International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

MONITORING SUSPICIOUS DISCUSSIONS ON ONLINE FORUMS USING DATA MINING

Shet Nitish Nagesh[1], Yashaswini[2], Rahul Anil Prabhu[3], Rajatha J Shetty[4]

[1,2,3]Canara Engineering College, Benjanapadavu
[4]Under the guidance of, Sushma M D (Assistant Professor)

Abstract - *People now-a-days are using forums as discussion medium. With the increasing years, the internet has changed the lives of many people for better or worse. As internet technology is progressing, many illegal activities have also increased exponentially. The Internet is an unacknowledged path for illegal activities such as hacking, trafficking, betting, fraud and scams etc. The cyber-crime branches are looking for provisions to detect these forums for illegal feedbacks, comments or reviews for their investigation. Our proposed system will monitor for suspicious postings, collect it from few discussion forums, implement technique of data mining and notify administrator about malicious user. In this concern, we focus on Data Mining and Sentimental Analysis to bring a awareness about such discussion and make user not to use them again.*

Key Words: Illegal Activities, Discussion forums, Sentimental Analysis.

1. INTRODUCTION

In recent days people are addicted to the social media like anything, it has become the part & parcel of our life. And we have started using it as a live platform to express our feelings, opinions, promotions of the current events on any topic. Fraud or misguided people don't leave any space to spread criminal activities & social media is one of the popular medium of them. Data mining & Data analysis is the technique by which we can keep eyes on social media. In this paper the suspicious word will be detected and get converted to *.

1.1 OBJECTIVES

- To Reduce the suspicious Activity on Online forums
- To notify the Admin about the malicious user
- To identify and convert negative words into *

2. Literature Survey

The paper elaborates about Stop-word Selection, Stemming algorithm, Brute-force algorithm, Learning Based algorithm and Matching algorithm. Matching algorithms use two constraints Stemmer Strength and Index Compression. Using these two constraints, stem words in database are compared and their value is calculated. Learning based algorithms include machine learning theories like SVM and conditional random field. This system also focuses on plan execution time, automated classification to identify more significant suspicious discussions. [1]

In this paper author describes about detection of emotion on online media. EmoTxt finds the emotions and categorize based on the input data provided in a comma separated value (CSV) file format. The output is in the form of CSV file. The file contains text id and predicted label for each input data set. The model classifies the emotions as, joy, sad, and anger etc. According to researchers [2], the model follows a tree structured hierarchical classification of emotions, where latter layers provides an understanding of emotions of the previous layers. The model includes six basic emotions, namely love, happy, anger, sad, fear, and surprise. The data is tested and trained on gold standard dataset using linear Support Vector Machine (SVM). [2]

The paper describes, the system will analyse data from few discussion forums and will classify the data into different groups i.e. legal and illegal data using Levenshtein algorithm. Levenshtein is used to measure similarity between two words. [3]

In this paper work, they have used Social Graph generation based approach for the identification of suspicious users and chat logs. Overall process of graph based suspicious activity detection is performed in seven steps. These steps are Generation of instant chat application, Storage of user chat logs, Data extraction from chat logs, Data pre-processing & normalization, Key Information Extraction, Social Graph Generation, Suspicious Group Identification. By using these steps, suspicious activity can be identified. Here, Concept of SVM approach is used for the extraction of key information like key users, key terms and key sessions.

Apriority algorithm is used for the social graph generation and final declaration of suspicious users is performed with decision tree approach. For the evaluation of this concept, user scores and normalized scores have been evaluated and compared for the different suspicious terms like terrorist, fraud, wrong and hack etc. From the evaluated user score and normalized score, we can say that proposed concept of SGTM is efficient for the suspicious session

identification. For future aspects, this concept can be compared based on the further evaluation parameters like accuracy, precision, recall etc. Also the considered concept can be integrated with other methods of classification like n-grams, naive Bayes etc. [4]

The author proposes research is being carried out using web mining. Using Web mining, the data set is collected by crawling large number of web pages. It requires a user interactive query interface intended for predicting crime hotspots from various web pages. The main techniques used are classification, sequential pattern mining, association analysis, outlier analysis and cluster analysis. Clustering and classification techniques identify the similar items and group them in classes. The association rules mining and sequential pattern mining techniques are similar. They both identify frequently occurring sets and extract a pattern. Using all these techniques in web mining make it more complex.

Along with the techniques, a conceptual network i.e. a dynamic structure of nodes connecting in a functional way is required for better visualization of criminal networks and to reveal the vulnerabilities inside the network. The biggest challenge faced by the researchers was collecting the data from the web pages which consist of hyperlinks, navigation links, advertisements, privacy policies etc. Theses noises should be removed from the data before processing. Another challenge was that on web the information is never constant. The model intends to concentrate on efficiency by using multiple processes, threads and asynchronous resources. [5]

3. Proposed Method

Data mining can be used to monitor social media as well as discussion forums for suspicious feedbacks or comments. Discussion forums can be used to spread any message to a large population almost instantly. Millions of people share their views and ideas on politics, religion and there are also people who intentionally hurt religious or racial sentiments through malicious posts. Hence it becomes important to monitor the posts on these forums.

In this paper we make use of collection of data from different online forums. This data is then passed into csv file. On the other part of this method user will be given by his own account and credentials of a website, where he need to logged in and can start a discussion with any topic. But whenever he /she make use of such words will be notified to admin of the particular site. And even user will be warned on his activity.

The technique we make use of data mining algorithm Naïve-Bayes theorem which is implemented in python library textblob. As Django is our frontend frame work helps the task in simple way. The algorithm used here will analyse the words into positive and negative classification based on polarity and subjectivity parameters. Fig 3.1 explains about Architectural design to monitor suspicious discussion.

Fig 3.1 Negative opinion converting into *

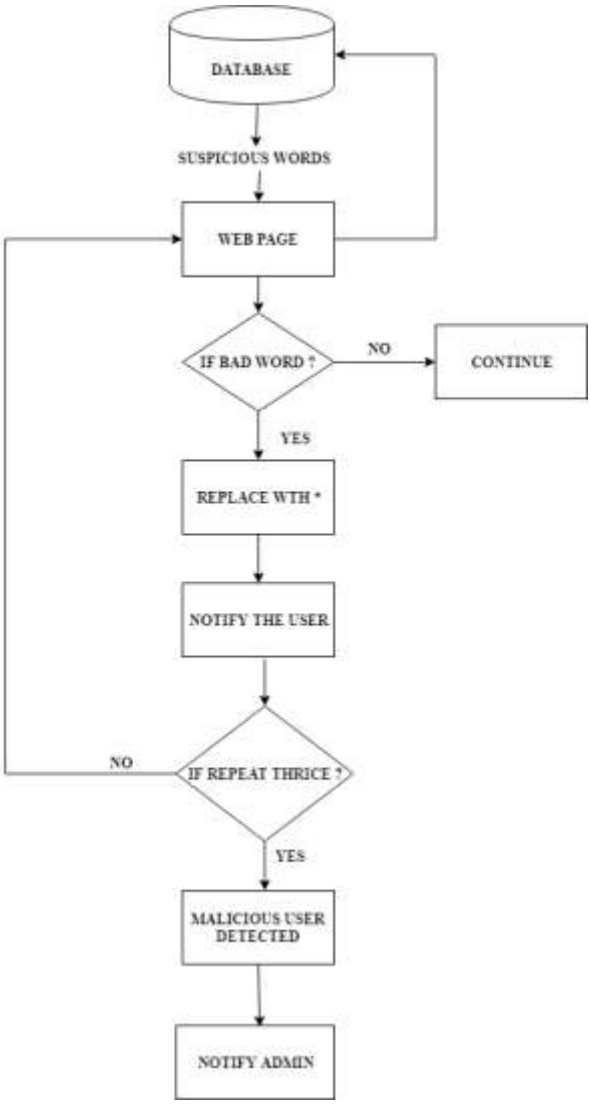

Fig 1.1 Architectural Design to monitor suspicious discussions

4. CONCLUSION

The main objective was to monitor the suspicious activity that occurs in various online forums. This application satisfies with our objectives. From the time of user login and his discussion on any topic available in online forum are monitored. Once the suspicious word is found it is replaced by the * and is notified to website administrator.

REFERENCES

[1] Murugesan, M. Sururthi, R. Pavitha Devi, S. Deepthi, V. Shri Lavanya, and Annie Princy. Automated Monitoring Suspicious Discussions on Online Formus Using Data Mining Statistical Corpus Based Approach. Imperial Journal of Interdisiplinary Research (IJIR) Vol2, Issue-5, 2016

[2] Javad Hosseinkhani, Mohammad Koochakazei, Solmaaz Keikhaee and Yahaya Hamedi Amin. Detecting Suspicion Information on Web Crime Using Crime Data Mining Techniques. International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol.-3, No. 1, 2014, Page 32-41

[3] Harika Upgaganlawar, Nilesh Sambhe. Surveillance of Suspicious Discussions on Online Forums Using Text Mining. International Journal of Advances in Electronics and Computer Science, Volume4, Issue-4, April-2017

[4] Social Graph Based Suspicious Chat Log Identification Using Apriori Algorithm and

Support Vector Machine Amit Verma1, Sonali Gupta2, Rahul Butail 3, 1Head of the Computer Science Department, 2Assistant Professor, 3Rahul Butail 1, 2, 3 Chandigarh Engineering College, Landran, Punjab, India.

[5] G.Vinodhini, R.M Chandrasekran. SentimentAnalysis and Opinion Mining: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering. Volume 2, Issue 6, June-2012

BIOGRAPHIES

Shet Nitish Nagesh

Department of Computer Science and Engg, Canara Engineering College, Benjanapadavu

Yashaswini

Department of Computer Science and Enggg, Canara Engineering College,Benjanapadavu

Rahul Anil Prabhu

Department of Computer Science and Engg, Canara Engineering College,Benjanapadavu

Rajatha J Shetty

Department of Computer Science and Engg,Canara Engineering College,Benjanapadavu