ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Framework for the Detection of Suspicious Discussion on Online Forums using Integrated approach of Support Vector Machine and Particle Swarm Optimization

Harsh Arora and Govind Murari Upadhyay
Assistant Professor, Department of Information Technology,
Institute of Innovation in Technology & Management, GGSIP University, New Delhi, India

Abstract: With the advancements of technology, human has become more efficient to share the information via internet. It has more often observed that the news stories are initially broken up on social media sites like Twitter, Facebook, etc. and later have taken up by the other channels like printing media or electronic media channels. Any social media based message board where information and user opinion can be discussed is considered as online forums. Most of the data of online forums are stored in text format, hence the present work make use of only text format of suspected postings as evidence for investigation. But there are many suspicious users such as spammers, fraudsters, and other types of attackers that use the latest technology for the criminal activities. So, there is the need to develop tools for the recognition of suspicious activities. There is the existence of tools and methods for the recognition of suspicious information available on internet in the form of user comments or views. In this research paper, we are presenting the existing research work for the detection of suspicious information. The key features and drawbacks of existing concepts have also presented. To improve the autonomous approach, we have also presented a framework using integrated approach of Support Vector Machine (SVM) and Particle Swarm Optimization (PSO). SVM is a statistical learning based data mining approach and PSO is swarm intelligence based concept considered to optimize the parameters of SVM.

Keywords: Suspicious Activities; Support Vector Machine; Particle Swarm Optimization; Data Mining

I. INTRODUCTION

Online Forums are the online virtual social networking space where users can share their opinion and review about some kind of products [1]. Online forum data mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. It has become very critical for keeping track of suspicious activities on online forums. The data in online forums can be expressed in the form of text, audio and video format. But on internet, the most common and useful format for the discussion is text corpus [2]. Text corpus is the way to use and discuss data in textual format with proper manner. Online forums based data can be used for positive terms and can also be used by criminal authorities to make people aggressive for the legal activities. The discussion forums are kept under constant surveillance which helps to identify the apprehensive activities. Various law enforcement agencies throughout world are looking for some solutions to observe these discussion forums and access the possibility of illegal activities [3]. But for the analysis of these suspicious activities, there are various challenges like uncovering of the suspicious published contents and publications posted by users and analysis of the users' behaviour in the social media.

In this research work, we have analysed the existing concept and methods for the detection of these suspicious discussions and activities. Different authors have adapted different approaches for the recognition of suspicious or criminal activities on online textual datasets [4]. The basic process of suspicious discussion recognition is shown in figure 1. Also a novel approach of SVM and PSO has been presented for the identification of illegal activities from this textual data. SVM is a statistical learning and classification based data mining approach. PSO is swarm intelligence based concept considered to evaluate the parameters in optimized manner.

Rest of the paper is organized in the following manner. Section II explains the work related to suspicious activity detection. Section III describes the basic concepts of Support Vector Machine and Particle Swarm Optimization. In section IV, proposed framework for the suspicious information detection is explained and Section V concludes the paper.

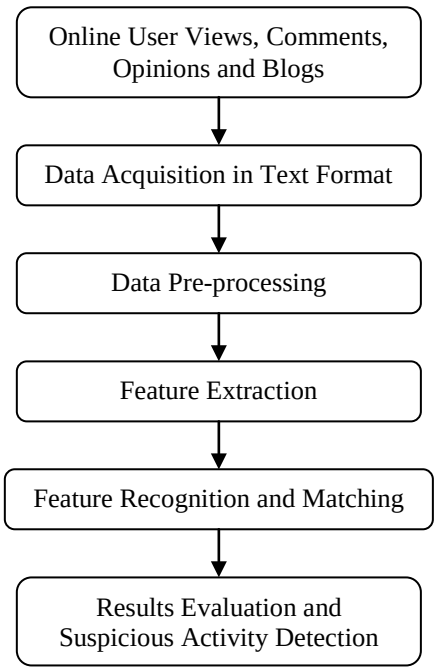

Figure 1: Basic Process of Suspicious Activity Detection

II. RELATED WORK

This section presents the existing work related to the analysis of suspicious discussion. The work of different authors has been presented here. Also a comparative analysis with the key features and drawbacks of the concepts is presented in table 1.

Jiang et al. (2016) [5] have introduced the novel algorithm of CrossSpot to spot the suspicious information and fraud deviations. Authors have used the metric based approach to define the suspiciousness of a block of information from multimodel data. Initially, authors have performed the experiment with the concept on the synthetic data which is based on Erdos-Renyi-Poisson model. Then, twitter based Hashtag hijacking dataset has been used for experimentation. The proposed concept of CrossSpot has been compared with the approach of Singular Value Decomposition (SVD) and High Order SVD (HOSVD). The proposed concept of CrossSpot shows efficient results for the experimentation with the presentation in terms of F1 Score.

Jain et al. (2016) [6] have proposed an approach to differentiate the twitter data as an actual information and rumor. Authors have extracted the twitter data for some particular topic with the help of Hashtag functions. For the validation of concept for any particular information, data of some well-known news channels has been considered and evaluated the results with semantic and sentiment analysis of tweets. For this proposed concept, authors have also presented a prototype "The Twitter Grapevine" to target the rumors specifically for Indian domains. The overall results has been evaluated based on the accuracy analysis initially for digital India & facebook.org rumors and then for KeralaHouse & Beef Rumor topics. In these results, favourable and unfavourable result predictions have been evaluated. Accuracy results for the later experiments are much lower than the former one of 76.99%.

Murugesan et al. (2016) [7] 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. The complete process of suspicious information extraction has been explained. After the pre-processing steps of stop words removal & stemming process with Brute Force algorithm, authors have used the matching algorithm for the suspicious keyword recognition. Finally authors have used the keyword spotting techniques, leaning based method and hybrid of defined approaches for the overall recognition of suspicious human activity.

Tayal et al. (2015) [8] 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. A graphical user interface has been designed for this CDCI approach and efficient accuracy results have been achieved for considered concept.

Alami and Beqqali (2015) [9] have proposed a similarity distance evaluation based approach to differentiate the suspicious content from the other authentic content/posts/blogs. The dataset of twitter text corpus has been used for the analysis. The considered concept is based on the evaluation of similarity index by comparing the social database. There are basically three steps of proposed concept as mentioned: text corpus, corpus processing and classification process using similarity approach. The concept is evaluated in terms of similarity distance index which leads to more execution time and lesser precision rate. So, there is need to improve the precision rates and execution time.

Hosseinkhani et al. (2014) [10] have presented a review for the existing concepts of crime data mining techniques for the detection of suspicious information on the web. The major challenge for the detection of suspicious concept is the day by day increase in volume of cyber data and increasing network traffic. There is the availability of various data formats like audio, video and textual data. In this research work, a theoretical review for the different mining concepts like data mining, web mining, crime data mining has discussed. On the basis of textual data mining concepts for the detection of crime activities are sequential pattern mining, classification, association rule mining and clustering etc.

Ali et al. (2014) [11] 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.

Kumar and Singh (2013) [12] have used Latent Sentiment Analysis (LSA) for the detection of suspicious users by identifying their sentiments over chat conversations on online social networking sites and chat messengers. The considered concept also identifies the cluster of people having similar suspicious sentiments for any topics. The proposed concept involves basic five steps for the identification of suspicious user clusters as mentioned: (a) Online data monitoring system and Database (b) Suspicious message identification using NLP/Keyword system (c) Latent semantic analysis (LSA) system (d) Suspicious users identification system and (e) Visual representation of suspicious users. For experimentation, authors have assumed

a private social network named "Manipal Net" as the social networking site so that suspicious users can be identified. The overall concept is well explained but tested on limited assumed network site.

Table I: Comparative Analysis of the Existing approaches for the Detection of Suspicious Discussion

Author and Year	Technique Used	Key Features	Drawbacks
Jiang et al. (2016)	CrossSpot Algorithm	The proposed concept shows improved results in terms of F1 score as compare to SVD and HOSVD	• Work is limited to only twitter dataset for some selected keywords with hashtags like 'HiJacking'
Jain et al. (2016)	The Twitter Grapevine: Web Application	Evaluates the originality of news/information available on twitter especially for Indian Rumors.	• Used the basic concept of Sentiment and Semantic analysis. So, Accuracy level is much lower. • Prototype experiment has performed for some limited Indian dataset of twitter.
Murugesan et al. (2016)	Traditional Corpus based approach	A framework for the suspicious activity recognition on online data has been presented.	• Authors did not perform the experiments on real time dataset. • Proposed concept is based on traditional corpus approach.
Alami and Beqqali (2015)	Similarity Distance based Differentiation Approach	Presented an autonomous approach for the detection of suspicious activity	• Concept lacks for the evaluation parameters like precision rate and execution time. • Also there is need to develop autonomous classification approach.
Tayal et al. (2015)	KNN approach in Crime Detection and Criminal Identification Concept	Proposed CDCI concept for the prediction and identification of Criminal activities in major Indian Cities	• Concept lacks due to increasing crime data volume and lesser privacy and data security.
Hosseinkhani et al. (2014)	Crime Data Mining	Presented the review of the traditional data mining approaches.	• Concepts are presented only in theoretical manner. • There is no actual detection of suspicious activity on web.
Ali et al. (2014)	Suspicious Pattern Detection Algorithm	Identification of suspected cyber threat in instant chat messenger available on Social Networking Websites and Instant Messengers	• Proposed concept is vulnerable to security attacks. • There is no encryption standard set for the chat conversations. • Not able to handle the big data. • Concept is limited to single language.
Kumar and Singh (2013)	Latent Sentiment Analysis	Detection of suspicious user cluster with similar behaviour by identifying their sentiments over chat conversations on online social networking sites and chat messengers	• Limited to assumed private social network • Evaluation parameters have not been considered by authors.

III. BASIC CONCEPTS

This section presents the basic concepts of Support Vector Machine and Particle Swarm Optimization.

A. Support Vector Machine

SVM algorithm [13] is superior learning models which are generally associated with a learning algorithm applied to it which analyses data, all this done to regroup or for categorization. It also performs the non linear classifications. It is considerable statistical approach used to resolve the problems related to supervised regression & classification with well-built theoretical fundamentals that follows the principle of structural risk minimization. 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 [14].

Support Vector machine works on the three steps based mechanism. First step is to take input data in a training phase, second step is to build a model using the input data and final step is output with a hypothesis that can predict the function with future data [15]. The goal of SVM is to produce a model which predicts target value of data instances in the testing set which are given only the attributes.

B. Particle Swarm Optimization

In 1995, Kennedy and Eberhart [16] developed a stochastic optimization technique known as PSO, inspired by the

intelligence of swarm behaviour such as fish and bird schooling. PSO uses the theory of social interaction for problem solving. The two main optimization factors of this technique are local search and global search optimization features. In local search, particles get their own individual best optimized solution using their own experiences. In global search, the experience of one bird is shared with the experience of another bird and finally gets result as a global best solution. [17].

This algorithm works in an iteration manner and moves closer to the best solution. Initially particles begin the process by the casual fly in the form of population of *N* particle solution. In the S-dimensional space, the position of the *i*th particle is represented as a point in this space where S is the number of variables participated [18]. In the entire process, particles try to find the global best solution.

IV. PROPOSED ALGORITHM

The integrated proposed approach of SVM and PSO is considered for the detection of suspicious discussion on online forums. In this integrated approach, Support Vector Machine is statistical learning concept used as the classification and regression models to differentiate the suspicious activity based keywords from the genuine information. Particle Swarm Optimization is swarm intelligence based concept well known for the local and global best optimized search solution. Here, we have considered the integrated approach to optimize the solution upto the maximum possible iterations. The flow chart for the work is shown in figure 2.

Input: Textual Data in online chat form
Output: Detection of Suspicious discussion

ALGORITHM

Step 1: Consider the unstructured training dataset available on online websites and chat messengers.

Step 2: Normalize the data on linguistic level by removing the numeric keywords from the dataset.

Step 3: Pre-process the data by applying the functions of stemming and stop work removal and tokenization of keywords.

3.1. Develop Array of Suffixes which are to be identified and removed to get a Root Word. A predefined list of root words is considered for the considered text database.

3.2. Boolean = Check word to be Stemmed Exists in the Dictionary.

3.3. If Boolean = true then no Stemming required.

Else

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

3.4. Check if Suffix exits in the Suffix Array developed in Step 3.1.

Replace Suffix (Word to be Stemmed Suffix)

Else

Go to next word.

3.5. Remove the words of stop function. For stop word removal a list of predefined keywords has been considered.

3.6. After stemming and stop word removal tokenize the data as a token of keyword.

Step 4: Apply the concept of Support Vector Machine for classification and matching of the suspicious keywords with expert dataset keyword list of values. The formulation for the suspicious keyword detection is considered below:

SVM = SVMtrain (suspicious_level, suspicious_type)

Where,

SVMtrain is the SVM training function.

suspicious_level maintains the levels of suspicious keywords by evaluating the count of repeating suspicious keywords

suspicious_type maintains the different types of suspicious keywords like rape case, terrorism activity, cyber crime, malware etc.

Step 5: Apply the Particle Swarm optimization for the optimization of classification and suspicious keyword detection.

5.1. All the keywords based tokens are considered as the particles. Initially these particles randomly fly and search for the food sources in the form of suspicious keywords in data. Then search for the local and global solution.

5.2. The performance of each particle depends upon the level of suspicious activity that has to optimize.

5.3. Each particle flies over the n- dimensional outer search space and keep updating the following information:

- $\boldsymbol{X_i}$ – current position of particle
- $\boldsymbol{P_i}$ – the personal best position of particle
- $\boldsymbol{V_i}$ – the current velocity of particle

5.4. The velocity updates in PSO can be calculated using the formula given below:

$$V_{i(t+1)} = \omega\, V_{i(t)} + c_1 r_1 \left(P_{i(t)} - X_{i(t)}\right) + c_2 r_2 \left(P_g - X_{i(t)}\right)$$

Now, V_i is the new velocity. So, the position of the particle updates with the velocity as below:

$$X_{i(t+1)} = X_{i(t)} + V_{i(t+1)}$$

5.5. Update the positions for each particle and store the global best solutions.

Step 6: Repeat the iteration steps and obtain the optimized results for the considered suspicious activity by matching the keywords with the expert dataset.

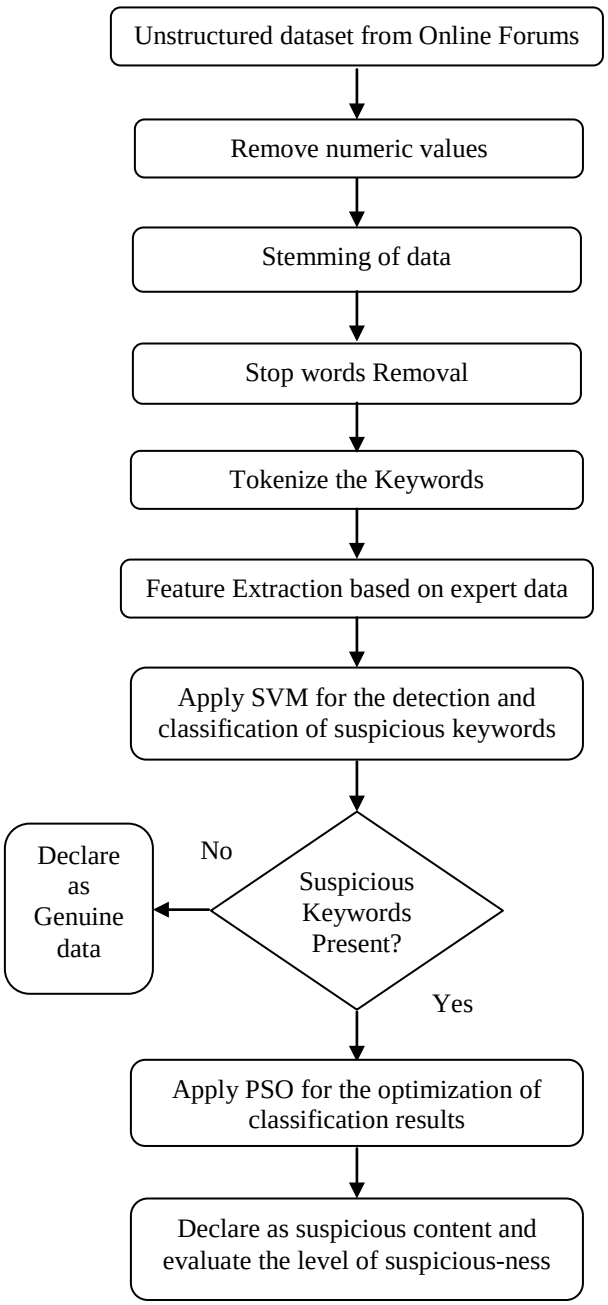

Figure 2: Workflow for Suspicious Activity Detection using integrated approach of SVM and PSO

V. CONCLUSIONS

Advancement in communication technology have decreased the advantage of its use due to its increasing use in illegal and suspicious activities. Agencies from all over the globe are looking for the valuable solution to observe these online forums. There is the existence of various traditional methods and tools for the detection of these suspicious activities. In this paper, we have presented the existing concepts for the identification of suspicious activities. The considered author's work concentrates on the use of concepts like CrossSpot Algorithm, The Twitter Grapevine: Web Application, Traditional Corpus based approach, Similarity Distance based Differentiation Approach, KNN approach in Crime Detection and Criminal Identification Concept, Crime Data Mining, Suspicious Pattern Detection Algorithm, Latent Sentiment Analysis etc. Due to various research gaps in the existing concepts, we have 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.

In this paper, an integrated approach of SVM and PSO has been proposed. In future, this technique can be applied for any real time dataset for the detection of suspicious keywords.

VI. REFERENCES

[1]. DeSanctis, Gerardine, Anne-Laure Fayard, Michael Roach, and Lu Jiang. "Learning in online forums." *European Management Journal* 21, no. 5 (2003): 565-577.

[2]. Li, Nan, and Desheng Dash Wu. "Using text mining and sentiment analysis for online forums hotspot detection and forecast." *Decision support systems*48, no. 2 (2010): 354-368.

[3]. Johnson, David R., and David Post. "Law and borders: The rise of law in cyberspace." *Stanford Law Review* (1996): 1367-1402.

[4]. Edwards, Matthew, Awais Rashid, and Paul Rayson. "A systematic survey of online data mining technology intended for law enforcement." *ACM Computing Surveys (CSUR)* 48, no. 1 (2015): 15.

[5]. 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.

[6]. Jain, Suchita, Vanya Sharma, and Rishabh Kaushal. "Towards automated real-time detection of misinformation on Twitter." In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on*, pp. 2015-2020. IEEE, 2016.

[7]. 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).

[8]. 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.

[9]. Alami, Salim, and Omar EL Beqqali. "Detecting Suspicious Profiles Using Text Analysis Within Social Media." *Journal of Theoretical & Applied Information Technology* 73, no. 3 (2015).

[10]. Hosseinkhani, Javad, Mohammad Koochakzaei, Solmaz Keikhaee, and Javid Hosseinkhani Naniz. "Detecting suspicion information on the Web using crime data mining techniques." *International Journal of Advanced Computer Science and Information Technology* 3, no. 1 (2014): 32-41.

[11]. 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.

[12]. 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.

[13]. Suykens, Johan AK, and Joos Vandewalle. "Least squares support vector machine classifiers." *Neural processing letters* 9, no. 3 (1999): 293-300.

[14]. Tong, Simon, and Daphne Koller. "Support vector machine active learning with applications to text classification." *Journal of machine learning research*2, no. Nov (2001): 45-66.

[15]. Amari, Shun-ichi, and Si Wu. "Improving support vector machine classifiers by modifying kernel functions." *Neural Networks* 12, no. 6 (1999): 783-789.

[16]. Kennedy, James. "Particle swarm optimization." In *Encyclopedia of machine learning*, pp. 760-766. Springer US, 2011.

[17]. Clerc, Maurice. *Particle swarm optimization*. Vol. 93. John Wiley & Sons, 2010.

[18]. Poli, Riccardo, James Kennedy, and Tim Blackwell. "Particle swarm optimization." *Swarm intelligence* 1, no. 1 (2007): 33-57.