[Bhuvaneswari*, 5(11): November 2018]
DOI- 10.5281/zenodo.1482354

ISSN 2348 – 8034
Impact Factor- 5.070

GLOBAL JOURNAL OF ENGINEERING SCIENCE AND RESEARCHES

PROACTIVE BIMODAL SPAM COMMUNITY DETECTION IN ONLINE SOCIAL NETWORKS

Bhuvaneswari A[*1], Pias Milton A[2], Kirubakaran V[3] & Soorya Prasad K[4]
[*1]Teaching Fellow, Department of Computer Technology, Madras Institute of Technology, Chennai, India
[2,3,4]UG Student, Department of Computer Technology, Madras Institute of Technology, Chennai, India

Abstract

In recent times, Online Social Networks like Twitter is facing more challenges in the direction of Spam accounts and spammer community. The popularity and open structure of twitter have attracted a large number of automated programs known as bots, which appear to be a double-edged sword to Twitter. Legitimate twitter users complain that their accounts are mistakenly suspended by twitter’s anti-spam action. Spammers are only identified after many benign users are affected. A comprehensive and empirical evaluation is performed in Twitter spam show that more and more spammers are evolving. The intention of the existing system is only to identify individual spam accounts. The spam accounts are classified using multi-class SVM classifier as namely benign, verified, spam, hacked accounts. In this paper, the suspicious spam community is without affecting the legitimate benign accounts using bimodal spam community detection algorithm whichinvolves graph network and content based features shows the effectiveness of the model.

Key words: *Online Social Network, Spam Identification, Community Detection, SVM classifier, Twitter.*

I. INTRODUCTION

Twitter is an Online Social Networking (OSN)service that enables users to post and access messages called “tweets". Registered users can read and post tweets, but unregistered users can only read them. Users can tweet via the Twitter website, compatible external applications (such as for smartphones), or by Short Message Service (SMS) available in certain countries. The service rapidly gained worldwide popularity, with more than 100 million users who in 2012 posted 340 million tweets per day. Retweeting is when a tweet is forwarded via Twitter by users. Both tweets and retweets were tracked to see which ones are most popular. Users may subscribe to other users' tweets—this is known as "following" and subscribers are known as "followers". In addition, users can block those who have followed them. Twitter has a challenging problem namely “spam”issue. The social network users propagate "follow anyone" model allows a steady influx of spammers to reach out to many people in a short amount of time, hoping at least one of them will click on a sketchy link. While a recent glitch preventing users from sending links in direct messages may have been Twitter's latest defense against spam, there's no doubt that it remains a major problem for the Twitter platform.

Various state-of-art methods tried to resolve the spam communities over Twitter.Anextroversivecommunity behaviors such as first activity, activity preference, activity sequence and action latency where introversive community behaviors include browsing preference, visit duration, request latency and browsing sequence [1]. Limitations are it is hard to trace the behavior patterns of users who access an OSN solely via APIs, thus, this method may not be applicable for those rare cases.Social network Aided Personalized and effective spam filter (SOAP) was introduce which greatly improved the performance of Bayesian spam filters in terms of accuracy, attack-resilience, and efficiency of spam detection [2]. Whitelists and blacklists both maintain a list of addresses of people whose emails should not and should be blocked by the spam filter respectively. Three types of Sybil attacks according to the spam attacker’s capabilities and some Mobile Spam Detection. It detects spam accounts by analyzing its behavior [3].

A botnet which is a collection of spam machines (bots) receiving and responding to commands from a server [4]. Binary obfuscation, Anti-analysis, Security suppression, Rootkit technology are some of the evasion tactics. It involves Active detection, Injection and Suppression. Binary Obfuscation and Encryption degrade the accuracy of detection approaches that rely on reverse engineering of bot binary. An algorithm which focuses on criteria such as harmful links, aggressive following behavior, posting repeatedly to trending topics, posting duplicated tweets, posting links with unrelated tweets etc [5]. A bipartite network between users and their corresponding tweets are constructed to compute spam scores for both the users and the tweets.The drawback is that the traditional twitter spam detectors focus on the spamming behaviors and cannot detect spam accounts which do not spam aggressively in social networks. Therefore, they are not effective to detect less active spam accounts and their spam tweets.

Twitter spammers where they achieve their malicious goals such as sending spam, spreading malware, hosting botnet channels, and launching other underground illicit activities [6]. The tactics includes gaining more followers, posting more tweets, mixing normal tweets and heterogeneous tweets. The graph-based features such as local clustering coefficient and betweenness centrality are relatively difficult to evade, these features are also expensive to extract. The main drawback of this paper is that the number of our identified spammers is only a lower bound and it is very difficult to obtain a perfect ground truth from such a big dataset.URL shortening technique [7] which twitter users usually use to reduce the URL length because tweets can contain only a restricted number of characters. They proposed warningbird, a suspicious URL detection system for Twitter in which they considered correlations of URL redirect chains extracted from a number of tweets. The two main drawbacks are attackers really had reduced the lengths of redirect chains because too long chains could be treated as malicious or they had applied dynamic redirections to prevent simple static crawlers.

The twitter's popularity and very open nature for exploitation by automated programs, i.e., bots. They have designed an automated classification system that consists of four main parts: the entropy component, the spam detection component, the account properties component, and the decision maker [8]. The entropy component checks for periodic or regular tweet timing patterns; the spam detection component checks for spam content. The decision maker summarizes the identified features and decides whether the user is a human, bot, or cyborg. The drawback is that the classification system does not add a comprehensive solution in order to suspend the various categorizations of automated programs. An approach automatically identifies software relevant tweets from a collection or stream of tweets. A corpus of tweets was then used to test the effectiveness of the trained language model [9]. With limited resources in the Opportunistic Networks where some selfish or malicious nodes can drop data packets quietly, degrading the network performance [10]. TRSS scheme is based on the observation that nodes move around and contact each other according to their common interests or social similarities. Every node can evaluate other nodes trustworthiness using direct or indirect trust model.

In order to improve the efficiency of blocking a Sybil attack by combining neighborhood similarity method and improved Knowledge Discovery tree based algorithm is used [11]. Using the symbolic identification by trust relationship, frequency and length of each node in specific interval and the variance are calculated. The Knowledge Detection tree finds the maximum variance between the connectivity of nodes and also calculates the length and frequency of nodes in a particular time interval. It mentioned threats and solutions that exist in these networks, including privacy violations, identity theft, and sexual harassment [12]. Classic threats include privacy and security threats, modern threats are unique to the environment of OSNs, combination threats combine various types of sophisticated and lethal attacks. Privacy-preserving OSNs, such as Safebook and developing solutions for privacy-preserving ad hoc social networks such as the Semantics based Mobile Social Network (SMSN) framework.

An optimization formulation that incorporates sentiment information into a novel social spammer detection framework [13]. An exploratory study on two Twitter datasets was used to examine the sentiment difference between spammers and normal users. The sentiment information are then modeled with a graph Laplacian and incorporated into an optimization formulation. The spammers send unwanted tweets to Twitter users to promote websites or services, which are harmful to normal users [14]. Spam detection [15] was then transformed to a binary classification problem in the feature space and solved by conventional machine learning algorithms.The web applications such as Hotmail, Facebook, Twitter, and Amazon have become important means of satisfying working, social, information seeking, and shopping tasks, suspicious user's care increasingly attempting to engage in dishonest activity, such as scamming money out of Internet users and faking popularity in political campaigns. In the present day world, people are so much habituated to OSN. Because of this, it is very easy to spread spam contents through them [16]. In this paper they are proposing an application which uses an integrated approach to the spam classification in Twitter.

II. PROPOSED BIMODAL SPAM COMMUNITY DETECTION

The intention of the proposed system is to identify individual spam accounts and its communities. The system uses bimodal techniques namely, (i) spam accounts - community network detection and (ii) keyword based spam content detection. Consider the case, when a spammer creates more number of spam accounts and gives bidirectional relationship between them. The community detector algorithm which checks for malicious tweets and its keywords (content based) that is identified by bots, heterogeneous tweets or different semantic tweets.If the malicious tweet is found, then the account is marked as "Detected". The detected account is suspended for a specified number of time period. The neighbors associated with the spammer account are separated and looked for the detected friends. If no detected friends are found, it is again checked for any verified follower and marks it as hacked account. There are four different classification of twitter accounts namely benign, verified, spam, hacked accounts using a multi-class SVM classifier [17].

Benign Accounts

Benign accounts are normal genuine twitter accounts. They follow and are followed by many other benign accounts and some verified accounts same as in Figure 1.If detected friends are found and the ratio is high, and then selects a random friend from the community and check whether if he has all the detected accounts as his friend and examining the previous tweets, he is identified as a spam. Inactive spammers identified in a spam community network are put in a halt state.

Verified Accounts

Twitter concentrate on highly sought users in music, acting, fashion, government, politics, religion, journalism, media, sports, business and other key interest areas. These accounts are given verified signature by the Twitter Inc. to ensure that they are official genuine accounts and they follow only known accounts and are followed many benign users such as in Figure 2.The FOFO Ratio and reputation score are widely used for identification of spammers. And by this way spam community is identified using spam community detector and if no spammers are identified, the tweets are posted to public timeline.

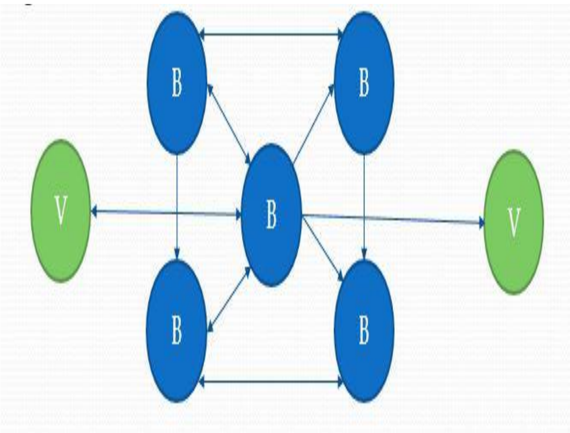

Fig 1. Benign Accounts in Twitter

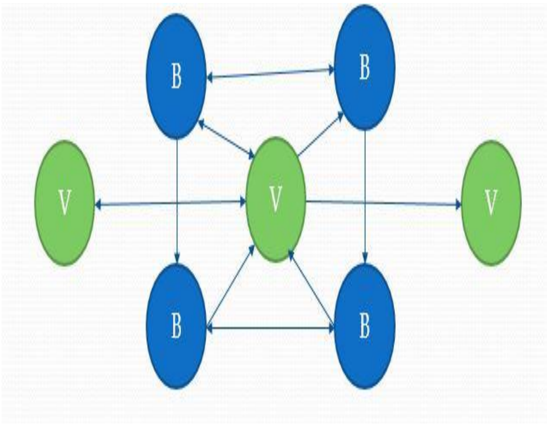

Fig 2. Verified Accounts in Twitter

Spam Accounts

Spam accounts are those which are created to spread malicious contents and URLs, degrade others, etc., in order to gain credibility, these fake accounts will try to become ‘friends’ or follow verified accounts as in Figure 3, e.g., celebrities and public figures with the hope that these accounts be friend or follow them back. When genuine accounts befriend or follow back fake accounts, it legitimizes the account and enables it to carry out spam activities. Another way for spammers to attack is to hack into and take over a user’s account, spreading fake messages to the user’s authentic followers.

Hacked Accounts

Hacked accounts are normal benign user’s accounts being hacked by the hackers or spammers for their personal gain which is showed in Figure 4.When spammers created a closed network like this then they evade the existing detecting features such as fofo ratio, reputation score, betweenness centrality etc.,FOFOratiois the ratio of the number of an account’s followings (F_1) to its followers (F_2).

$$F = \frac{F_1}{F_2} \qquad (1)$$

Reputation score (RS), which is the ratio of the number of an account’s followers to the sum of its followers and followings, could be viewed as a variant of FOFO Ratio.

$$RS = \frac{F_2}{F_1 + F_2} \qquad (2)$$

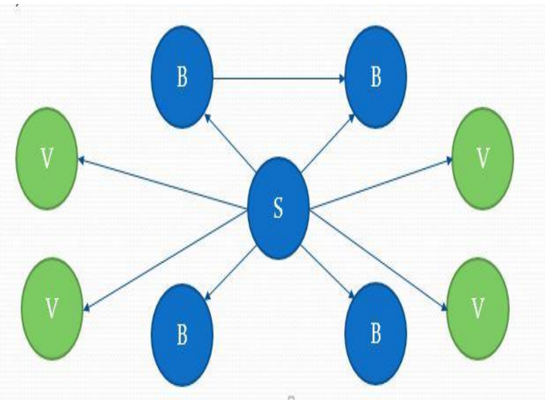

Fig 3. Spam Accounts in Twitter

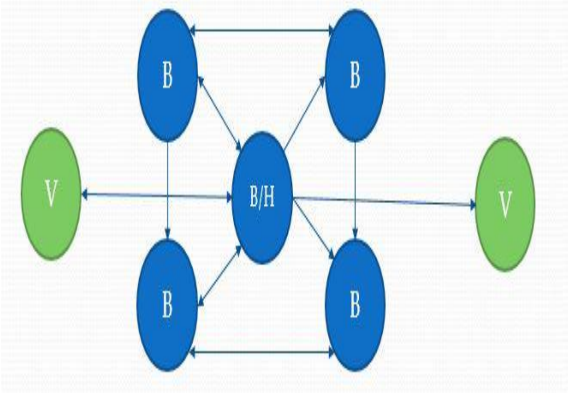

Fig 4. Hacked Accounts in Twitter

Betweenness Centrality is a centrality measure of a vertex within a graph. Vertices that occur on many shortest paths between other vertices have a higher betweenness than those that do not. A Twitter spammer will typically use a shotgun approach to finding victims, which means it will randomly follow many unrelated accounts. As a result, when the Twitter spammer follows these unrelated accounts, the spammer creates a new shortest path between those accounts through the spam account.Thus, the betweenness centrality of the spammer will be high. The fewest shortest path passing describes a normal or benign account while many shortest path passing denotes a spam account. This clearly illustrates the betweenness centrality. After identifying a particular spammer through his bot tweets or heterogeneous tweets, it is checked with the detected spammers and as a whole a community are identified. Hence it is easy to identify spammers as a whole and restricted them before they make impact in the social media using algorithm (1) below.

Algorithm (1) :Bimodal Spam Community Detection Algorithm

Input

```
 • $U = \{u_1, u_2, ...u_n\}$ set of tweets
 • $X = \{x_1, x_2, ...x_m\}$ set of users
 • $F = \{f_1 , f_2 , ...f_{fcount}\}$ set of followers
1: Set $u^{0}_{j} \leftarrow 0$ $\forall j= 1..n$, $i \leftarrow 0$, $result \leftarrow benign$, $no\_detected \leftarrow 0$
2: Stage 1:
```

```
for each $u_j \in U$ do
if malicious($u_j$) == true then
Stage 2:
        Mark the account as detected
        if the account is verified by twitter
            Set result ← hacked
        end if
Set fcount ← followers count of the account
Stage 3:
         for each $f_k \in F$ do
             if it is detected earlier then
            no_detected ← no_detected + 1
         end if
         end for
Compute detect_ratio = no_detected / fcount
         if detect_ratio >= 0.50 then
             Set result ← spam
              Hold other undetected followers' tweets
         end if
     end if
end for
return result
```

If two accounts follow each other, it consider there is a bidirectional link between them. The number of bidirectional links of an account reflects the reciprocity between an account and its followings. Since Twitter spammers usually follow a large number of legitimate accounts and cannot force those legitimate accounts to follow back, the number of bidirectional links that a spammer has is low.Since the existing features find out the spam accounts only after they made a huge impact on other accounts in social media as shown in Figure 5.

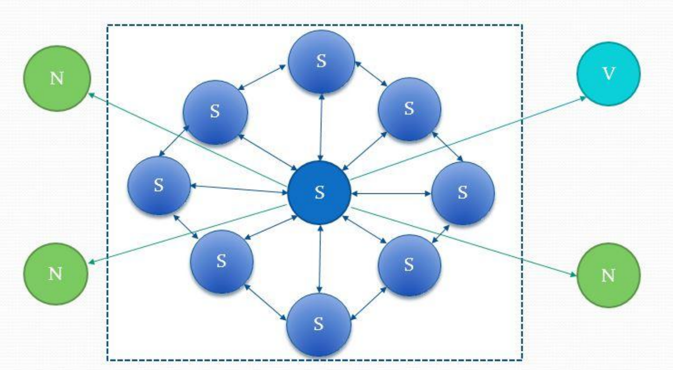

Fig 5. Closed Spam community which evades existing features

Finding a Malicious Tweet

Local Clustering Coefficient of a vertex is the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. When they evade the existing features first the accounts needs to be checked whether they are hacked or not. For finding it ensure any verified accounts by twitter follows the suspicious account. If it is true then the account is hacked by someone.The Spam Tweet Ratio (STR) is the ratio of keyword matched to N - total no of words in extracted tweet

$$STR = \frac{Keyword\_Matched}{N} \quad (3)$$

It concludes that the account is a benign one and it is hacked by some spammer or else they are spam accounts. Then the account is marked as detected. The malicious tweet are identified using keyword based which is shown in algorithm (2) below.

Algorithm (2): Keyword based Spam Detection Algorithm

Input

```
• $K= \{k_1, k_2, ...k_n\}$ set of keywords
• $U = \{u_1, u_2, ...u_n\}$ set of tweets
• $t$ = current tweet
Stage 1:
Remove all non – alpha numeric characters such as digits 0 – 9 or characters like *,!,@,#.
Remove links that starts with http or https
Remove mentions of other users that start with @user or any hashtags that start with #hashtag
Stage 2:
Set keyword_matched ← 0
for each $u_j \in U$ do
    if t matches with the jth previous tweets of that user
        return true
    end if
end for
for each $k_j \in K$do
    if the pattern word matches with words in tweets
        keyword_matched ← keyword_matched + 1
    end if
end for
Compute spam_tweet_ratio = keyword matched / total no of words in extracted tweet
if spam_tweet_ratio >= 0.25
    return true
else
return false
```

There are three major areas of identifying the spammers. First it identify the type of accounts. Initially every account is considered to be benign accounts or verified accounts. Then separately every account is checked with the ratios such as FOFO ratio and reputation score ratio. And accordingly a set of spammers will be identified and then a particular spammer is chosen with the help of spam tweets. And the spammer is now checked for any detected spammer follower. And the follower is also marked as spammer. The same process is repeated until every spammer is identified.The spam tweet ratio (STR) uses the tuning threshold of 0.25 which identify the spam nodes in most simple way in the proposed method.

III. EXPERIMENTS &RESULTS

The Twitter Streaming API allows to receive tweets and notifications in real time from Twitter. However, it requires a high-performance, persistent, always-on connection between your server and Twitter. The experimental result makes an account of its suspicious followers. The implementation searches the entire followers list and makes sure whether there is any detected friends or not. If there are any friends and their ratio is too high then the total network of that account is considered as a spam network. The ratio should be considerable that is it should be greater than equal to half the count of its followers. Since it is found that the account's network as a spam network in previous module then the accounts in the network are taken into consideration one by one and simultaneously checked for spam accounts. If spam accounts are found, the users are blocked whereas if it is a normal account and if it did not satisfy the proposed properties, the tweet is posted on the public timeline. Spam Community Identification – Twitter Handling consists of several modules. Processing datasets module handle processing json API response to give individual information.

During pre-processing, the extra characters and links are removed to check the common word ratio to check whether two tweets are same and whether the tweets are posted by the same person are checked. Searching a particular spam defining keyword gives various details such as Relationships, URLs in tweets, Domains in tweet, and Hash tags in tweet. The data is parsed that gives the value of a particular data which is used for future computations. After the computations, the objects are parsed and store it for future computations. The crawled data is parsed for required information and the tweet information is found. It crawl the friends or followers list of a person who posted some irrelevant tweets. A particular keyword are searched in the social network to find some malicious behavior and so again it is given as a JSON object and parsed it. The particular keyword containing tweets are now obtained and found the accounts who posted that particular tweet. The dataset from the twitter is crawled using Twitter API and obtained the response in JSON object format with 20 Twitter user's network and 700 tweet content. Based on the betweeness centrality measure, FOFO ratio and reputation score, the Figure 6 identifies the node {9, 12, 13, 14 and 19} as spam community nodes which follows algorithm (1) and algorithm (2).

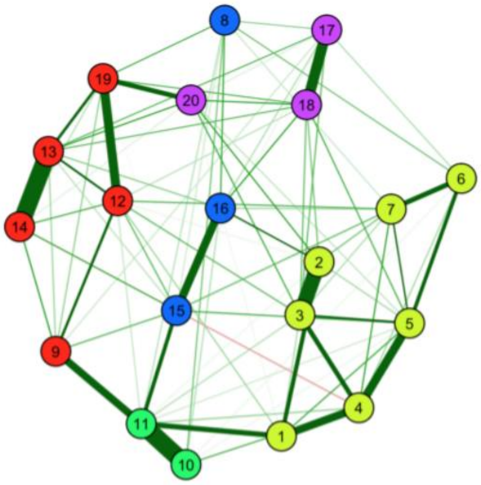

Fig 5. Spam Nodes Community Detection – Graph and Keyword based Method

The tweet information gives us two factors: FOFO Ratio and Reputation Score. And this two factors are checked for individual users. The same tweets cannot be posted in Twitter as it does not allow us to post it. But to vary a tweet from another tweet spammers use some characters and links which will be considered as a separate tweet same.The detected account is suspended for a while. The neighbors associated with the identified spammer account are separated and looked for the detected friends. If no detected friends are found, it is again checked for any verified follower and marks it as hacked account. If detected friends are found and the ratio is high, it selects a random friend from the community and check whether if he/she has all the detected accounts as their friend and examining the previous tweets, he is identified as a spam. Inactive spammers identified in a spam community network are put in a halt state. And by this way spam community is identified using spam community detector and if no spammers are identified, the tweets are posted to public timeline.Hence the keyword based method and network based method is combined in the proposed work establish the high FOFO ratio, reputation score and SVM classifier accuracy is compared in Table 1.

Table 1. Classifier accuracy of Bimodal Spam Community Detection

Method Used	**FOFO ratio**	**Reputation Score**	**SVM Classifier Accuracy**
Spam Network based Method	0.54	0.65	71.4%
Keyword based Method	0.59	0.72	78.7%
Proposed Bimodal (Network + Keyword) Method	**0.69**	**0.79**	**84.5%**

IV. CONCLUSION

The proposed system that paves way for finding the spam community identifies associated spammers in the beginning stage itself. Most of the existing studies utilize machine learning techniques to detect Twitter spammers. Hackers can also create shortened URLs to easily redirect you to malicious sites, since the URL itself gives you no indication of the site name. Twitter spammers are evolving to evade existing detection features. This are largely controlled by several new detection features to detect more Twitter spammers. To deeply understand the effectiveness and difficulties of using machine learning features to detect spammers, an analysis of the robustness of detection features is required. When a spammer creates more number of spam accounts and gives bidirectional relationship between them, the identification of spammers will become a complex task. Even the spammers who are identified are detected only after huge impact on normal users. The comparison of proposed method with SVM classifier accuracy of 84.5% which is comparatively higher than state-of-art method which shows the effectiveness of the proposed bimodal spam community detection.

REFERENCES

1. *X.Ruan, Z.Wu, H Wang and S.Jajodia, 2016. "Profiling Online Social Behaviors for Compromised Account Detection". IEEE Trans. Information Forensics and Security, 11(1), pp.176-187.*
2. *H. Shen, and Z. Li, 2014. "Leveraging social networks for effective spam filtering". IEEE Transactions on Computers, 63(11), pp.2743-2759.*
3. *K. Zhang, X.Liang, R. Lu, and X.Shen, 2014. "Sybil attacks and their defenses in the internet of things". IEEE Internet of Things Journal, 1(5), pp.372-383.*
4. *S.Khattak, N.R.Ramay, K.R.Khan, A.A. Syed, and S.A.Khayam, 2014. "A taxonomy of botnet behavior, detection, and defense". IEEE communications surveys & tutorials, 16(2), pp.898-924.*
5. *I.A.Bara, C.J. Fung, and T.Dinh, 2015, May. "Enhancing Twitter spam accounts discovery using cross-account pattern mining". In Integrated Network Management (IM), 2015 IFIP/IEEE International Symposium on (pp. 491-496). IEEE.*
6. *C.Yang, R. Harkreader, and G.Gu, 2013. "Empirical evaluation and new design for fighting evolving twitter spammers". IEEE Transactions on Information Forensics and Security, 8(8), pp.1280-1293.*
7. *S. Lee, and J.Kim, 2013. "Warningbird: A near real-time detection system for suspicious urls in twitter stream". IEEE transactions on dependable and secure computing, 10(3), pp.183-195.*
8. *Z.Chu, S.Gianvecchio, H. Wang, and S.Jajodia, 2012. "Detecting automation of twitter accounts: Are you a human, bot, or cyborg?". IEEE Transactions on Dependable and Secure Computing, 9(6), pp.811-824.*
9. *A.Sharma, Y. Tian, and D.Lo, 2015, March. "Nirmal: Automatic identification of software relevant tweets leveraging language model". In Software Analysis, Evolution and Reengineering (SANER), 2015 IEEE 22nd International Conference on (pp. 449-458). IEEE.*
10. *L.Yao, Y.Man, Z.Huang, J. Deng, and X.Wang, 2016. "Secure routing based on social similarity in opportunistic networks". IEEE Transactions on Wireless Communications, 15(1), pp.594-605.*
11. *S.J. Samuel, and B.Dhivya, 2015, March. "An efficient technique to detect and prevent Sybil attacks in social network applications". In Electrical, Computer and Communication Technologies (ICECCT), 2015 IEEE International Conference on (pp. 1-3). IEEE.*
12. *M.Fire, R. Goldschmidt, and Y.Elovici, 2014. "Online social networks: threats and solutions". IEEE Communications Surveys & Tutorials, 16(4), pp.2019-2036.*
13. *X.Hu, J.Tang, H. Gao and H., Liu, 2014, December. "Social spammer detection with sentiment information". In Data Mining (ICDM), 2014 IEEE International Conference on (pp. 180-189). IEEE.*
14. *C., Chen, J., Zhang, Y., Xie, Y., Xiang, W., Zhou, M.M., Hassan, A. AlElaiwi, and M., Alrubaian, 2015. "A performance evaluation of machine learning-based streaming spam tweets detection". IEEE Transactions on Computational social systems, 2(3), pp.65-76.*
15. *M.Jiang, P. Cui, and C.Faloutsos, 2016. "Suspicious behavior detection: Current trends and future directions". IEEE Intelligent Systems, 31(1), pp.31-39.*
16. *K. Kandasamy, and P.Koroth, 2014, March. "An integrated approach to spam classification on Twitter using URL analysis, natural language processing and machine learning techniques". In Electrical, Electronics and Computer Science (SCEECS), 2014 IEEE Students' Conference on (pp. 1-5). IEEE.*

17. *G.Madzarov, D. Gjorgjevikj, and I.Chorbev, 2009. "A multi-class SVM classifier utilizing binary decision tree". Informatica, 33(2).*