# MALICIOUS DATA MINING FROM CYBER TEXT DATA

[1]M.Bhavana, [2]M.Ashok Kumar, [3]K.Nikhil ,[4]A.Naga Kiran,[5]Mrs. Y.Padma,

*[1, 2,3, 4]IV/IVB.Tech, Department of IT, P. V.P.Siddhartha Institute of Technology, Vijayawada.*

*[5]Asst. Professor, Department of IT, P. V.P.Siddhartha Institute of Technology, A.P, INDIA.*

**Abstract**—*Due to the increase in technology, there are chances of performing the crimes in newer ways. In recent trends, we have seen a tremendous usage of social networking sites. Due to these social networking sites, there is a high chance of carrying out criminal activities like robbery, killing, abetting suicides etc, as these are user protected and has ability to transfer messages and mails among several number of people in the form of mails and documents. The main objective of our project is to analyse such criminal information from mails and documents in order to aid the criminal department investigators. This model helps the investigators by displaying the malicious messages contents of respective users and their word frequencies in order to solve the mysteries in a very short span of time.*

***Keywords:*****Malicious, Crime, Email, Investigation, Data Mining, Forensic.**

## 1. INTRODUCTION

At present days, everyone like professionals, students, professors, teachers including criminals is communicating through internet via emails, social networking sites, messengers etc, Because of this trouble free communication means, criminals are performing many more illegal activates very easily which includes bomb blasts, robbery, fraud drug dealings and many more. In order to find the culprits, forensic experts and investigators often going through various chatting sites to analyse chat data between suspects and to find out the actual culprits. The main concept in our project is to collect chat data between such suspects and to help the crime department investigators by analysing large amount of chat data and find out the hidden malicious data. In our project, not only structured format data, but also unstructured formatted data can also be analysed. The main concept in our project is to collect chat data between such suspects and to help the crime department investigators by analysing large amount of chat data & finding out the hidden malicious data. In our project, not only structured format data, but also unstructured formatted data can also be analysed. The main objective of our project is to assist investigation departments by obtaining the information in advance, which a culprit is transferring through internet based communication. The following process shows the process of how to detect malicious messages between suspects and their count frequencies by performing several actions like pre processing, extraction of pre processed keywords, comparison of that extracted keywords with suspected words loaded in dictionary and finding out the actual culprits.

### Problem Statement:-

In order to investigate, several crime departments are going through internet based information transferring applications like emails & chat messengers. Since, such data will be in large volumes and in unstructured format, finding the actual culprit from suspicious persons is a very challenging and tedious task. The problem can be classified into following:

- Defining that the information as malicious.
- Defining that resultant information associate with specific email or person.

- Detecting the criminals and their impactness.

## 2. DATA ANALYSIS

Generally the chat data between the users will be in unstructured format and is of high volume. Using of artificial intelligent machines is very difficult for such data. So we used data mining techniques and classified that data with respect to certain attributes. We have analysed that data by considering the email addresses of users and their chat content in an excel format. We have used Net beans IDE software to handle that classified data.Net beans IDE uses java language for performing data mining approach. For data pre-processing and extracting, porter stemmer algorithm has been used which easily removes all the clause forms from the chat content and converts that words into its root forms. Stop words has also been deleted in pre-process steps and resultant words has been extracted for further checking. Since we have already stored many suspicious words in the database, we have compared our resultant extracted words with that dictionary words which are finely in structured format. It is observed that the output is very accurate as it is displaying each users email addresses and their chat content along with obtained malicious terms frequencies.

## 3. OUR APPROACHES

### 3.1 Collection of emails and messages information datasets:

Fundamental step is to collect datasets which have large amount of information regarding mails to suspects. We have drawn out a huge amount of sample datasets which contains this information in order to implement the data analysis and data mining.

### 3.2 Creation of database & uploading suspected words into our created database:

Initially we have created a database using my slowed have collected several suspected words from various sources in web. This is a text file which contains all the malicious words in a organized from and it contains words in several formats like jumbled, mixed words etc. We have delivered all these malicious words into our database to detect suspicious users and their respective data and emails information.

### 3.3 Uploading chat datasets into our model:

Data need to be adapted to the software we are using. Data sets which are in MS-Excel format have been uploaded into our model which contains to and from attributes along with the message contents between those suspects. Since we are using java language to implement our model, we have converted our datasets into table format.

### 3.4 Data pre-processing and extracting:

Information that is gathered have to be converted into the form which is to be understandable by the software or language we are using. For that purpose, we are performing several tasks like removal of unwanted texts, symbols, as well as words which are generally not useful for performing text mining .Since text files contains unwanted and inconsistent data, initially there is a need to perform cleaning procedure. Pre-processing and extracting will be done with the help of program code which helps in removing punctuation marks, stop words and some particular information which is not required for further checking process .For this Data Pre-Processing we had used two algorithms which were used to help for cleaning the data .The main algorithms are:

#### 3.4.1 Porter Stemming algorithm:

Stemming is the term used in linguistic morphology and information retrieval to describe the process for reducing inflected (or sometimes derived) words to their word stem, base or root form. Generally a

written word form. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Algorithms for stemming have been studied in computer science since the 1960s. Many search engines treat words with the same stem as synonyms as a kind of query expansion, a process called conflation.

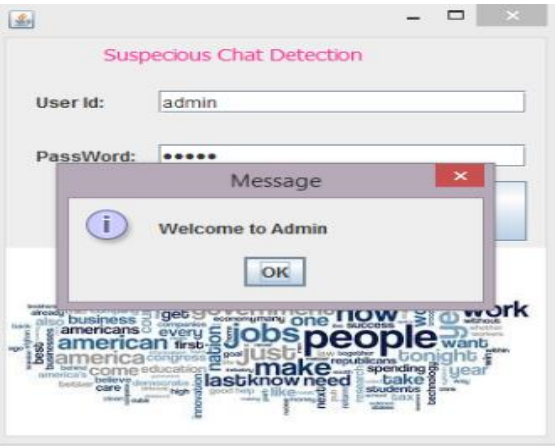
### 3.4.2 Catching the Stop words

Most Search Engines do not consider extremely common words in order to save disk space or to speed up search results. These filtered words are known as 'Stop Words' .These are more generally used words in real life these words are used make the meaning of the conversation .We don't have use of such words keeping in the chat data .So in order to eliminate we use this algorithm for eradication of general words.

## 3.5 Inspecting the data

After the pre-processing and extraction, the resultant data will be scrutinised by applying different mining procedures like comparing the resultant words with the suspected words that are resided in the dictionary and presenting the output. The entire process will be shown in the following sector.

First we have some chat datasets from online and suspected data words to be loaded into database. We have also created some chat datasets on our own for better understanding. Our entire model will be perceived by using following screens:

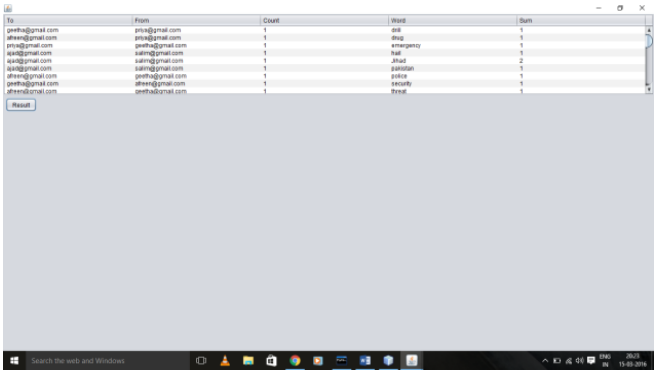


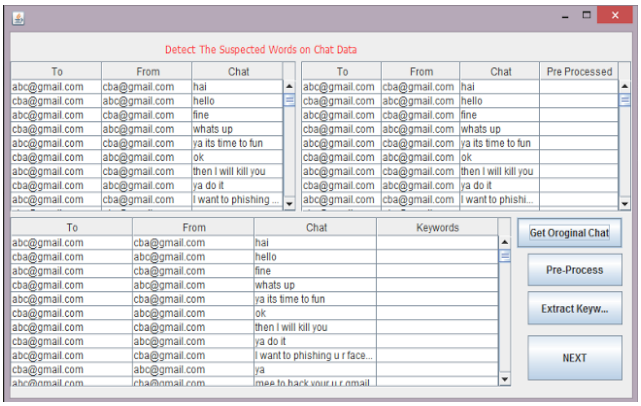

Admin will login into the account in order to perform further actions.

List of suspected words feeding into the database

```
Drill,Domestic security
Exercise,Domestic security
Cops,Domestic security
Law enforcement,Domestic security
Authorities,Domestic security
Disaster assistance,Domestic security
Disaster management,Domestic security
DNDO (Domestic Nuclear Detection Office),Domestic security
National preparedness,Domestic security
Mitigation,Domestic security
Prevention,Domestic security
Response,Domestic security
Recovery,Domestic security
Dirty Bomb,Domestic security
Domestic nuclear detection,Domestic security
Emergency management,Domestic security
Emergency response,Domestic security
First responder,Domestic security
Homeland security,Domestic security
Maritime domain awareness (MDA),Domestic security
National preparedness initiative,Domestic security
Militia,Domestic security
Shooting,Domestic security
Shots fired,Domestic security
Evacuation,Domestic security
Deaths,Domestic security
Hostage,Domestic security
Explosion (explosive),Domestic security
Police,Domestic security
Disaster medical assistance team (DMAT),Domestic security
Organized crime,Domestic security
Gangs,Domestic security
```

Performing data preprocessing steps on the extracted and analyzed data.






Showing results of suspicious users message contents and their respective mail ids.

## 4 .CONCLUSION

In this paper, we have brought a model into existence which can perform term mining on criminal's cyber text data. This model takes suspicious persons emails and chat messages content in an excel table format and performs pre-processing and analysing after extraction. Finally it provides information regarding actual culprit's malicious actions as shown in the above figures which will helps investigation departments to search over specific mails rather than examine all suspicious people's cyber text data.

## 5 .FUTURE WORKS

As our model is only used for data which was collected after a crime scene was done, we need to overcome such large criminal activities by creating a model where mining will be done over online streaming data. Model have to be created where we can place certain limitations to users like, whenever the suspicious word count increases between the users , that respective users will automatically get blocked from sending and receiving mails or chats after certain limited warning alerts. To achieve this, we have to give some malicious words with high priority and if those words which is of highest priority are often gets used, that respective user's details and locations will automatically get transferred to criminal investigation organizations. Remaining emails and users details would get removed from the database so the increasing volume of database can be controlled. With this it would also become easy for forensic department to check only specified mail rather than all.

## REFERENCES

[1]S.Gowri, 2G.S.Anandha Mala, 3G.Divya,"Suspicious Data Mining From Chat And Email Data" 2012/ International Journal of Advances in Science Engineering and Technology.

[2]Net Beans IDE installation procedure https://netbeans.org/downloads

[3]https://dev.mysql.com/downloads/mysql/ MySQL server and command line client command prompt installation.

[4]N. Pendar, "Toward spotting the pedophile telling victim from predator in text chats," IEEE Internet Computing, pp. 235–241, 2007.

[5]Farkhund Iqbal, Benjamin C. M. Fung, Mourad Debbabi "Mining Criminal Networks from Chat Log"

2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[6]Martin Halvey and Mark T. Keane, "An Assessment of Tag Presentation Techniques" poster presentation at WWW 2007, 2007.

[7]Salvatore J. Stolfo and Shlomo Hershkop, "Email Mining Toolkit Supporting Law Enforcement Forensic Analyses" Columbia University. 500 West 120th St. New York, NY 10027.

[8]Vadher, Bhargav, "EMail Data Mining: An Approach to Construct an Organization Position-wise Structure While Performing Email Analysis" (2010).Master's Projects. Paper 63, San Jose State University.