iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** VI **Month of publication:** June 2017

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Social Graph Based Suspicious Chat Log Identification Using Apriori Algorithm and Support Vector Machine

Amit Verma[1], Sonali Gupta[2], Rahul Butail[3]
[1]Head of the Computer Science Department, [2]Assistant Professor, [3]Rahul Butail
[1, 2, 3] Chandigarh Engineering College, Landran, Punjab, India

Abstract: ***With the increasing use of Instant Chat Messengers to share information, suspicious activities has also increased. There are many sources to share the information but instant chat messengers and social networking websites are the quick and easy means to share anything. Sometimes, even news stories are initially broken up on social media sites and further on chat messengers instead of any news channel and newspaper etc. Due to these technology advancements, some people are misusing these instant chat messengers to share suspicious activities and make plans to do something suspicious. This kind of chat is mainly available in textual format. In this paper, a social graph based concept is used to identify suspicious terms, chat sessions and users. Here, users are considered as nodes and the relation of user with any chat log is considered as edge of the graph. In this process, Support Vector Machine is used to identify & classify the suspicious key terms. Apriori algorithm is used for the social graph generation. Suspiciousness of any chat group can be predicted based on the support & confidence level of Apriori algorithm. Experiment results are evaluated in order to identify suspicious key terms, key users and key sessions. Also the weightage of key terms, key user score and normalized score has been evaluated. The declaration of user as suspicious or authentic is based on weightage that is evaluated using decision tree approach.***

Keywords— Decision Tree Algorithm, Social Graph generation, Support Vector Machine, Apriori Algorithm, Suspicious Activity

I. INTRODUCTION

Now-a-days, internet users are addicted to instant chat messengers. There is the availability of thousands of chat messengers through which more than trillion of messages are shared by users. This evolution in internet and instant chat messengers led to expansion of cyber crimes. Cyber crime departments are continuously working on the approaches to detect these growing suspicious activities through instant chat messengers [1]. The available concepts/tools with cyber departments are only limited to detect malicious content having malicious URL links. So, there is need to enhance existing systems or development of some novel approaches to stop the growth of cyber crime activities. In this paper, the existing concepts for suspicious behaviour detection are presented in section II. By considering the key features and eliminating drawbacks of existing concepts, a social graph based approach is adapted by us for the identification of suspicious profiles. Here, concepts of SVM [2] and Apriori algorithm [3] have been used for key terms identification and social graph generation respectively. To identify the suspiciousness of any chat conversation, Apriori algorithm based support & confidence terms are evaluated. Further, weightage of key terms is determined and the decision tree [4] is structured for the different key terms. Finally, suspicious user is declared based on the decision tree prediction. For the evaluation of this concept, we have also generated a real time chat conversation application. In this chat application, multiple users can chat and discuss about their interest. Here, suspicious keywords are pre-installed (trained) in the application. So, if someone would use any kind of suspicious words, then it can be identified on the spot. Overall process of graph based suspicious activity detection is performed in seven steps as mentioned: (1) Generation of instant chat application, (2) Storage of user chat logs, (3) Data extraction from chat logs, (4) Data pre-processing & normalization, (5) Key Information Extraction, (6) Social Graph Generation, (7) Suspicious Group Identification. By using these steps, suspicious activity can be identified. Chat logs' data is mainly available in textual format. But people generally chat in some informal manner on chat messengers. Due to this informal manner of textual chat data, it is not easy to use this data for textual mining process. Due to absence of language & grammatical rules, textual mining process has to face various challenges [5] [6]. Some key challenges that abrupt the textual mining process are emotions, inter-wined communication threads, noise, slangs and Multilinguality etc. Emotions are smiley symbols composed of mainly punctuations marks. They are generally used to represent the mood of the user in the form of some facial expressions. But in case of textual mining, it is difficult to identify. Inter-wined communication thread is occurred in communication process when one user starts

communication on some other topic without closing the current/earlier conversation thread. Noise is the use of extra characters while writing. As users on instant messengers are habitual for the instant reply, so they respond in very casual manner without any verification of grammatical or spelling of text. Generally, used noise messages by users are “g888888”, “f999999” etc. Here, extra 8 and 9 are considered as noise during the textual information extraction. The actual keywords for these “g888888” and “f999999” are ‘great’ and ‘fine’ respectively. Slangs are the quick expressions that are used to convey some messages with lesser typed characters. Some commonly used slangs in chat conversation are sry, ROFL, k, L8, TY etc., where, ‘sry’ represents ‘sorry’, ‘ROFL’ represents ‘Rolling On Floor Laughing’, ‘k’ represents ‘okay’, ‘L8’ represents ‘Late’ and ‘TY’ represents ‘Thank You’. Another challenge with textual mining of textual data is the use of multilinguality. Multiliguality is the use of multiple languages in a chat conversation. Sometimes, the use of common alphabets in different languages can interrupt the textual information extraction process. So, these are the common challenges that may someone have to face with, while textual data mining. The organization of the paper is structured in the following manner. Current section has described about some instant chat messengers, brief about the proposed concept and challenges for the mining of informal textual chat data. Further, section II presents the work replated to identification of suspicious behaviour in chat messengers. Section III brief about the basic of Apriori algorithm, SVM approach and decision tree algorithm. Section IV presents the Social Graph based Text Mining (SGTM) approach. Section V deals with evaluated results and comparative analysis of different suspicious terms. And Section VI presents the conclusion of research paper along with some future directions.

II. RELATED WORK

This section deals with the brief review of existing research work that pertain the concepts of suspicious chat logs identification. The authors in the considered work have shown the suspicious chat logs detection process in both the chat logs and social networking websites.

A. Anwar et al. [7] have proposed a social graph based approach for the detection of user interests as suspicious or authentic. In their proposed framework, they made use of the n gram technique and with the help of HITS (hyperlink induced topic search) algorithm, recognized all the key words which were giving us the user’s interests in the conversations. The graph was generated, in which they made use of the self generated concept of ties (edges) between the pair of users (nodes) were established. Authors have considered three cyber crime investigation scenarios and each having their own very view of user group identification. Authors have used 1100 chat logs of the 11,143 chat sessions and collected from a single computer, but if the chat logs were to be collected from multiple computers then the graph could have been more detailed and enriched. Overall results were evaluated in terms of graph depicting the user interests.

B. Ali et al. [8] have proposed a Suspicious Pattern Detection (SPD) algorithm for the identification of suspected cyber threat in instant chat messenger available on Social Networking Websites and Instant Messengers. The proposed framework has considered the Ontology based Information Extraction technique (OBIE) with a pre-defined knowledge base data mining approach of Association Rule Mining (ARM). The proposed concept involves three major steps as mentioned: (a) word extraction from unstructured text (b) e-crime monitoring system program (c) SPD algorithm. The proposed concept has been tested for the Global Terrorist Database (GTD). The proposed concept has been compared with other Instant Messengers, Mobile Phone Apps and Social Networking Sites based on the ability to detect suspicious information during online chats. As per considered parameters, proposed concept shows efficient results.

C. Murugesan et al. [9] have used statistical corpus based data mining approach for the detection of suspicious activities on online forums. Authors have presented the work on textual data of online forums. Authors have used the keyword spotting techniques, leaning based method and hybrid of defined approaches for the overall recognition of suspicious human activity.

D. Badea et al. [10] proposed a tool to determine the chat logs. The concepts of Text Mining and Time series analysis have been used to develop the tools. The tools were developed through checking the chat logs, which had particular words repeating, and in what time, they are occurring and how repetitive the frequency of the words is taking place. Then these co-relations between how Rhythmic behavior all these were developed through the time series Model.

E. Mostafa [11] focused on collecting text data logs from social media website (twitter) and checking the responses of users to various famous brands like DHL, KLM, IBM, T-Mobile and Nokia. The research highlighted that users had positive feedback of the certain big known brands due to their well known name and the trust people showcased on that brand. In his work mining of sentiments was done which gave a true and positive feedback of the customers to the brand. This study showed how brands can quantitatively and qualitatively monitor their brands success.

F. Tayal et al. [12] have proposed the approach of crime detection and criminal identification (CDCI) using data mining approach. Authors have used the Indian dataset if criminal acts like Delhi rape cases, national crime records, committee to protest journalists, crime alerts etc. for the period of 2002-2012. Further, data has distributed into 35 major categories with their attributes. The results have been evaluated for the seven major Indian cities as mentioned: Bangalore, Hyderabad, Jaipur, Pune, Mumbai, Kolkata and Delhi. For the simulation of results authors have used Java based Netbeans platform for the identification and prediction of crime and criminal activities. Also, WEKA tool has been used for the verification of crime activities. KNN approach has been considered for the criminal identification and Google maps have been embedded to enhance the k-means clustering.

G. Benveniste et al. [13] proposed the work which was based on the concept of text analysis and data mining tools to search out the main topics of texts, chat conversations and also web posting, all are in high demand due to the rapid increase in the web information worldwide. The steps followed by them to get the work done is on the lines of (FVS) feature vector selection in the first step and liner Discriminant Analysis as the second step to categorize the raw unstructured text documentations. When the output is out it gets compared with the help of latent semantic analysis (LSA)which is done for the applications in which text categorization is taking place.

H. Haythornthwaite and Gruzd [14] focused on the mining of the text of communal conversations. During this work certain attributes were drawn like that certain words were repeating for a particular community and all were based on the noun based level, like certain communities make use of more peculiar types of words. They found that certain level of words like “agree”, “thanks” and other smiles were more of use and other “hateful” words like “disagree’ were less of use.

I. Setiawan [15] focused on collecting the users reviews and text mining them to see which brand is popular among the customers. This research work was focused on checking the levels of how big companies are branding themselves in front of the customers and how an ordinary man evaluates which telephone brand operator should he trust .The data or the text which was to be mined was from the micro blogging website twitter.

J. Cheney [16] Work focused on the aspect of how the social media his transforming how the news is traveling in the World Wide Web. The content posted on these social media websites needed to be mined and examined as to hoe people like such peculiar type of information as compared to the traditional form of newspapers. His work also tells how people are more inclined to knowing information through the social media as compared to the old methods.

K. Jiang et al. [17] have introduced the novel algorithm of CrossSpot to determine the suspicious information and fraud deviations. uthors have used the metric based approach to define the suspiciousness of a block of information from multimodel data. Twitter based Hashtag hijacking dataset has been used for experimentation.

III. BASIC CONCEPTS

This section provides the overview of the basic concepts like Decision tree algorithm, Apriori Algorithm and Support Vector Machine. These three concepts are used for the generation of social graph depicting suspicious chat logs.

A. Apriori Algorithm

In 1994, Agarwal and Srikankhave developed the Apriori algorithm for the data mining processes. Initially, authors have successfully implemented the concept for market basket transactions and generated the frequent itemsets [18]. Frequent itemsets are those itemsets that satisfy the support threshold values as defined by user. As per the principles of Apriori algorithm, an itemset can be considered as frequent if it’s all subsets will also be frequent. In this algorithm, one by one candidates of subset are extended in the frequent manner. Each step is known as candidate solution. If there would not be any availability of further successful extensions, algorithm terminates automatically. To efficiently count the candidate items, Hash tree structure and breadth-first search are follows by Apriori algorithm. It was only the Apriori algorithm in mining that initially uses the support based pruning to reduce the exponential growth of candidates itself. To determine the frequency of items occurred in the context are defined by Confidence. So, confidence can be defined as how frequent data items in the context. In data mining process, Apriori algorithm uses Boolean association rules for the frequent itemsets. Here, prior knowledge of frequent data items is used by following the bottom up approach. This employs in the iteration manner. To generate the association rules of Apriori algorithm, there should be minimum support and confidence threshold.

B. Support Vector Machine

Support Vector Machine is a supervised learning approach. In 1963, SVM algorithm was initially introduced by Vapnik & Chervonenkis. In 1993, Cortex and Vapnik proposed the first Support Vector Network which was published in 1995 [19]. SVM

algorithm bears a property that it can simultaneously maximize the geometric margin and minimize the empirical classification errors. Due to this property, SVM algorithm is also known as maximum margin classifier. These successful properties of SVM algorithm in machine learning extend its work to data mining approaches. SVM is not much capable to perform the accurate classification in case of large data but it performs well for the fine quality of organized data. In SVM, proper selection of parameters is most important. However, improper selecting of SVM parameters usually leads to very poor generalization capabilities. Searching the optimal SVM parameters is decisive for achieving exceptional performance. In this research work, SVM algorithm is used to extract the key feature information includes the classification of suspicious key users, key terms and key sessions from the authentic one's.

C. Decision Tree Algorithm

Decision Tree Algorithm is also a supervised learning approach that can be used for both continuous and categorical variables. It is an inductive inference approach widely accepted for the classification of patterns specially the knowledge representation in hierarchical manner [20]. Decision tree approach classifies the patter in sorting manner of tree structure from root to leaf nodes. The commonly known examples of decision tree are ID3 and C4.5 [21]. The key strengths of decision tree algorithm are its ability to decide most important classification & prediction fields, ability to handle categorical & numerical attributes and ability to generate decision rules. The basic algorithm for the set of S rules is represented here.

1) Cleave the set S based on rules into feature values of f_1, f_2, …f_n.
2) Check the outcomes of cleaved set S.
3) If every detachment belongs to same class, then assign the same class to each leaf node.
4) If detachment found to be different, then recursively cleave those detachments.

The outcomes of this algorithm will be a decision tree having test nodes and leaf nodes with class labels. Here, Decision tree approach is used for the final declaration of user as suspicious or authentic.

IV. SOCIAL GRAPH BASED TEXT MINING (SGTM)

This section presents the proposed concept Social Graph based Text Mining (SGTM) concept for the detection of suspicious chat log. SGTM process is includes the seven steps as mentioned: (1) Generation of instant chat application, (2) Storage of user chat logs, (3) Data extraction from chat logs, (4) Data pre-processing & normalization, (5) Key Information Extraction, (6) Social Graph Generation, (7) Suspicious Group Identification. The step wise explanation of these seven steps is as below:

A. STEP 1: GENERATION OF INSTANT CHAT MESSENGER APPLICATION

Initially, a java based instant chat messenger is developed. For this, net beans IDE 8.1 version is used with the functions of Java Swing. Instant chat messenger includes the generation of class thread, server module for class server and using standard API's and TCP/IP protocols.

B. Step 2: Storage of User Chat Logs

As the users start using chat messenger then the data got stored on the server. Chat messenger includes the data in textual format. This dataset is further used for the experimentation.

C. Step 3: Data Extraction from Chat Logs

As the Chat log's data mainly in textual form, so there is need to convert this raw into machine readable data in XML format so that it can be further used for the analysis. This extracted information also includes the chat log credentials like their chat id, password, start time of chat and end time of chat sessions etc.

D. Step 4: Data Pre-processing & Normalization

As there is no particular linguistic rules for the chat session, so analyser's faced the common challenges posed by the noisy and informal nature of textual conversation data. Before starting any further experimentation, there is need to perform some pre-processing steps and need to normalize the dataset. It includes the steps of Segmentation, Tokenization, Noise Normalization, Normalization of slangs, removal of numeric data, stop words removal and stemming. These sub-steps are discussed below:

1) Segmentation of dataset is performed if there would be availability of long text data in paragraph or lines. So, segmentation is performed using the separators like '.' ',', '?' etc. to segment the long paragraphs.
2) This segmented data is further tokenized and considered each data word as a separate token.

3) This tokenized data is normalized from the noise values. Noise is the use of extra characters in tokenized words. For example: 'g8888888', here extra 8 is removed and data is normalized from the slang word 'g8' which further represents 'great'.
4) After normalization of noise level, token words are normalized from slangs words. Slangs like 'g8', 'f9', 'sry' are normalized with words 'great', 'fine', 'sorry' respectively. In this way, slangs words are normalized by replacing with their respective complete words. For experimentation, a pre-defined list to normalize these slangs is considered.
5) After the normalization of noise and slangs, if there would be any other numeric character, then that will be removed in this step.
6) Further, Stop Words are removed from the data obtained in the previous step. Stop words are non linguistic words that can be analysed by calculating the frequency count. There is also a predefined list of Stop Words to perform this step.
7) The last step of pre-processing & normalization is stemming. Stemming is the process to obtain the root word from their extended different form with suffix or prefix. For example, 'suffer' can be available in other form like suffered, suffering etc. So, stemming process converts these extended forms into their original root/base word (from suffering/suffered to suffer). This is generated represented with Boolean logic.

If Boolean = true

then no Stemming required.

Else

Root Word=get Root Word (Word to be Stemmed)

Also, check if Suffix exits in the Suffix Array based previous word.

Replace Suffix (Word to be Stemmed Suffix)

Else

Go to next word.

E. Step 5: Key Information Extraction

This step involves the extraction of key information by using Support Vector Machine. Key information includes the expressions of key users, key terms and key sessions. In normalized chat sessions, each chat word has some specific level of prominence or implication in the whole chat discussion. By using SVM algorithm, this step extracts feature values for all terms existing in the extracted normalized data to characterize their prominence in the whole chat discussion. Apply SVM for the Key feature extraction as shown in equation below:

$$SVM = SVMtrain\ (Suspicious\_Key\_terms, Suspicious\_Key\_Sessions) \quad \text{…Equation (1)}$$

Where,

SVMtrain is the SVM training function.

Suspicious_Key_terms maintains the values of words which are detected as suspicious

Suspicious_Key_Sessions maintains the corresponding suspicious key sessions.

F. Step 6: Social Graph Generation

After the extraction of key information and weightage of suspicious key words in each session, Apriori algorithm is used to depict the bipartite graph generation. Suspiciousness of detected chat groups from the previous steps are further predicted based on the support & confidence level of Apriori algorithm. In bipartite graph [22], the vertices are disjoint in two subsets and edges are defined to show the existence of relationship of any one entity with another. The expression for two different disjoint subsets of X and Y with relationship edges is shown in equation below

$$G = (X, Y, E) \quad \text{…Equation (2)}$$

As per the Apriori algorithm, the support of one entity for another entity is evaluated, and then final declaration of suspicious user is based on confidence level of relation of one entity with another and showing their suspiciousness.

G. Step 7: Suspicious Group Identification

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Final identification of suspicious user is declared based on the weightage of key user relation with suspicious key terms using decision tree algorithm. Here, we are considering the scenario that we are having the exact number of user's chat group information. Decisions of decision tree are based on the below defined decision tree rules for the set of 'S' Sessions having 'n' number of users (U_1, U_2, …U_n) are as described

1) If S contains no suspicious user, then decision tree will be leaf.
2) If S contains homogenous users, then corresponding tree will be a deaf that identifies class U_h
3) If S contains heterogeneous Users, then test case will be chosen based on the corresponding required outcome.

In this way, final users will be declaring as suspicious with respective chat session by using the above mentioned steps. The complete process of suspicious user and chat session identification is explained with DFD level 0, DFD level 1 and DFD level 2 in figure 1 to figure 3 respectively.As shown in figure 1, level 0 DFD diagram has considered the end components as User and Analyzer/Admin. Here, user uses the Instant Chat Messenger (IM) Application to share their reviews with other users by creating groups and individual chat. On the other hand, application developed by Admin is used by user and admin analyses the information shared by user is suspicious or genuine. For this, Social Graph based text mining approach is used and finally suspicious user and chat sessions are identified. Figure 2 shows the DFD level 1. In level 1 DFD, information is further elaborated. User initially uses the developed chat application and their textual data is stored by admin. This data is further extracted, pre-processed and normalized so that it can be further used for text mining. As discussed in proposed concept of SGTM, concepts of SVM, Apriori algorithm and Decision tree approach are used for the final identification of suspicious users & chat sessions.Figure 3 shows the DFD level 2. Complete process of SGTM is explained here in DFD level 2. The basic concept explained in DFD level 0, 1 and 2 are same, just the change in level of explanation. In DFD level 2 all steps of SGTM are mentioned. The explained 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.

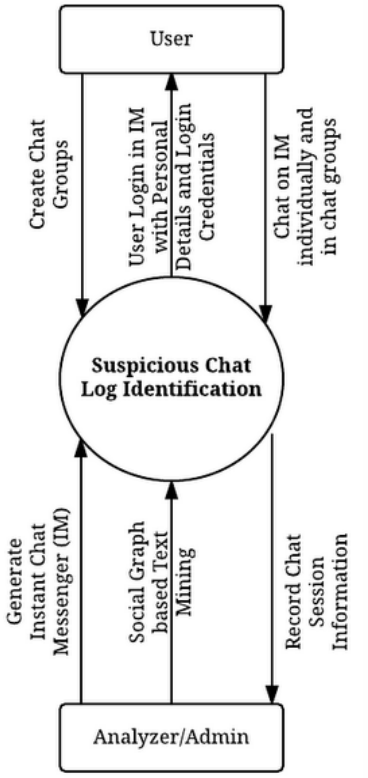


Figure 1: Level 0 DFD for Suspicious Chat Logs Identification

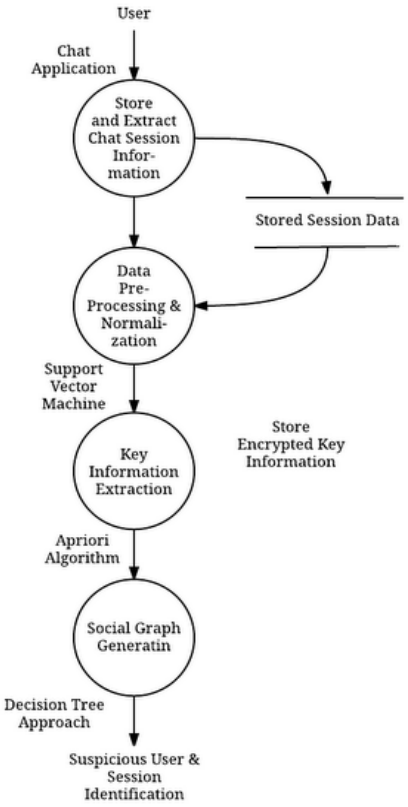


Figure 2: Level 1 DFD for Suspicious Chat Logs Identification

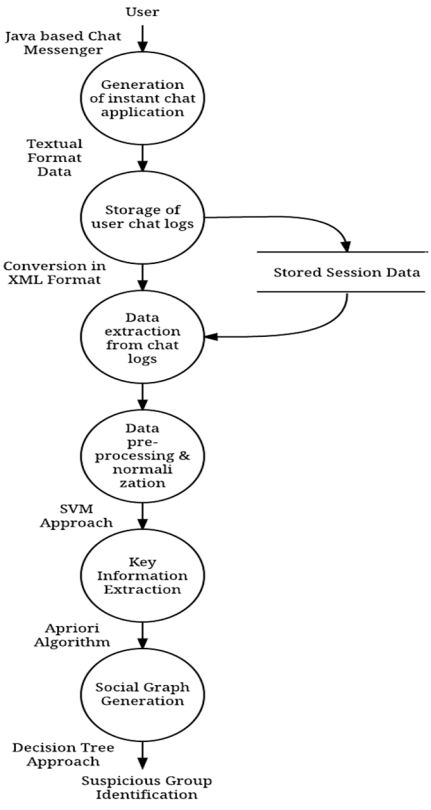


Figure 3: Level 2 DFD for Suspicious Chat Logs Identification

Volume 5 Issue VI, June 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

V. RESULTS AND DISCUSSION

Here the evaluated results and discussion is made for the graph based suspicious chat logs identification. For the implementation of SGTM, windows based system having 6GB of RAM, Intel (R) Core (TM) i5 CPU and Java based Net Beans IDE 8.1 platform have been used. The SGTM concept is evaluated based on the user score and normalized score for different users and chat session with different keywords. In this research work, users have used some suspicious terms of different level like terrorist, fraud, wrong and hack. For these terms, SGTM approach evaluated the user score and normalized score. The evaluated results are shown below in table 1.

Table I
User Score and Normalized Score for Suspicious Terms

Suspicious Terms	User Score	Normalized Score
Terrorist	0.25	3.75
Fraud	0.125	1.75
Wrong	0.075	1.25
Hack	0.175	2.75

This information of evaluated user score and normalized score is further presented in figure 4 and figure 5.

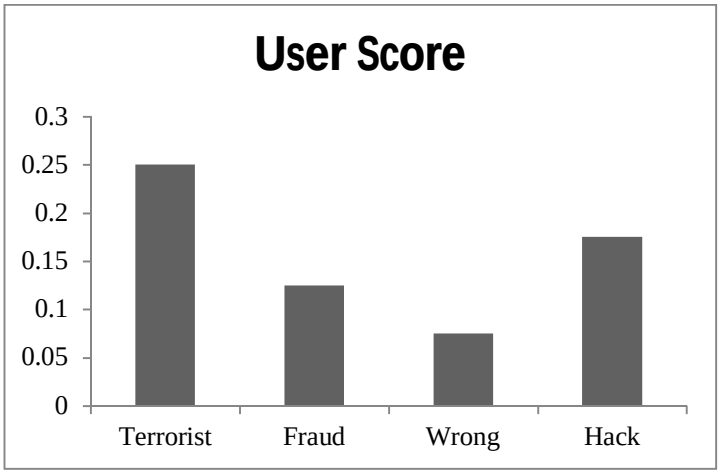


Figure 4: Comparison of Different Terms based on User Score

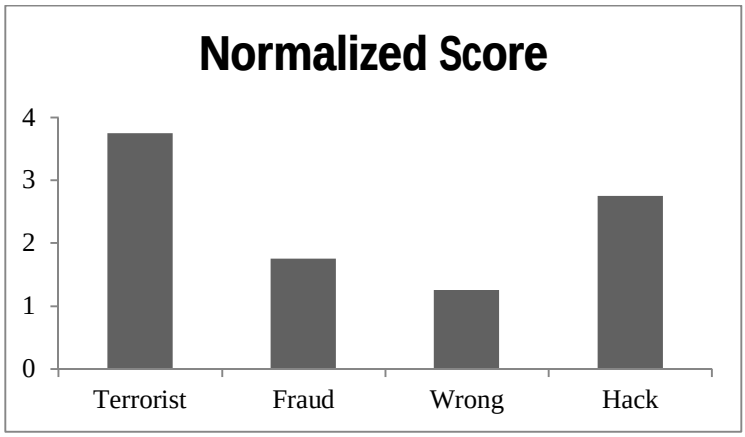


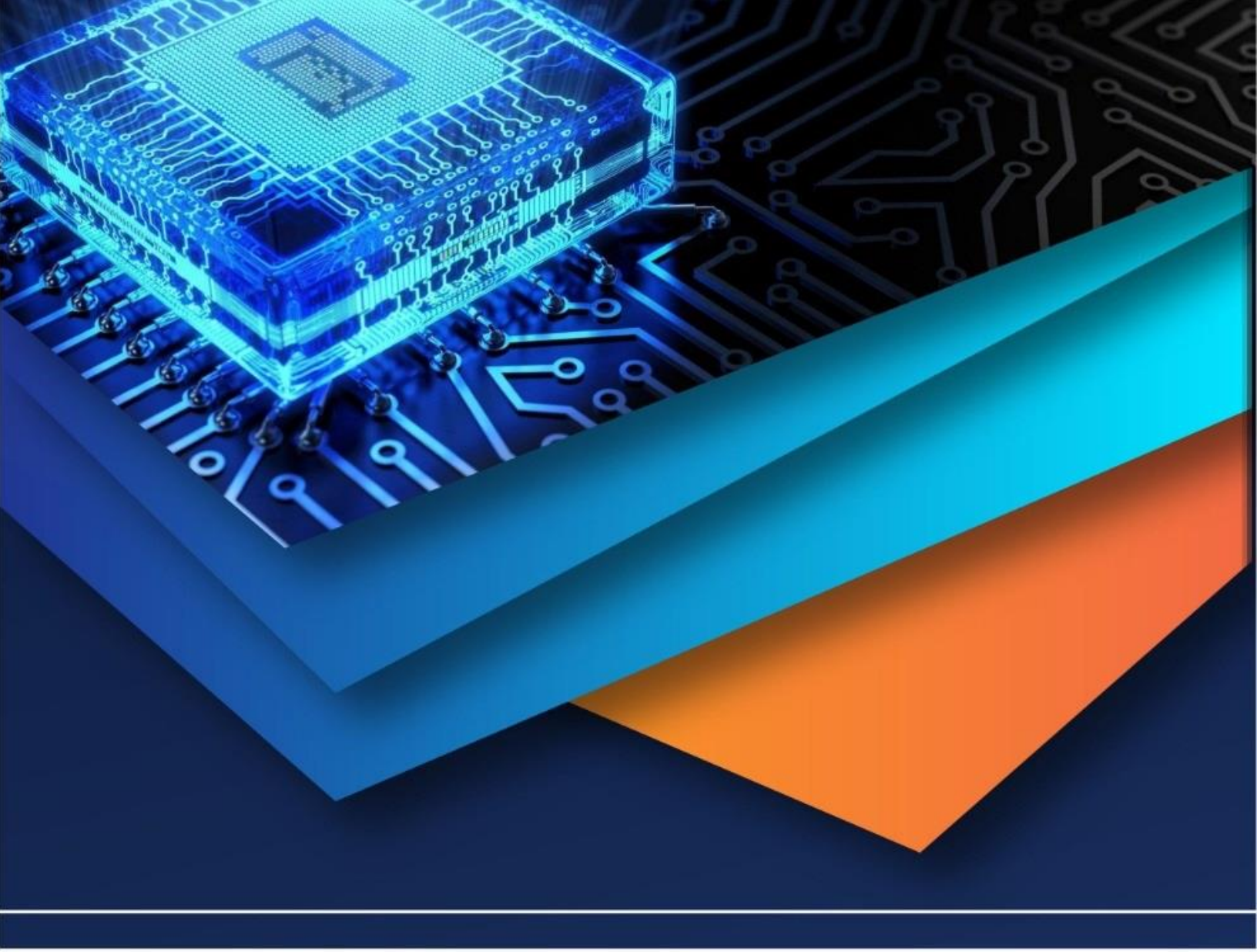
Figure 5: Comparison of Different Terms based on Normalized Score

VI. CONCLUSION

In this research work, we 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. Apriori 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.

REFERENCES

[1] Ali, Mohd Mahmood, and Lakshmi Rajamani. "APD: ARM deceptive phishing detector system phishing detection in instant messengers using data mining approach." Global Trends in Computing and Communication Systems (2012): 490-502.
[2] Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text Classification. " Journal of machine learning research2, no. Nov (2001): 45-66.
[3] Inokuchi, Akihiro, Takashi Washio, and Hiroshi Motoda. "An apriori-based algorithm for mining frequent substructures from graph data." Principles of Data Mining and Knowledge Discovery (2000): 13-23.
[4] Rokach, Lior, and Oded Maimon. Data mining with decision trees: theory and applications. World scientific, 2014.
[5] Agarwal, Sumeet, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. "How much noise is too much: A study in automatic text classification." In Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on, pp. 3-12. IEEE, 2007.
[6] Aw, Ai Ti, and Lian Hau Lee. "Personalized normalization for a multilingual chat system." In Proceedings of the ACL 2012 System Demonstrations, pp. 31-36. Association for Computational Linguistics, 2012.
[7] Anwar, Tarique, and Muhammad Abulaish. "A social graph based text mining framework for chat log investigation." Digital Investigation 11, no. 4 (2014): 349-362.
[8] Ali, Mohammed Mahmood, Khaja Moizuddin Mohammed, and Lakshmi Rajamani. "Framework for surveillance of instant messages in instant messengers and social neworking sites using data mining and ontology." In Students' Technology Symposium (TechSym), 2014 IEEE, pp. 297-302. IEEE, 2014.
[9] Murugesan, M. Suruthi, R. Pavitha Devi, S. Deepthi, V. Sri Lavanya, and Annie Princy. "Automated Monitoring Suspicious Discussions on Online Forums Using Data Mining Statistical Corpus Based Approach." Imperial Journal of Interdisciplinary Research 2, no. 5 (2016)
[10] Badea, Iulia, and Stefan Trausan-Matu. "CSCL chats' analysis using time series." In RoEduNet Conference 13th Edition: Networking in Education and Research Joint Event RENAM 8th Conference, 2014, pp. 1-3. IEEE, 2014.
[11] Mostafa, Mohamed M. "More than words: Social networks' text mining for consumer brand sentiments." Expert Systems with Applications 40, no. 10 (2013): 4241-4251.
[12] Tayal, Devendra Kumar, Arti Jain, Surbhi Arora, Surbhi Agarwal, Tushar Gupta, and Nikhil Tyagi. "Crime detection and criminal identification in India using data mining techniques." AI & society 30, no. 1 (2015): 117-127.
[13] Benveniste, Steven M., and Monique P. Fargues. "Automated text content identification for document processing using a kernel-based support Vector Selection approach." In Signals, Systems and Computers, 2009 Conference Record of the Forty-Third Asilomar Conference on, pp. 366-370. IEEE, 2009.
[14] Haythornthwaite, Caroline, and Anatoliy Gruzd. "A noun phrase analysis tool for mining online community conversations." In Communities and Technologies 2007, pp. 67-86. Springer London, 2007.
[15] Setiawan,johan Using Text Mining to Analyze Mobile Phone Provider Service Quality (Case Study: Social Media Twitter) International Journal of Machine Learning and Computing 4, no.1 (2014):106.
[16] Cheney,Debora. "Text mining newspapers and news content: new trends and research Methodologies."(2013).
[17] Jiang, Meng, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, and Christos Faloutsos. "Spotting Suspicious Behaviors in Multimodal Data: A General Metric and Algorithms." IEEE Transactions on Knowledge and Data Engineering 28, no. 8 (2016): 2187-2200.
[18] Perego, Raffaele, Salvatore Orlando, and P. Palmerini. "Enhancing the apriori algorithm for frequent set counting." In International Conference on Data Warehousing and Knowledge Discovery, pp. 71-82. Springer Berlin Heidelberg, 2001.
[19] Cortes, Corinna, and Vladimir Vapnik. "Support-vector networks." Machine learning 20, no. 3 (1995): 273-297.
[20] Freund, Yoav, and Llew Mason. "The alternating decision tree learning algorithm." In icml, vol. 99, pp. 124-133. 1999.
[21] Jin, Chen, Luo De-Lin, and Mu Fen-Xiang. "An improved ID3 decision tree algorithm." In Computer Science & Education, 2009. ICCSE'09. 4th International Conference on, pp. 127-130. IEEE, 2009.
[22] Zha, Hongyuan, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. "Bipartite graph partitioning and data clustering." In Proceedings of the tenth international conference on Information and knowledge management, pp. 25-32. ACM, 2001.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY