International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

Identification of Suspicious Activities in Chat Logs using Support Vector Machine and Optimization with Genetic Algorithm

Amandeep Singh Khangura[1], Maninderpal Singh Dhaliwal[2], Mansi Sehgal[3]
[1,2,3]*Assistant professor, Department of Computer Science, Ludhiana College of Engineering & Technology, Katani Kalan, India.*

Abstract: ***In the era of modern technology, the advancements in communication technology are leading to riveting trends in daily lives through instant messengers, social networking websites and many other popular communication technologies. Unfortunately, with this advancement the misuse of technology has also been proliferated which leads to the increase in suspicious activities. Some people misuse the technology to spread violence, share criminal activity, bullying other people and thereby enhances the suspicious content on the internet. The communication may be available in the form of text, audio or even in video format. In this research paper, text based chat logs are being used. Features from pre-processed data are extracted and bipartite graph is applied for the computation of feature values. The algorithm which is used is SVM (Support Vector Machine) and the concept is optimized using Genetic algorithm.***

Keywords: Chat logs, Suspicious Activities, Support Vector Machine, bipartite graph, Genetic algorithm

I. INTRODUCTION

The most astounding aid of Internet technology is the ability of the individuals to interact with each other [1]. Millions of users around the globe communicate using instant messengers, social networking websites and many more. People are fond of using technology. Unfortunately, with this growth many illegal activities have also been increased. This research paper focuses on identifying the suspicious activities in chat logs. Suspicious activities can be any criminal activity or cybercrime including fraud and financial crimes, cyber-terrorism, cyber-extortion or it can be cyber-bullying [2] [3]. The proposed system works on text based chat log files. A rule based engine is used for extracting key features which are vocabulary terms, users and sessions are identified [4]. The classification is done with the help of SVM (Support Vector Machine) classifier and optimized using Genetic algorithm.

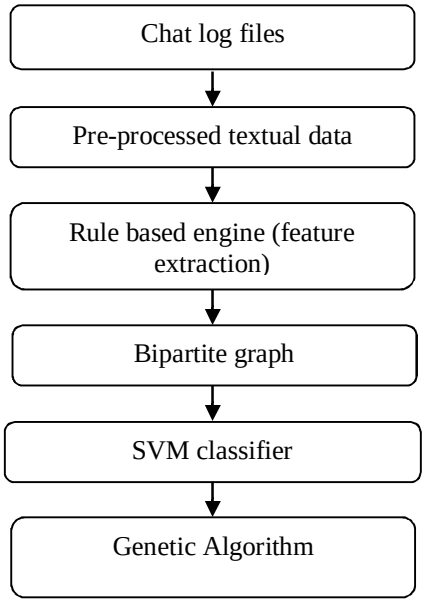

Fig. 1: Basic Process of Suspicious Activity Detection

Firstly, the chat logs files based textual dataset is collected which is available online. Then the data is pre-processed for the noise removal and to acquire valuable data. A rule based engine is used for feature extraction set and the computation of extracted key

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

features is done by constructing bipartite graph and then SVM is applied. The concept is optimized as well as the user profile showing some suspicious activity is discerned with the help of Genetic algorithm. The basic process of identification of suspicious activity is shown in figure 1.

The rest of the paper is structured as follows: Section II includes the work related to suspicious activity detection. Section III describes the basic concepts used in proposed work. In section IV, proposed system for the suspicious activity detection is explained and Section V describes the results & discussion and Section VI concludes the paper.

II. RELATED WORK

This section presents the work of the researchers related to the suspicious activity detection. A comparative analysis of their work is done which includes the key features as well as the different techniques used for the identification and is shown in table 1.

A. Sahasrabuddhe et al. (2017) [5]

have done a survey on intrusion detection system using data mining techniques. An intrusion detection system is a software based application which monitors the system for any malicious activity as well as suspicious transactions. If either of the activity is encountered then is reported to the database administrator. In the proposed system Naïve Bayes classifier is used to detect and classify all the SQL injection attacks. The training dataset consists of an XML file containing 1024 patterns with total number of 10 features. In the proposed system user input is compared to the XML file and if attack is detected then an alert message is sent to the administrator indicating the presence as well as the type of attack.

B. Sahoo et al. (2017) [6]

used text mining techniques for investigating different chat logs. The author has proposed an illegal activity detection system from the data available on social media. It also identifies the associated criminal profiles. The framework which is proposed in this paper works on the message data publically available on social networking websites. The data is extracted from the web and then it is normalized by removing punctuation marks, stop words, slangs etc. They have applied n-gram technique and the algorithm used for the identification is HITS (Hyperlink Induced Topic Search). The three features which are considered in the framework are key terms, key sessions and key users. Thus they are not only identifying the suspicious activity but also the user profile associated.

C. Jiang et al. (2016) [7]

have proposed a system to identify suspicious information as well as fraud deviations. The algorithm which is used is a novel algorithm of CrossSpot. To identify whether the information block contains suspicious data, a metric based approach is used. Initially, an experiment is performed based on Erdos-Renyi-Poisson model deploying the concept of synthetic data. For the experimentation the dataset which is used is based on the twitter's hashtag hijacking. Further, the proposed concept of CrossSpot is compared with two approaches named Singular Value Decomposition (SVD) and High Order SVD (HOSVD). F1 score is calculated and the concept of CrossSpot is sowing efficient results.

D. Oprea et al. (2015) [8]

have worked on large scale log data to detect the malware infections at their early stage in the enterprise network. As the amount of traffic generated on a network is increasing day by day, the cyber attacks are also increasing. The authors have proposed a system to detect these attacks early so as to prevent from the further damage. They have considered two datasets; LANL dataset and AC dataset. LANL dataset consists of DNS traffic including 3.81 billion DNS queries and 3.89 billion DNS responses. The size of dataset is 1.15TB which is collection of only two months network traffic. On the other hand, AC dataset amounts 38.14TB over two months consisting of logs of web proxies. The authors have to face some challenges in order to work on such large volume of data. They have applied a suite of technologies in order to normalize as well as to reduce the volume of data. They have only considered the infection patterns which are general in different types of network. The training module consists of normalized data and then host profiling is done. The next step in training is C&C communication modelling. The algorithm used for the implementation is Belief Propagation algorithm which works in two modes. The former mode consists of the SOC which is comprised of hints of hosts and the latter is no hint mode. The feature set consists of domain connectivity features, web connection features, registration data features and many more. The algorithm is applied in both of the modes and showing efficient results.

E. Tayal et al. (2015) [9]

have used data mining approach to develop a CDCI system i.e. crime detection and crime identification system. The authors have considered Indian criminal dataset over the time period of 2002-2012 consisting of criminal cases like national crime records, rape

cases, criminal alerts, protest cases, journalist's cases etc. The dataset collected is categorised into 35 major category groups. The evaluation of the results is done for the seven major Indian cities namely: Delhi, Mumbai, Kolkata, Pune, Jaipur, Hyderabad and Bangalore. For the classification and identification of criminal cases is KNN (K nearest neighbour) classifier is used and for the verification of the criminal activities WEKA tool is used. For the crime detection and crime identification a graphical user interface is designed and to enhance the clustering process of k-means Google maps are considered. The whole process of the identification as well as the classification of the crime and criminal activities is implemented in Java Netbeans and the proposed system has achieved efficient results.

F. Dogra et al. (2015) [10]

have considered video based data to detect the anomalous activities. The proposed system is implemented using weighted RAG i.e. region association graph. The proposed framework comprised of surveillance scenes which are geometrically represented. The proposed system consists of three phases. The initial phase consists of scene segmentation as well as the feature extraction. The scenes are segmented into regions with similar characteristics. Thus the approach adopted for the scene segmentation is region growing approach. The second step consists of the representation of the segmented scene using RAG. A node is assigned to each region in the connected graph with their respective weights and the nodes which are not assigned any region are initialized with zero weight. The third and final phase consists of the detection of the anomalous activities. The authors have considered two datasets CAVIAR and ViSOR. The former dataset consists of 240 second video clip and the latter consists of multiple long duration videos with an average of 40-60 minutes clip. The algorithm has shown efficient results to detect the anomalous activities.

G. Kumar and Singh (2013) [11]

have considered the chat conversation of users over the chat messengers and the social networking websites available online. The authors have used the latent sentiment analysis approach to detect the suspicious user profiles by identifying the sentiments of the users associated to the chat conversation. The proposed concept also identifies the group of users showing similar sentiments for any particular topic. The authors have divided the work in five basic steps. The initial step is the database/data monitoring system available online. The second step is the identification of the suspicious messages using keyword based approach. The third step is the LSA i.e. latent sentiment analysis. Then comes the identification of the suspicious users and last step is the visual representation of the identified users. The social networking website which is considered for the experimentation is "Manipal Net" which is a private network. The proposed concept is explained competently but is tested on limited websites.

H. Anwar et al. (2014) [12]

proposed that with the help of chat logs and text mining techniques, a graph can be generated which will help us to find the users interests during their particular chat log conversations. 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 .When they would be taking part in at least one common group chat session. To that they gave weights to the given ties and when these would overlap in user sessions, a particular set of user emotion was established. And at the end they showcased three cyber crime investigation scenarios and each having its own very view of user group identification. These following experiments were based on the data set which was composed of nearly 1100 chat logs of the 11,143 chat sessions. The following chat logs were collected for duration of 29 months, which gave away all the vital key words, during the sessions hence helping during the crime investigations. The chat logs are 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.

I. Iqbal et al. (2012) [13]

have considered chat logs for mining the criminal networks. The proposed framework is designed to extract the activities as well as the social network of the suspects. The framework is designed is such a way that it can easily analyse the unstructured data and also analyse the content of the messages. The authors have divided the framework in three modules. The first module is the clique miner. This module is working in three steps. In the first step the chat log data is getting divided into sessions and then with the help of Named Entity Recognition (NER) tool the entities are extracted. In the third step a naïve approach is used to identify all the cliques present in the chat log. The second module is topic miner. In this module the common chat topics are extracted. Keyword based approach is used for the extraction after applying some pre-processing procedures. The pre-processing includes tokenization,

stemming and stop-word removal. The third and final module is information visualizer. The evaluation of the experiment is done on synthetic dataset which is collected from MSN chat logs and thus showing efficient results.

TABLE i

COMPARATIVE ANALYSIS OF THE WORK FOR THE DETECTION OF SUSPICIOUS ACTIVITIES

Author and Year	Work Done	Technique Used	Key Features
Sahasrabuddhe et al. (2017) [5]	Software based application named as Intrusion detection system to detect malicious activities	The proposed concept is implemented with the help of Naïve Bayes classifier.	• Dataset consist of XML file with 1024 patterns with features like probability calculator and signature comparison.
Sahoo et al. (2017) [6]	Proposed a system to detect Illegal activity and criminal profile from data available on social media	N-gram technique in association with HITS (Hyperlink Induced Topic Search) algorithm.	• Three features are used which are: key terms, key users and key session.
Jiang et al. (2016) [7]	Proposed a system to identify fraud deviations and suspicious information from twitter's hashtag hijacking dataset	Algorithm used is CrossSpot algorithm and the experiment is based on Erdos-Renyi-Poisson model.	• As compared to SVD and HOSVD the proposed concept shows improved results in terms of F1 score.
Oprea et al. (2015) [8]	The authors have worked on large scale enterprise log data of two datasets; LANL and AC dataset , to detect malicious infections	Belief Propagation algorithm is used to process two modes of data i.e. with hint and without hint.	• Feature set consist of web connection features, domain features and registration features.
Tayal et al. (2015) [9]	The authors have developed crime detection and crime investigation system (CDCI) from Indian criminal dataset.	The proposed CDCI system is implemented using Java Netbeans with the help of KNN (K-nearest neighbour) classifier.	• Different criminal cases from the seven Indian major cities are considered.
Dogra et al. (2015) [10]	The authors have considered two video based datasets CAVIAR and ViSOR to detect anomalous activities.	The proposed system is implemented using weighted Region Association graph i.e. RAG.	• The scenes from the video clip are segmented and then the segmented scenes are converted into regions with similar characteristics.
Kumar and Singh (2013) [11]	The authors have considered a private network i.e. "Manipal Net" to detect the suspicious user profiles by identifying their associated sentiments.	The approach which is used is LSA i.e. Latent sentiment analysis.	• The clusters of the user profiles showing same sentiment on particular topic are also encountered.
Anwar et al. (2014) [12]	Used HITS approach for the identification of user interests.	N Gram and HITS (hyperlink induced topic search) algorithm	• Graph based theory for the analysis of user's interests during their particular chat log conversations.
Iqbal et al. (2012) [13]	The authors have designed a framework to encounter criminal network from chat logs.	Framework is divided into three modules; clique miner, topic miner and information visualize	• Word tokenization, stop word removal and stemming. • Name entity recognition is used for entities extraction.

III. BASIC CONCEPTS

This section presents the basic concepts of Support Vector Machine and Rule based engine and Genetic Algorithm.

A. Support Vector Machine

Support Vector Machine is a supervised machine based learning approach which was initially introduced in 1963 by Vapnik and Chervonenkis [14]. SVM is a linear classifier but it also performs non-linear classifications. In order to separate the different associated classes, SVM determines the linear separations in the search space. SVM algorithm is entirely based on structural risk minimization principle. Thereby, it is used to resolve the problems of classification, clustering and regression statistically. SVM classifies the data into different classes and the classification decision is entirely based on support vectors [15].

The mechanism of the Support Vector Machine consists of three modules. The first phase is training phase in which the classifier is trained with the input data. In the second phase a model is built based on the input data and in the third phase classification is done by the SVM classifier. The most important step in the whole mechanism is the selection of the parameters. Failure in proper parameter selection generally leads to the failure in classification capabilities.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 5 Issue VII, July 2017- Available at www.ijraset.com

B. Rule Based Engine

Rule based engine [16] is a concept of storing data or information in order to elucidate information in a functional way. A rule based engine consists of some set of rules on the basis which decides what actions are to be taken. A knowledge base is there which is nothing but the list of rules based on the input information. Inference engine is there which is responsible for taking and performing actions based on the knowledge base information. It is the processing element which processes the data and takes action according to the given input and thus delivers the corresponding results.

C. Genetic Algorithm

Genetic algorithms [17] are nature inspired search procedures which are based on natural phenomenon, genetics and selections. Genetic algorithms are mainly used to solve optimization problems. The optimization problem can be of two types constrained optimization problem and unconstrained optimization problem. Genetic algorithm solves both types of problems.

The initial phase of algorithm consists of set of solutions which are basically input to genetic algorithm. These set of solutions are also known as population. The algorithm works on iterations until the optimal solution is achieved. At each iteration, individuals from current population are being selected by the algorithm which is further used as parents thereby producing children for the next population. In this way the process continues. The main reason behind the selection of the solution from one population and then use it to form new population is the hope that the new population is going to be better than the old one. A population point is generated at each iteration and the only the best population point proceed towards optimal solution.

IV. PROPOSED ALGORITHM

This section represents the concept of suspicious activity detection from chat logs data available online. The proposed concept is implemented with the help of Support Vector Machine and Genetic Algorithm. Support Vector machine is a statistical supervised machine learning classifier which is used to classify the different suspicious activities based on the keywords. A rule based engine is used to extract the key features which are vocabulary terms, users and session. Further bipartite graph is constructed for the computation of the key features. The concept of genetic algorithm is used to get the optimal result as well as for the identification of the user profiles associated with the suspicious activity. Figure 2 shows the flow chat of the work.

Input: Text based data from chat log

Output: Suspicious Activity detected and associated User profile

V. ALGORITHM

A. Consider the chat log file available online

B. Extract the unstructured message data from the log file.

C. After the extraction, the data is normalized by removing the different punctuation marks and the numeric digits.

D. The normalized data is then pre-processed by applying removing stop words. The functions of tokenization and stemming are also applied to process the data.

1) A pre-defined list of root words is considered corresponding to the dataset. An array of Suffix is developed which is to be identified and then removed to extract the root word.

2) Boolean == word to be stemmed exist?

3) if Boolean == true, No Stemming required
Else
getRootword(word_to_be_stemmed);

4) If Suffix exists in the Suffix array Replace_Suffice(with _the_Stemmed_word); Else
Go to next word.

5) Remove Stop words. A list of stop words is prepared on basis of which stop words are removed from dataset.

6) After stop word removal and stemming, word tokenization is done i.e. the data is converted into tokens of words.

7) Apply the Rule based engine to extract the key features which are vocabulary terms, users and sessions and construct a bipartite graph to compute the features.

8) Apply Support Vector Machine classifier to identify and to classify the dataset into different suspicious categories by matching the dataset keywords with the expert keyword list. The working and the formulation for detection of suspicious keywords by SVM is shown below:

SVM = SVMtrain (suspicious_count, suspicious_category)

Where,

SVMtrain is the training function of SVM.
suspicious_count is the frequency count of the suspicious keywords i.e. number of times a keyword has occurred.
suspicious_category defines the category of the suspicious keyword i.e. whether it is related to cyber security, terrorism or cyber bullying etc.

E. Apply the Genetic algorithm for the optimization of classification and detection of user profile with suspicious keyword.

1) All the tokens i.e. keywords are considered as population P. the mutation probability is considered as Pm and the crossover probability as Pc

2) The performance of population is measured by fitness function.

3) Initial population is randomly generated of size N as shown:

x1, x2, x3,…….Xn

4) Calculate the fitness function as shown:

f(x1),f(x2),f(x3),…….f(xN)

5) Apply the genetic operators i.e. mutation and crossover.

6) Repeat the process until the size of new population becomes equal to the size of old population i.e. N.

7) Replace the initial population with the new population

8) Go to step 7.4 and repeat the whole process until a satisfactory result is achieved.

VI. RESULTS AND DISCUSSION

This section presents the performance evaluation of the proposed suspicious activity detection system from online available chat logs. A list of keywords is prepared related to different suspicious activities. Figure 3 & 4 shows some keywords related to terrorism and cyber security respective.

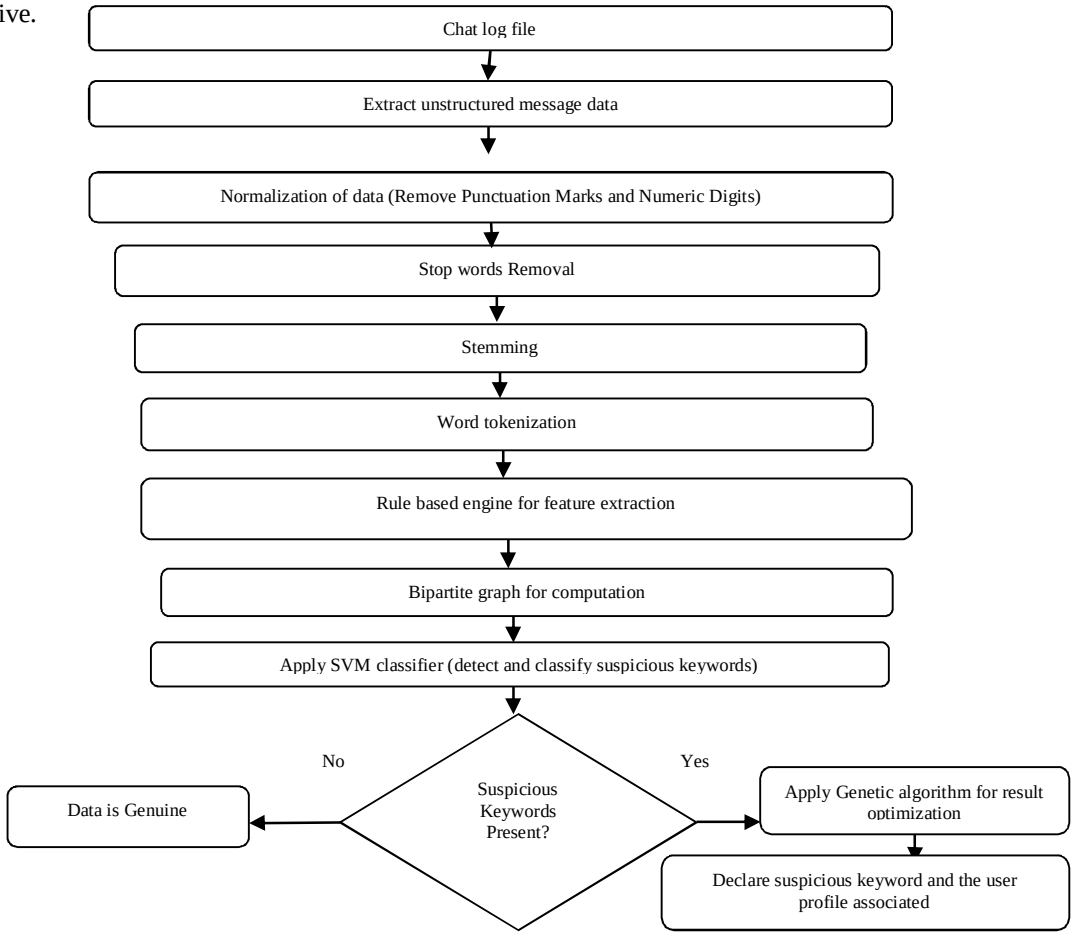

Figure 2: Workflow for Detection of Suspicious Activity using SVM and Genetic Algorithm

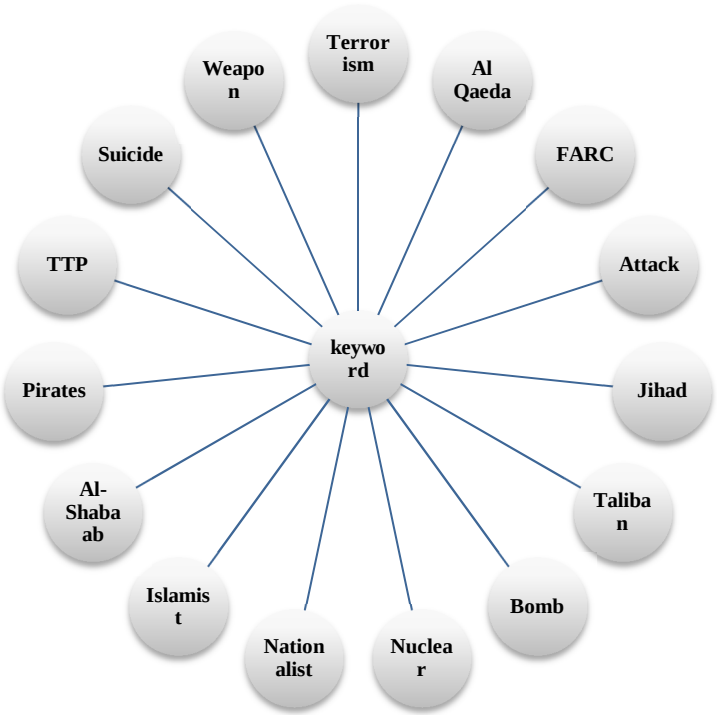

Figure 3: Terrorism related Keywords

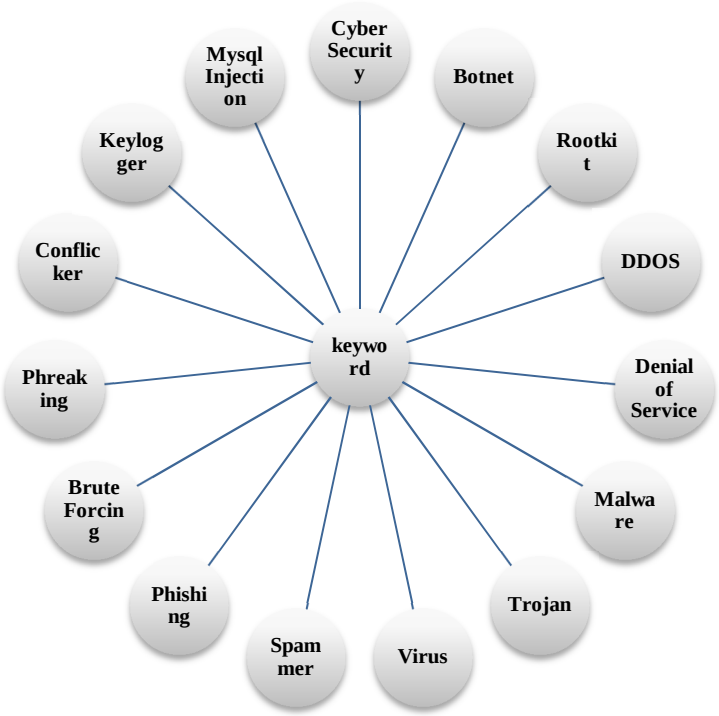

Figure 4: Cyber Security related Keywords

The frequency count of each keyword is calculated which is shown in table 2. It indicates the presence of the keywords in the dataset. The proposed concept is implemented using SVM classifier and then it is optimized and further user profiles are also

identified with the help of Genetic algorithm. The test results are show for both SVM classifier as well as the integration of both SVM and Genetic Algorithm. Figure 5 shows the comprehensive results.

TABLE 2
FREQUENCY COUNT OF KEYWORDS IN DATASET

Keyword	F_Count	Keyword	F_Count	Keyword	F_Count
Terrorism	25	Al-Shabaab	01	Malware	23
Al-Qaeda	10	Pirates	06	Trojan	45
FARC	03	TTP	02	Virus	54
Attack	35	Suicide	49	Spammer	11
Jihad	10	Weapon	53	Phishing	26
Taliban	27	Security	67	Brute Forcing	31
Bomb	43	Botnet	14	Phreaking	09
Nuclear	22	Rootkit	05	Conflicker	06
Nationalist	07	DDOS	14	Keylogger	02
Islamist	31	Denial of Service	48	Mysql Injection	19

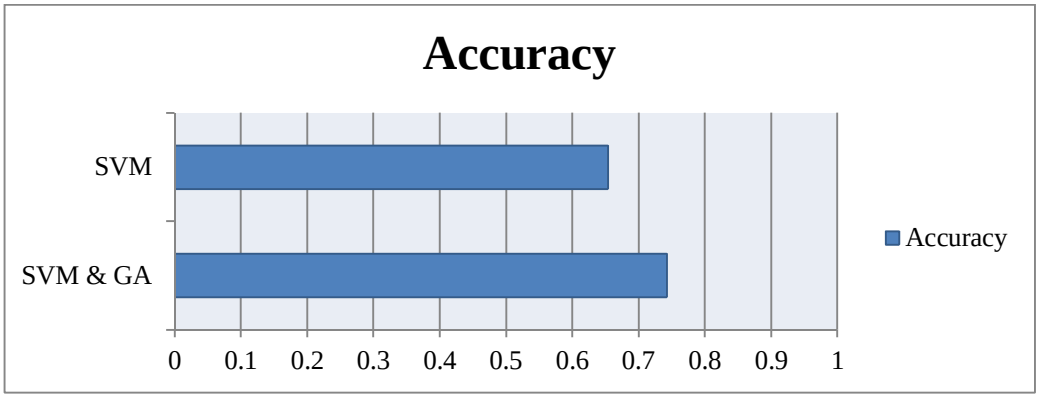

Figure 5: Comprehensive Results of SVM and SVM & GA

Figure 5 clearly states that the results procured by the integrated approach of SVM & GA with the accuracy of 74.3% are more accurate than the results obtained by SVM classifier which shows an accuracy of only 65.4%.presented an integrated approach of SVM and PSO algorithm for the detection of suspicious activities on online forums. The proposed framework presents the work flow as mentioned in figure 2.

VII. CONCLUSION

The ability to communicate online anytime, anywhere and with anyone is one of the most remarkable technological achievement. But unfortunately, this achievement is being misused by some people. Organizations from all over the globe are looking for solutions to prevent the cyber crimes or suspicious activities. In this paper, we have proposed a concept to identify the suspicious activities like terrorism or cyber security related activities as well as the user profiles associated to particular suspicious keyword. The dataset of chat logs is considered which is in textual format available online. The proposed concept is implemented using Support Vector machine classifier and then further results are optimized with the help of the concepts of Genetic Algorithm. The results obtained from genetic algorithm show more accuracy than the results procured by SVM classifier.

In future the proposed concept can also be used for some real time datasets or applications to detect and to prevent different suspicious activities.

REFERENCES

[1]. Tapscott, Don. Growing up digital: The rise of the net generation. Vol. 352. New York: McGraw-Hill, 1998.
[2]. Burden, Kit, and Creole Palmer. "Internet crime: Cyber Crime—A new breed of criminal?." Computer Law & Security Review 19, no. 3 (2003): 222-227.
[3]. Gordon, Sarah, and Richard Ford. "On the definition and classification of cybercrime." Journal in Computer Virology 2, no. 1 (2006): 13-20.
[4]. Moore, Robert. Cybercrime: Investigating high-technology computer crime. Routledge, 2010.
[5]. Sahasrabuddhe, S. Naikade, A. Ramaswamy, B. Sadliwala, and P. P. Futane, "Survey on Intrusion Detection System using Data Mining Techniques," pp. 1780–1784, 2017.
[6]. G. Sahoo, P. Pawar, K. Malvi, A. Jaladi, and K. Khithani, "Environment Monitoring System based on IoT," pp. 1083–1091, 2017.
[7]. M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "Spotting Suspicious Behaviors in Multimodal Data: A General Metric and Algorithms," IEEE Trans. Knowl. Data Eng., vol. 28, no. 8, pp. 2187–2200, 2016.
[8]. Oprea, Z. Li, T. F. Yen, S. H. Chin, and S. Alrwais, "Detection of Early-Stage Enterprise Infection by Mining Large-Scale Log Data," Proc. Int. Conf. Dependable Syst. Networks, vol. 2015-September, pp. 45–56, 2015.
[9]. 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.
[10]. D. P. Dogra, A. A. Sk, and H. Bhaskar, "Scene Representation and Anomalous Activity Detection usingWeighted Region Association," no. August 2016, 2015.
[11]. Kumar, A. Sharath, and Sanjay Singh. "Detection of User Cluster with Suspicious Activity in Online Social Networking Sites." In Advanced Computing, Networking and Security (ADCONS), 2013 2nd International Conference on, pp. 220-225. IEEE, 2013.
[12]. Anwar, Tarique, and Muhammad Abulaish. "A social graph based text mining framework for chat log investigation." Digital Investigation 11, no. 4 (2014): 349-362.
[13]. F. Iqbal, B. C. M. Fung, and M. Debbabi, "Mining criminal networks from chat log," Proc. - 2012 IEEE/WIC/ACM Int. Conf. Web Intell. WI 2012, pp. 332–337, 2012.
[14]. Cortes, Corinna, and Vladimir Vapnik. "Support vector machine." Machine learning 20, no. 3 (1995): 273-297.
[15]. 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.
[16]. Peuschel, Burkhard, and Wilhelm Schäfer. "Concepts and implementation of a rule-based process engine." In Proceedings of the 14th international conference on Software engineering, pp. 262-279. ACM, 1992.
[17]. Horn, Jeffrey, Nicholas Nafpliotis, and David E. Goldberg. "A niched Pareto genetic algorithm for multiobjective optimization." In Evolutionary Computation, 1994. IEEE World Congress on Computational Intelligence., Proceedings of the First IEEE Conference on, pp. 82-87. Ieee, 1994.