ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 3, March 2015**

# Analysis and Design of Efficient generalized Forensic framework for Detecting Twitter Spammers

Ankita M. Ghate, L. G. Malik

Project Scholar, Dept of CSE, G. H. Raisoni College of Engineering, Nagpur (M.S), India

Professor and Head of Department, Dept of CSE, G. H. Raisoni College of Engineering, Nagpur (M.S), India

**ABSTRACT**: Asocial networking web site could be a platform to make social networks or social relations among those who share interests, activities, backgrounds or real-life connections. Users pay a good deal of your time on known social networks(e.g.Facebook,Twitter, SinaWeibo, etc.), reading news, discussing events and posting their message. Unfortunately, this quality conjointly attracts a big quantity of spammers whoincessantly expose malicious behaviours (e.g. post messages containing commercial topics or URLs, following a bigger quantity of users, etc.), resulting in nice inconvenience on traditional users' social activities. The propose system will workAIER (Artificial intelligence for emergency response) approach for detecting twitter spammer .We first collected and labelled a large dataset with 34 K trending topics and 20 million tweets then, construct a labelled dataset of users and manually classify users into spammers and non-spammers; after that, abstract a set of novel features from message content and users' social behavior. Our experiments show that true positive rate of spammers and non-spammers could reach 99.1% and 99.9%.

**KEYWORDS**: Twitter,ArtificialIntelligence,Data Collection Log Data collection spammers, network forensic.

## I. INTRODUCTION

In the last decade social media has turned into one of the most essential parts of the people's daily lives. Besides keeping track and communicating with people, it gave people a chance toreach the masses and interact with them. Some highly popular social media sites such as Twitter,Facebook, Myspace and YouTube made it possible for people to create their own communitysharing common interests. Especially on Twitter, users can follow the people relevant to theirinterests so that within the community they can keep up to date with the most recent news, getinvolved in the discussions and share ideas. However, people can also get notified from people
Whomthey don't follow in some other ways such as unsolicited messaging or through Twitter's public timeline. Twitter messages are not longer than 140 characters, and are known as tweets.

Even though Twitter giving an opportunity to reach people is quite favorable, this isconsidered to be an opportunity for spammers. This created a new way for them to reach theirpotential victims by sending them spam tweets. Twitter spammers have malicious intents such as unfair advertising of products, spread of malware, and stealing user's private information(phishing). This is made easier nowadays by usage of URL shortening services like bit.ly. Twitter is unable to filter malicious links which are shortened using such services. The maliciousactivities of the spammers have an adverse impact on the business of genuine e-commerce websites; moreover, wastes valuable human time. Therefore, in this paper we investigatecharacteristics of Twitter accounts which can distinguish between spam and legitimate Twitteraccounts.

**Our Contribution:**1] In our proposed system spammer detection from real time database and find out the accuracy or generate graph. And provide this application in the cloud.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

## II. RELATED WORK

As a part of the literature review several papers have been examined and analyzed to form a set of metrics that could depict the features of Twitter spam account. First paper reviewed was published by Benevenuto et al. [1] based on a research that aimed to capture the spam accounts through some features related to tweet content and users' social behavior. These features and a dataset consisting of over 54 million users and 1.8 billion tweets was the input to a classification algorithm. As a result, approximately 70% spam accounts and 96% legitimate accounts have been identified, and a set of important common characteristics of spam accounts were proposed. Second paper published by McCord et al. [2] proposed a set of classifications based on user and content features which are used in spam detection. As for the dataset collection active Twitter users were crawled and their most recent 100 tweets, followers and following information was gathered, and through the user and content based features the dataset was analyzed. Consequently, Random Forest which is one of the four classification techniques performed in the study gives the best precision of 95.7%. Stringhini et al. [3] presented a study which managed to detect over 15 thousand spam accounts in collaboration with Twitter, and have them suspended. This study aimed to find out how spam accounts operate not only in Twitter, but also two other major social network sites Facebook and Myspace. Wang et al. [4][5] showed how machine learning can be adapted for the spam detection purpose in Twitter. For this experiment several features (graph-based and content based) were extracted from most recent 20 tweets and user's social features in order to form a real dataset. This study was proven to be efficient for detecting spam accounts in Twitter. Lee et al. [6] presents a 7 months long study for making long term observations on spammers in social media. Through deployment of 60 honeypot accounts they managed to find 36.000 candidate spam accounts. As a result of this study they found out that the spammers revealed some key distinguishing characteristics in their behaviors. One of the disadvantages found out about the reviewed papers is that some of them require metrics that are very expensive to compute. For example, Stringhini et al. [7] was able to do collaboration with Twitter, so they managed to get over computational difficulties.

## III. PROPOSED FRAMEWORK FOR DATA COLLECTION IN SYSTEM

In this paper we will implement system analysis data from twitter dataset. It collects data from dataset and detects spammers from user tweets.

### A. *Data Collection:*

In the Data collection step the forensic method is to identify possible information of source node and collect all information related to source, destination, and time of activity, process ID, port address. Major sources of data are personal computers, browsing log data information, digital storage media which store image of data, Routers in network device, Cell Phones, Digital Camera, Network traffic in system etc

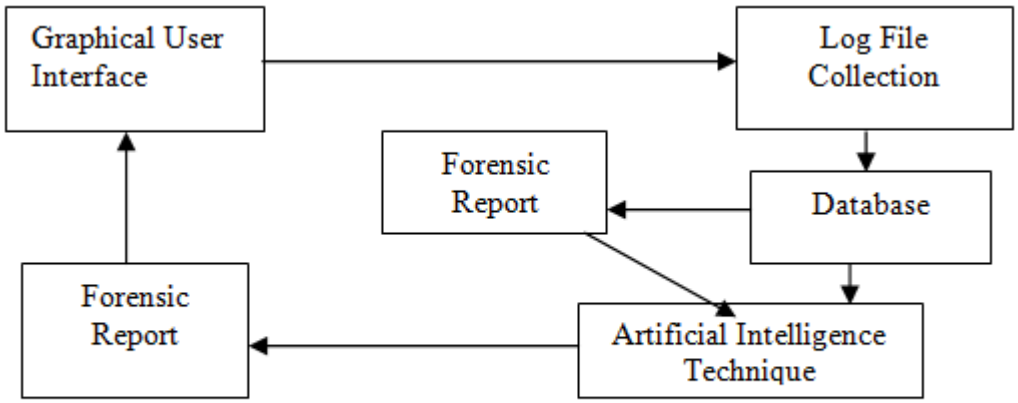



Fig. 1. Block Diagram of proposed system

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

## Graphical Interface:

The graphical interface provides investigators to provide ease in find out the evidence from the computer system where crime is occur. This is also shows the result of processing data in presentable form. The representation of data can be able to analysis with using Graph, tabular form or in form of report.

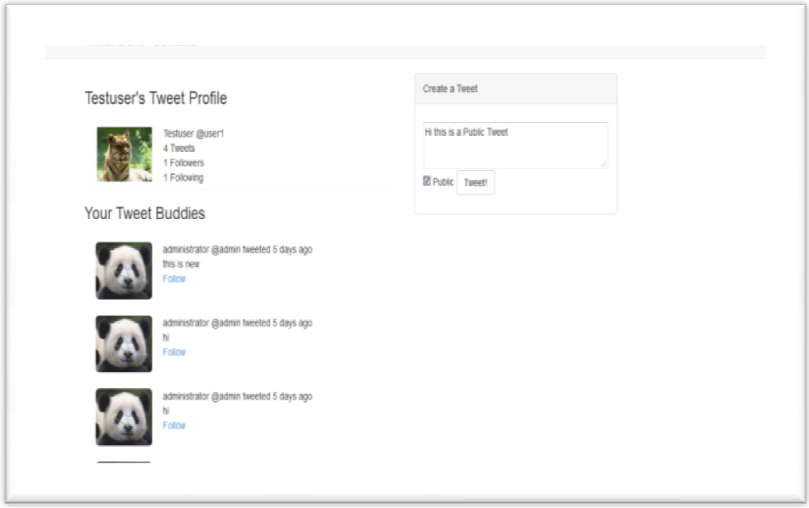



Fig.2. GUI of twitter

B. *Database:*

Database is used to store and collects evidence from crime side using proposed system and stored required data for investigation. The database provides flexibility to store the data in table format. The attribute gives more information about the normal user and the malicious user, also it able to know the source IP address of user. We have download and used twitter dataset that size is 230K from Amazon Web Service.

C. *LogFile Collection:*

In this step log data of browser have been captured and stored in database for investigation. This file log contains history of browser in which user log information retrieved.
The figure 2 shows the running process in the computer system including running process name, session name and usage of memory in system. The all process file related information is retrieve in text file for further examination analysis of evidence.Figure 3 shows log information data from the web browser which include browsing date, URL that are recently browsed. In crime investigation log history of web browser gives more information about thread or normal user file analysis.

D. *Artificial Intelligent Technique:*

This approach identifying the network spammers from network using artificial intelligence technique.In this approach analysis, detecting and investigate the twitter spammers.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

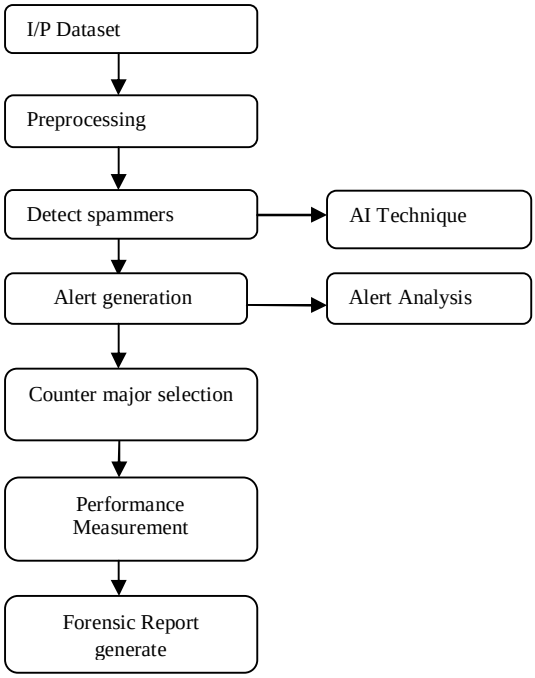



Fig 3: Steps for investigation of Spammers

## IV. GENERAL METHODOLOGY FOR SPAM DETECTION APPROACH

In our system we use the artificial intelligence technique for detection of spammers. We use two approaches i.e. 1]user keyword violation and 2]find out repeated keywords.

### A. *User Keyword* Voilation*:*

In this approach finding per day spammers according user tweets.and generate graph as per user tweets.It also finding the unwanted keywords or restricted keywords.This keyword creates problem for user.our approach is advantage for that particular person.our approch also generates forensic report.

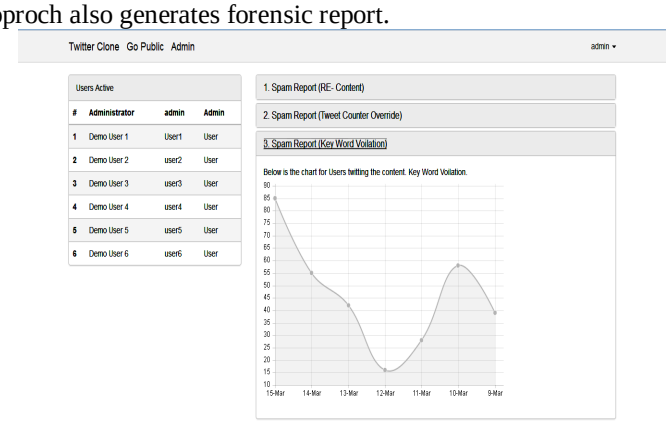



Fig 4:Keyword violation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 3, March 2015

B. *User Keyword* Override*:*

*In this approach finding repeated tweets.when same tweets sending number of times per day this tweets called as a spammers.This keyword analysis and detected from server and generate forensic report.*

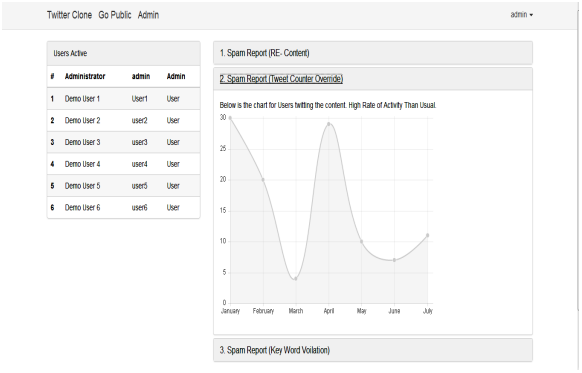
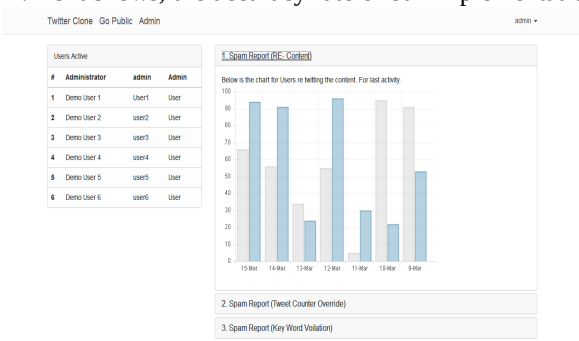



Fig 5:Keyword Override

## V. RECOUNT GRAPH

Thisgraph counting the spam keyword.The user rewetting the content from above graph. Following graph shows the performance of implemented system. As it shows, the accuracy rate of our implementation in percentage.




## VI. CONCLUSION AND FUTURE ENHANCEMENT

In this paper cybercrime data is collected withusing of proposed methodology, For evidence the log file is captured from real time database and detecting spammers using artificial intelligence approach. Here two techniques used keyword violation and override keyword and detecting spammers for that particular technique. Also generating forensic report from that technique.

In future work, the proposed method can suggest investigating under real time network settings for forensic investigation and generating forensic report.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 3, Issue 3, March 2015**

## REFERENCES

1. F. Benevenuto, et al., "Detecting spammers on Twitter," in Collaboration, electronic messaging, anti-abuse and spam conference (CEAS), 2010.
2. K. Lee, et al., "Seven Months with the Devils: A Long-Term Study of Content Polluters onTwitter," in ICWSM, 2011.
3. M. McCord and M. Chuah, "Spam detection on Twitter using traditional classifiers," in Autonomic and Trusted Computing, ed: Springer, 2011, pp. 175-186.
4. G. Stringhini, et al., "Detecting spammers on social networks," in Proceedings of the 26th Annual Computer Security Applications Conference, 2010, pp. 1-9.
5. A. H. Wang, "Detecting spam bots in online social networking sites: a machine learning approach," in Data and Applications Security and Privacy XXIV, ed: Springer, 2010, pp. 335-342.
6. A. H. Wang, "Don't follow me: Spam detection in Twitter," in Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on, 2010, pp. 1-10.
7. http://www.alexa.com/topsites
8. Monika Verma andSanjeevSofat "Techniques to Detect Spammers in Twitter- A Survey"International Journal of Computer ApplicationsJanuary 2014.
9. Jeremy Davis,JoeMaclLean,David Damper" Methods of information hiding and detection in file systems." 2010 Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering.
10. Ying Zhu, Member, IEEE" Attack Pattern Discovery in Forensic Investigation of Network Attacks2011"Ieee Journal On Selected Areas In Communications, Vol. 29, No. 7, August 2011.

## BIOGRAPHY

**Ankita M. Ghate**received the B.E degree in Information technology from the Nagpur University, India, in 2013 and pursuing M.Tech degree in Computer Science & Engineering through G.H.Raisoni College of Engineering, Nagpur, India.

**Dr.Latesh Malik**is an Professor & Head of Computer Science & Engineeringin Computer Science & Engineering department, G. H. Raisoni College of Engineering, Nagpur, India.