be more effective for the detection of spammers, while metadata-based features are less effective.

A. Need for the Study

Many researchers from industry as well as academia are continuously working to reduce the number of cyber-criminals to make the usage of Online Social Network (OSN) a pleasant and delightful experience. As a result, a plenty of spam detection methods were proposed. However, as methods are mature and innovative, spammers use more intricate mechanisms to avoid detection.

The current spam and other malicious behavior detection strategies utilize either feature-based or graph partitioning based strategies. Feature-based strategy includes features, such as number of followers, number of tweets which are easy to bypass, while few advanced features are challenging to bypass. However, features are generally based on user activities and spammers can adjust their behavior to imitate those of benign users.

Although, these methods are traditional detection approaches, spammers can try to evade them by creating adequate attack links between malevolent users and legitimate users. Hence, a technique is proposed which is a fusion of *metadata-based* features with *interaction-based* and *content-based* features. Legitimate users usually follow and react to requests from known users and avoid interactions with strangers. On the other hand, spammers follow random users, that forms very limited connections amidst followers and affect interaction-based features.

B. Objectives of the Study

A mixed technique for the detection of automated spammers in Twitter utilizes a combination of *metadata-based, content-based, interaction*-based features is proposed. In the evaluation of depicting the features of current approaches, most of the network-based features are undefined using user followers, as a result, ignoring the fact that the reputation of a user in a network is transmitted from the followers (instead of the ones user is following).

The features are classified into three different categories, which are, *metadata-based*, *content-based*, and *interaction-based* features. Metadata-based features are derived from already existing extra information regarding the tweets, while content-based

features study the tweet posting behavior of a user and quality of the text used in posts and interaction-based features are derived from user interactions in a network.

II. Related Works

Spammers have been a problem from the very beginning of the internet. Through time, detection of spammers became very complex. We know that Twitter consists of many malicious tweets with URLs to perform various cyber crimes. In [1], a hybrid approach for detecting spammers in Twitter was proposed which uses a dataset containing 11000 labeled users, including 10000 benign users and 1000 spammers. It uses 19 different features, including 6 new and 2 redefined features. Our work is an extension to [1] which identifies the automated spammers and also blocks the spammer account from future signing in to the Social Network. In [2], Sangho Lee and Jong Kim, proposed WarningBird, a system which detects suspicious URLs and investigates the correlations of URL redirect chains which are extracted from various tweets.In [3], the tactics used for evasion, utilized by the spammers are classified into two types: profile-based and content-based feature evasion tactics by Chao Yang, Robert Harkreader, and Guofei Gu. In [4], Twitter spams are detected with the datasets which are collected by using various Machine Learning algorithms like KNN, k-kNN, Random Forest, c5.0, Stochastic GBM and Naïve Bayes. Two features are proposed namely: Account-based features and Tweet content-based features. In [5], a review of spam detection methods is discussed. The Twitter spam detection features can be classified as follows: Account-based, Tweet-based features and the relationship between the sender and the receiver of tweets. In [6], the features used for deciding whether a tweet is a spam or not are as follows: Number of Unique Hash-tags (#), Number of Unique URLs and Number of Unique mentions (@). The different stages of spam detection in [5] are Collection of live tweets, Pre-processing, Feature extraction and Classification. In [7], Spam Detection on Twitter is done by using Use-based and Content-based features. Random Forest Classifier gave the best results among SMO, Naive Bayes and K-NN neighbor. In [8], many features such as user profile features, user-activity features, content features and location-based features are proposed. After extracting the features, clusters are formed which group similar trending topics of a tweet user profile. In [9], the authors proposed an integrated approach for the classification of spammers using URL analysis, Machine Learning techniques and Natural Language Processing (NLP). In [10], Spam detection is proposed by extracting two features namely: User-based and Content-based features. User-based features are further classified into number of friends, number of followers and reputation of the user. Content-based features are further classified into number of URLs, replies/mentions, keywords/wordweight, retweets and hash-tags.

III. Proposed Approach

In the proposed system, a mixed technique for the detection of automated spammers in Twitter, which utilizes a combination of various features, is designed.

The proposed features are categorized into three categories namely- metadata, content and interaction. Metadata-based features are derived from already existing extra information regarding the tweets. Content-based features study the tweet posting performance of a user and quality of the text used in posts. Interaction-based features are derived from user interactions in a network. A summary of all categories including the features in each category is shown in the below table.

Table 1: Nine features with their categories

Category	Feature
Metadata	1) Retweet Ratio
	2) Tweet Time Interval Standard Deviation
	3) Tweet Time Standard Deviation
Content	1) URL Ratio
	2) Unique URL Ratio
	3) Mention Ratio
	4) Hashtag Ratio
Interaction	1) Follower Ratio
	2) Reputation

A. Metadata-based features

Many researchers from industry as well as academia are continuously working to reduce the number of cyber-criminals to make the usage of Online Social Network (OSN) a pleasant and delightful experience. As a result, a plenty of spam detection methods were proposed. However, as methods are mature and innovative, spammers use more intricate mechanisms to avoid detection.

1. Retweet Ratio (RR)

$$RR(u) = \frac{RT(u)}{N(u)} \quad (1)$$

where RT(u) represent the total number of tweets retweeted by user u and N(u) is the total number of tweets tweeted by the user u.
Retweet Ratio (RR) is high for automated spammers and low for legitimate users.

2. Tweet time Interval Standard Deviation (TISD)

$$TISD(u) = \frac{\sum_{i=1}^{n}(T_i - \bar{T})^2}{N(u)} \quad (2)$$

where $T_1, T_2, \ldots T_n$ is the time interval between the consecutive tweets, $\bar{T}$ is the mean of time interval and N(u) is the total number of tweets tweeted by the user u.
Tweet Time Interval Standard Deviation (TISD) is low for automated spammers and it is high for legitimate users.

3. Tweet Time Standard Deviation (TSD)

$$TSD(u) = \frac{\sum_{i=1}^{N(u)}(t_i - \bar{t})^2}{N(u)} \quad (3)$$

where t_i represents the tweet time of i^{th} tweet, $\bar{t}$ is the mean tweet time $\bar{T}$ and N(u) is the total number of tweets tweeted by the user u.
Tweet Time Standard Deviation (TSD) is low for automated spammers and it is high for legitimate users.

B. Content-based features

In the existing methods of detection of spammers, content quality is considered as the most important. Spammers usually post attractive tweets to deceive users. In the proposed method, four content-based features are identified and are defined below.

1. URL Ratio (UR)

$$UR(u) = \frac{U(u)}{N(u)} \quad (4)$$

where U(u) represents the total number of URLs used in the tweets of user u and N(u) is the total number of tweets tweeted by the user u.
URL Ratio (UR) is nearly 1 for spammers and low (nearly 0) for legitimate users.

2. Unique URL Ratio (UUR)

$$UUR(u) = \frac{UU(u)}{U(u)} \quad (5)$$

where UU(u) represents the total number of unique URLs in the tweets of user u and U(u) is the total number of URLs used in the tweets of u.
Unique URL Ratio (UUR) is low for automated spammers and it is high for legitimate users.

3. Mention Ratio (MR)

$$MR(u) = \frac{M(u)}{N(u)} \quad (6)$$

where M(u) is the overall number of mentions in the tweets and N(u) is the total number of tweets tweeted by the user u.
Mention Ratio (MR) is high for automated spammers and low for legitimate users.

4. Hashtag Ratio (HTR)

$$HTR(u) = \frac{HT(u)}{N(u)} \quad (7)$$

where HT(u) is the total number of hashtags used in tweets of user u and N(u) is the number of tweets tweeted by the user u.
Hashtag Ratio (HTR) is high for automated spammers and low for legitimate users.

C. Interaction-based features

The interaction data available in Twitter can be used in spammer detection. In Twitter, there is an option to follow other users through which a user can interact with other users and create a trusted network among the users. Two features are identified under this category and are defined below.

1. Follower Ratio (FR)

$$FR(u) = \frac{|\bar{u}|}{|\vec{u} \cup \bar{u}|} \quad (8)$$

where $\vec{u}$ is the group of followings of the user u and $\bar{u}$ is the group of followers of the user u.
Follower Ratio (FR) is low for automated spammers and high for legitimate users.

2. Reputation (R)

$$R(u) = \frac{|\bar{u} \cap \vec{u}|}{|\vec{u}|} \quad (9)$$

where $\vec{u}$ is the group of followings of the user u and $\bar{u}$ is the group of followers of the user u.
Reputation (R) is low for automated spammers and high for legitimate users.

IV. System Architecture

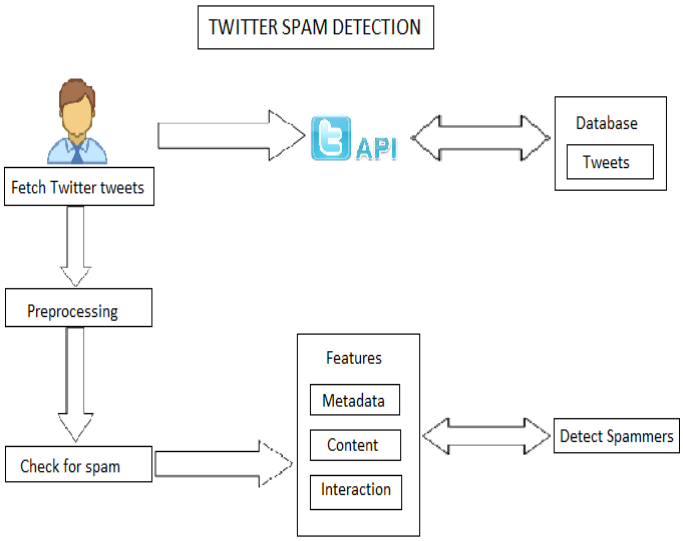

Fig.1: System Architecture

In the proposed method, the system fetches all the tweets posted by the users and then preprocesses them and searches for automated spammers. The different features through which an admin can identify an automated spammer are categorized into three categories i.e. Metadata-based features, Content-based features and Interaction-based features which are again classified into 9 different features as indicated in Table 1. The system compares each tweet posted by the users against the proposed features and detects automated spammers in any social network.

V. Applications Of Proposed System

In the proposed technique, an approach on exploiting metadata-based features along with content-based and interaction-based features to detect automated spammers was developed. The approach of the proposed system lies in the depiction of users, which depend on their interactions with their *followers.*

By using this proposed system and with the help of the proposed 9 features, the automated spammers can be easily detected which help to reduce the cyber-crimes in Online Social Networks.

VI. Enhancement

Many approaches were made to detect the automated spammers in Social Networks but the number of cyber crimes due to spammer is also increasing consistently.[1] is one of the approach to detect spammers through 19 different features. This paper is an enhancement to [1] in which automated spammers are detected through 9 different features which are classified into three categories namely Metadata, Content and Interaction. Our approach not only detects and identifies spammers but the admin also has an option to block the spammer account. When a particular user is identified as a spammer, the admin can block the user from signing in into the Social Network. When the corresponding user tries to log in into his/her account, they will not be able to sign in into their accounts. In this way, if the spammers are blocked from signing in into their accounts, there will be a decrease in the number of cyber crimes in Social Networks.

VII. Results and Discussion

Twitter, a most popular micro-blogging platform, is viewed an Online Social Network (OSN) which has a wide range of users. Twitter allows its users to follow their favorite actors, athletes, celebrities, political leaders, and news channels, and also to view to their content without interventions. By *following* activity, a user can view updates of the corresponding account. Though most of the social networking sites are used for several legitimate purposes, their open-nature and a large user base have made them targets for cyber criminals.

Hence, a method is proposed, in which 9 features are grouped into three categories which are helpful to detect automated spammers in a network. Interaction and Content based features are more effective for the detection of spammers, while metadata-based features are less effective. By using the proposed method, automated spammers can be easily detected in an Online Social Network.

The following are some of the output screenshots of the proposed method.

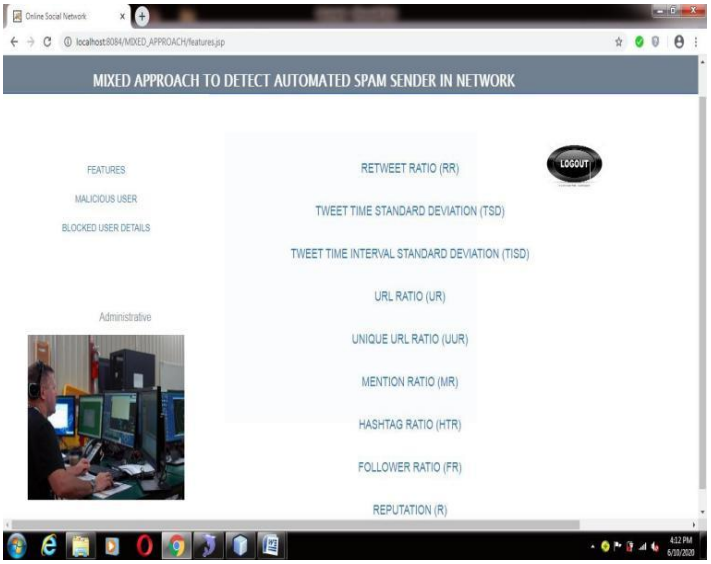

Fig. 2: Features

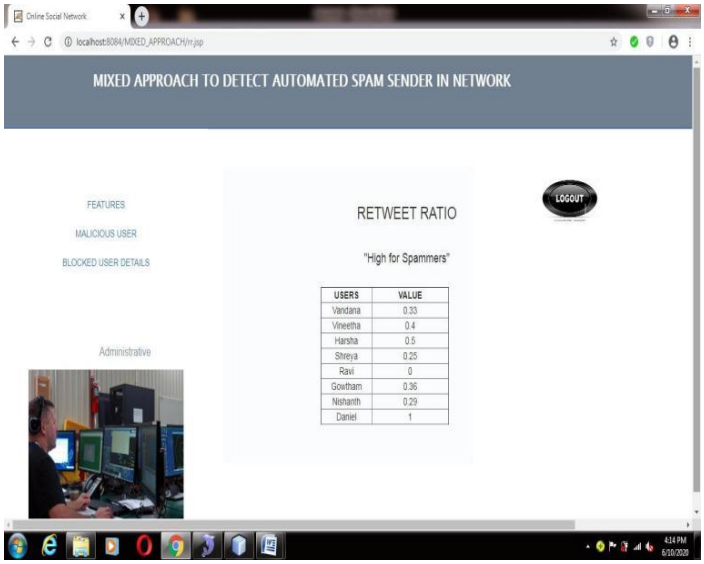

Fig. 3: Retweet Ratio

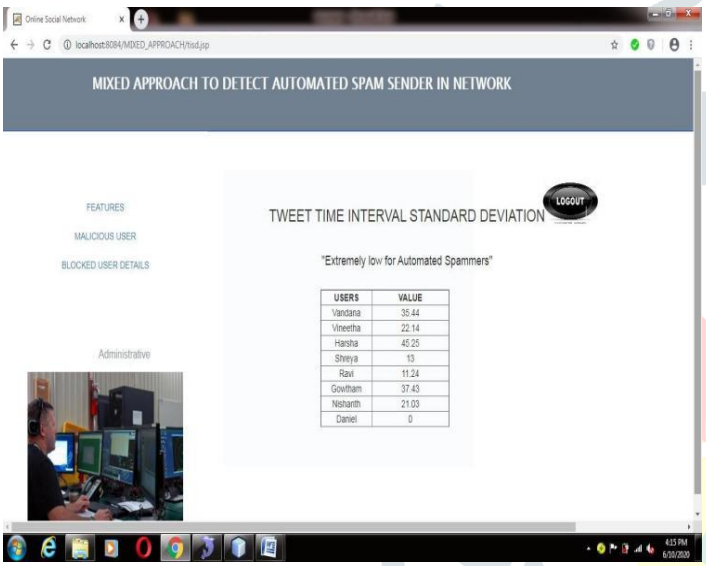

Fig. 4: Tweet time Interval Standard Deviation

Fig. 5: Blocked Users

Fig. 2 shows all the features which are observed in the homepage of Admin. Fig. 3 and Fig. 4 shows the Retweet Ratio (RR) values and Tweet time Interval Standard Deviation (TISD) values of all the users registered with the Social Network. Fig.5 represents the enhancement in which the admin can block the spammer from getting logged in to their corresponding account.

VIII. Conclusions and Future Work

In the proposed method, 9 features are grouped into three classes which are helpful to detect automated spammers in a network. Interaction and Content-based features are more effective for the detection of spammers, while Metadata-based features are less effective. By using the proposed method, automated spammers can be easily detected in an Online Social Network.

The three classes into which the features are classified are metadata, interaction and content and the various features are shown in Table 1.

In this project, the system gives a solution for detecting automated spammers on Online Social Networks. Future enhancement can be done by taking the real data from any Online Social Networks like Facebook, Twitter, Whatsapp, etc. and detect spammer accounts upto a great extent.

References

[1] Mohd Fazil, Muhammad Abulaish, "A Hybrid Approach for Detecting Automated Spammers in Twitter," *IEEE Transactions on Information Forensics and Security*, Vol. 13 No. 11, pp. 2707-2719, Nov. 2018.

[2] Sangho Lee and Jong Kim, "WarningBird: A near real-time detection system for suspicious URLs in Twitter stream," *IEEE Transactions on Dependable and Secure Computing.*,vol. 10, no. 3, pp. 183–195, May 2013.

[3] Chao Yang, Robert Harkreader, and Guofei Gu, "Empirical evaluation and new design for fighting evolving Twitter spammers," *IEEE Transactions on Information Forensics and Security, vol. 8, no. 8, pp. 1280–1293, Aug. 2013.*

[4] M. Sangeetha, S. Nithyanantham, M. Jayanthi, "Comparison of Twitter spam detection using various machine learning algorithms," International Journal of Engineering & Technology 7(1-3):61, Dec.2017.

[5] Abdullah Talha, Resul Kara, "A Survey of Spam Detection Methods on Twitter," International Journal of Advanced Computer Science and Applications (IJACSA) 8(3), March 2017.

[6] Arpna Dhingra, Shruti Mittal, "Content Based Spam Classification in Twitter using MultiLayer Perceptron Learning," International Journal of Latest Trends in Engineering and Technology (IJLTET), Vol. 5 Issue 4, July 2015.

[7] Cristina Radulescu, Mihaela Dinsoreanu, Rodica Potolea, "Identification of Spam Comments using Natural Language Processing Techniques," IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP), (p. 7), Oct. 2014.

[8] Saini Jacob Soman, S. Murugappan, "Detecting Malicious Tweets in Trending Topics using Clustering and Classification," 2014 International Conference on Recent Trends in Information Technology (p. 6), IEEE, Dec. 2014.

[9] K. Kandasamy, Preethi Koroth, "An Integrated Approach to Spam Classification on Twitter using URL Analysis, Natural Language Processing and Machine Learning Techniques," 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science (p. 5), IEEE, Apr. 2014.

[10] M.McCord, M.Chuah, "Spam Detection on Twitter using Traditional Classifiers," (p. 7), ATC'11: Proceedings of the 8th international conference on Autonomic and trusted computing, Canada: IEEE, Sept. 2011.