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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 4, April 2017

A Survey on Chat Log Investigation Using Text Mining

Khan Sameera, Pinki Vishwakarma

M.E 2nd Year, Dept. of Computer Engineering, Shah & Anchor Kutchhi Engineering College, Mumbai, India

Professor, Dept. of Computer, Shah & Anchor Kutchhi Engineering College, Mumbai, India

ABSTRACT: In this ever growing era of modern technology, it has become essentially necessary for people to communicate with each other. Various medium like social media are used for communicating business, organizational details, etc. Due to the increase in technology, there are chances of performing the crimes in newer ways. These social networking sites, there is a high chance of carrying out criminal activities like phishing, spamming, cyber predation, cyber threatening, blackmail, robbery, killing and drug trafficking etc. has ability to transfer suspicious messages via mobile phones, Instant Messengers and Social Networking Sites, most of the crime related information on the web are in text format, which is difficult to trace their criminal activities dynamically. Detecting and exploring crimes and investigating their relationship with criminals are involved in the analyzing crime process. Criminology is a suitable field for using text mining techniques that shows the high volume and the complexity of relationships between crime datasets. Text mining techniques is an effective way to detect and predict criminal activities. Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. The proposed framework applies n-gram technique in association with a self-customized Hyperlink Induced Topic Search (HITS) algorithm to identify key-terms, key-users and key-sessions.

KEYWORDS: Text Mining, Chat logs Mining, Social Network Analysis (SNA), Social graph generation, Cyber Crime Investigation.

I. INTRODUCTION

In this busy modern world, people have become fond of internet and its technologies. The internet is being used extensively for most of the real life applications such as sending e-mails, distant learning, online searching, and chatting in collaborative environment etc. in the past few years. At present, there are various chat tools as well as chat rooms that are available on Internet. Due to the emergence of such rooms and tools, communications between internet users all over the world have been enriched.

Due to the growth of Internet Technology many legal as well as illegal activities have been increasing. The evolution of Internet has led to the growth of innumerable cybercrimes. Most of the offenders are exploiting the convenience, speed and anonymity of this technology to commit a wide range of activities that knows no borders, either virtual or physical. An important aspect that has emerged because of the break-through of Internet Technology is the misuse of this technology in communicating abduction of young teenagers and children via chat rooms or e-mails that are difficult to monitor. Monitoring such chat rooms will be helpful in the detection of crime and even to some extent crime prevention.

Text mining techniques is an effective way to detect and predict criminal activities [1] [2]. Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. According to new statistics, more than 16,000 criminal activities on social media, including Face book and Twitter have been reported over the last year. It has become very important to monitor online forums in order to detect criminal or more broadly terrorist activities. There is not much research on monitoring chat room conversations for potentially harmful activities. The current monitoring techniques are basically manual which is costly, tedious, difficult, and time consuming. The Proposed system is very good to detect such illegal activities on Social media by using text mining algorithm and also find criminal profiles.

DOI: 10.15680/IJIRCCE.2017. 0504039

II. RELATED WORK

Cybercrime is increase day by day on Social media. One of the most amazing boons of Internet Technology is its ability to connect with various individuals around the globe with the help of various Social Networking Sites. Analyzing these sites for various criminal activities again becomes difficult .Social network analysis deals with analyzing the behaviours of groups, organizations and individuals and determining its behavioural patterns. Social network analysis is becoming a major tool for counter terrorism applications. Social Networking Sites (SNS) are a web-based service which facilitates individuals to construct a profile, which is either public or semi-public. SNS consists of a list of users with whom we can converse, share a connection and also view their activities in a network. SNS users communicate by blogs, messages, chatting with music files and video. SNS plays a very important role in the human life; it has evolved to become the main communication medium among individuals and organizations after Oral and Non Verbal Communication. In social media, the users produce several and various formats of suspicious posts (text, image, video…) and exchange them online with other people. The data in most social media sites are stored in text format, so in this work we will focus only on text posts.

Text mining techniques is an effective way to detect and predict criminal activities. Thus, our project is a very good option to detect such activities since it uses text mining algorithm to continuously check for suspicious words even if they are in the form of code words or short forms [3]. Since many Social networking sites allow information to be publicly available, many criminal cases can be solved by analyzing this publicly available information on social media.

III. PROPOSED SYSTEM

Suspicious messages are sent through Instant Message (IM) and Social Networking Sites (SNS) which are untraced, leading to hindrance for network communications and cyber security. The existing general forensic search tools have some short comes. We proposed a Framework that discover and predict such messages that are sent using SNS are suspicious chat log even if they are in short forms or code word. The proposed framework applies n-gram technique in association with a self-customized Hyperlink Induced Topic Search (HITS) algorithm to identify key-terms, key-users and key-sessions [4]. The main objective is to develop a system to detect and monitor suspicious activity over a network and to find suspicious user profile.

The framework unifies user interaction and conversation data together to identify key information components and different user groups. Fig 1 presents the architecture of the proposed framework for chat log investigation which perform six different task i) data extraction and normalization ii) vocabulary extraction iii) key information extraction iv) social graph construction v) user group identification vi) user profile detection [5] [6]. First of all, the chat logs are processed to identify different information components and normalize them for noise removal and slang neutralization. The second step applies a n-gram technique to extract a vocabulary set of the chat community, which is followed by the extraction of key-information and computation of feature values by constructing two bipartite graphs and applying HITS on them. Thereafter a social graph of users is constructed as a weighted graph using their interaction patterns in the group chat sessions. Finally, the social graph is used to identify user-groups using clustering techniques and generate user's profile. The following are details of tasks.

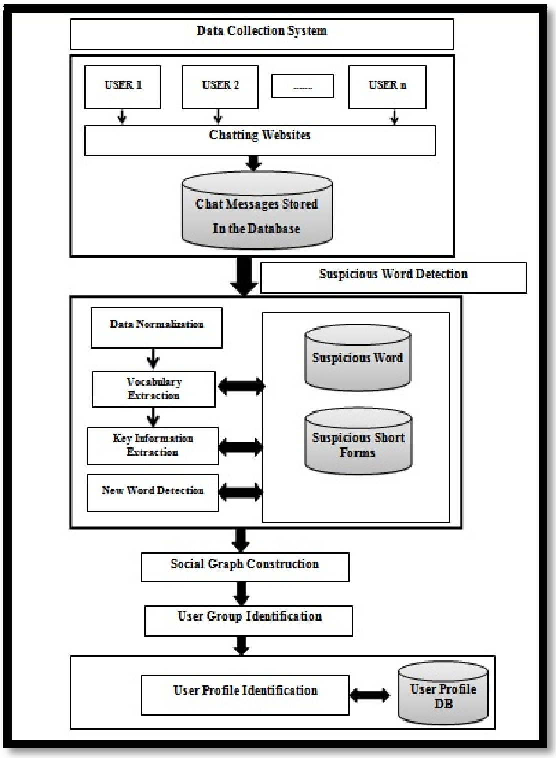

Fig 1: System Architecture

A. Data Extraction and Normalization:

This task aims to transform raw chat logs into a machine-readable format. It consists of three sub-tasks that are described in the following sub-sections.

1) Information Component Extraction:

First step is to extract the data consists of messages, chat from web. After the formation of system the security agency and investigation department can use the data what they are gathering including chat logs. We have extracted a large amount of sample dataset consists of message and chat to perform the text mining and data analysis.

Chat messages are logged in various formats depending on the application platforms and their settings. The data is a collection of chat logs archived through Messenger Plus! which is a third party extension of Windows live messenger. The logs are available as HTML files, each of which contains discussions of one or more chat sessions.

2) Noise Normalization

As we are having a text files containing dirty data, for such type of records first we need to perform cleaning procedure. The most common form of noise in chat messages is the unnecessary repeated use of punctuation marks or letters, stop words and some particular information which is not required during the further checking process. The repeat may occur at any position e start, middle, or end of a word. For example, “okkkk” and “really?????”

3) Slang Normalization

Slang expressions commonly used in chat messages have no booked place in standard dictionaries. Therefore, we compiled a list of slang expressions and their equivalent standard terms from different sources and personal surveys. For example, "⟨f9, fine⟩" replaces each occurrence of "f9" by "fine", and "⟨lol, Laugh out loud⟩" replaces each occurrence of "lol" by "Laugh out loud".

B. Vocabulary Extraction:

A vocabulary is a set of terms or lexicons that completely cover a user's or a community's communicative knowledge. In context of chat communications, we define vocabulary as a set of valuable information containing key terms exchanged among participating users during its complete life of communication. Therefore, the vocabulary extraction process aims to identify vocabulary of the community involved in chat discussions and find suspicious words.

C. Key Information Extraction:

Once we complete extraction process and pre-processing over data. It can be sent for analysis by applying different data mining technique and algorithm as per the information needed. In this study we apply HITS algorithm [4].

D. Social Graph Construction:

A social graph of users is constructed as a weighted graph using their interaction patterns in the group chat sessions. A chat session generally contains a group of users interacting with each other, and such interactions establish a kind of tie or bond between them. The motive behind social graph construction is to model the participating users and their interaction patterns into a rich structure which could represent the ties among the participating users.

E. User-Group Identification:

For a proper investigation, it is very important to identify how the chat users are related among themselves. It is also crucial to identify different closely associated groups among the chat participants. In this project we identify both users' conversation and interaction data in group-chats to discover overlapping users' interests and their social ties.

F. User Profile Identification:

In this block we find suspicious user profile. The first target the extraction and integration of information related to a given profile from the web and the second target the analysis of user's behaviour [7].

IV. CONCLUSION

Social media is becoming an integral part of life online as social websites and Web based communication. E-communication through the chat server, Instant Messaging System, Internet Relay Chat (IRC) are one of the rapid growing communication type but suspicious message are sent through Instant Messengers (IM) and Social Networking Sites (SNS) which are untraced, leading to hindrance for network communications and cyber security.

Text mining techniques is an effective way to detect and predict criminal activities. Thus we have studying different methodologies such as Clique Miner, Ontology Based Information Extraction technique and Hits algorithm with their advantages and disadvantages. Mining data for Chat log Investigation HITS algorithm can be effective for crime investigation because it has high efficiency.

REFERENCES

1. Divya Nasa "Text Mining Techniques- A Survey" International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 4, April 2012.
2. Mahesh T R, Suresh M B, M Vinayababu "Text Mining: Advancements, Challenges and Future Directions" International Journal of Reviews in Computing 2009-2010 IJRIC& LLS.
3. A Lew and H. Mauch ― Intoduction to Data Mining Principles‖ ,SCI, springer, 2006.
4. Kleinberg JM. Authoritative sources in a hyperlinked environment. J ACM 1999; 46(5):604e32.
5. Abulaish M, Anwar T. A web content mining approach for tag cloud generation. In: Proc. of the 13th Int'l Conf. on IIWAS; 2011. p. 52e9.

6. Anwar T, Abulaish M. Web content mining for alias identification: a first step towards suspect tracking. In: Proc. of the IEEE Int'l Conf. on ISI; 2011. p. 195e7.
7. Mark Pollitt, PhD Associate Professor,"The Narratives of Digital Evidence". AAFS 66th Annual Scientific Meeting, Seattle, WA February 17, 2014.

BIOGRAPHY

Khan Sameera Mohammed Ajaz is a M.E 2nd year student in Shah & Anchor Kutchhi Engineering College Chembur, University of Mumbai. Her research interests are Data Mining, Computer Networks etc.