Framework for Surveillance of Instant Messages in Instant messengers and Social neworking sites using Data Mining and Ontology

Mohammed Mahmood Ali
Department of CSE
MJCET, Osmania University
Hyderabad, Andhra Pradesh, India
mahmoodedu@gmail.com

Khaja Moizuddin Mohammed
Department of CS & SE
University of Hail
Hail, Saudi Arabia
moizuddin17@gmail.com

Lakshmi Rajamani
Department of CSE
UCE, Osmania University
Hyderabad, Andhra Pradesh, India
drlakshmiraja@gmail.com

Abstract— **Innumerable terror and suspicious messages are sent through Instant Messengers (IM) and Social Networking Sites (SNS) which are untraced, leading to hindrance for network communications and cyber security. We propose a Framework that discover and predict such messages that are sent using IM or SNS like *Facebook, Twitter, LinkedIn*, and others. Further, these instant messages are put under surveillance that identifies the type of suspected cyber threat activity by culprit along with their personnel details. Framework is developed using Ontology based Information Extraction technique (OBIE), Association rule mining (ARM) a data mining technique with set of pre-defined Knowledge-based rules (logical), for decision making process that are learned from domain experts and past learning experiences of suspicious dataset like *GTD* (Global Terrorist Database). The experimental results obtained will aid to take prompt decision for eradicating cyber crimes.**

Keywords-Instant Messengers(IM); Social Networking Sites(SNS); Ontology based Information Extraction; Association Rule Mining(ARM); Knowledge based rules.

I. INTRODUCTION

Internet evolutions led to the growth of innumerable cybercrimes. Criminals adapted to send suspicious messages via mobile phones, Instant Messengers and Social Networking Sites, which is difficult to trace their criminal activities dynamically. The E-crime department must be improvised with the development of technology to find criminals. Many of the Instant Messaging Systems (IMS) developed restricted their limit for sending messages, video and audio conferencing. They are not well equipped to detect online suspicious messages.

Cybercrime activities are increasing day by day. The CIA, FBI and other federal agencies are actively collecting domestic and foreign intelligence information to prevent future cyber attacks. Recently the Internet Crime Complaint Center (IC3) released the report in 2012 of cybercrimes, with the latest data and trends of online criminal activity [1].We surveyed various architectures of Mobile Phones, Instant messengers and Social Networking sites [2, 3]. These studies helped us to develop a new Framework. WordNet, is a lexical database, contains a huge amount of information consisting of (155287 words organized in over 117000 Synsets for a total of 207000 word-sense pairs) words that is useful for our study for scanning and filtering the text messages stored in TDB (Text Database) [4]. WordNet is used as features for classification of words from unstructured text. Similarly, WordNet Ontology based on information extraction technique is discussed in [5]. Our Contribution includes improving the existing IMS using data mining technique of Associative rules [6], Ontology based information retrieval technique (probabilistic models), which is guided with pre-defined Knowledge based rules and ARM. Early detection of suspicious messages from instant messaging systems (Mobile Phone, IM and SNS) is possible with our proposed Framework to identify and predict the type of cyber threat activity and trace the criminal details.

This Section gives an overview of cybercrimes performed in IMS and deficiencies exist. The remainder of this paper is organized as follows: In Section II, we reviewed recent research advances in identifying criminals from cyberspace. The Section III illustrates the operational phases and its implementation of our proposed Framework to validate these instant messages sent are suspicious or not and steps for tracing the culprits. The experimental results are shown in Section IV, when tested with dataset collected from Global Terrorism Database (GTD) [25]. Finally, Section V concludes the paper with an outlook towards future research directions for adding the features of our proposed Framework to current IMS.

II. PROBLEM STATEMENT AND RELATED WORK

Nowadays, it's difficult to survive without IMS as users are addicted to. Trillions of messages are sent each day through emails and IMS. Popular IMS such as AOL, MSN, ICQ, Yahoo, Google Talk, Skype, Facebook, Twitter, and LinkedIn have changed the way of communication with friends, acquaintances, and business colleagues. Once limited to desktops, popular instant messaging systems are finding their way onto handheld devices and cell phones, allowing users to chat virtually from anywhere.

The Social approach to detect malicious web content for Facebook, with security heuristics is limited to identify malicious URL links [8]. Recently the Facebook static messages are scanned to identify criminal's behavior [9]. Detection of suspicious emails from static messages using decision tree induction proposed which is purely dependent on highest information entropy that identifies the messages are

 978-1-4799-2608-4/14/$31.00 ©2014 IEEE

deceptive or non-deceptive [10]. Similarly, architecture for detection of phishing attacks from IM for text messages using data mining technique was proposed [11]. All of these techniques are inefficient to predict the type of suspicious cyber threat activities like Murder, Terrorist attack, Match Fixation, Drug smuggling, Kidnap, Robbery and theft, Corruption charges and Sexual harassment in its entirety, similar to this we proposed a framework for surveillance of instant messages in our earlier work without data mining technique [7]. In this section, we described the significant vulnerabilities in IMS and the type of cyber threat to be detected which criminal's use that lead to cyber crimes.

III. Proposed Framework For Suspicious Messages

In this Section we explore the operational phases of proposed Framework as shown in Fig. 1. The Suspicious Pattern Detection (SPD) algorithm initiates the steps to capture the instant messages that are communicated between the users and then, stores them into database for identifying suspicious messages. SPD algorithm is shown in Fig. 8, apart from that, it also instigates the e-crime monitoring system program to trace the culprit details for E-crime department.

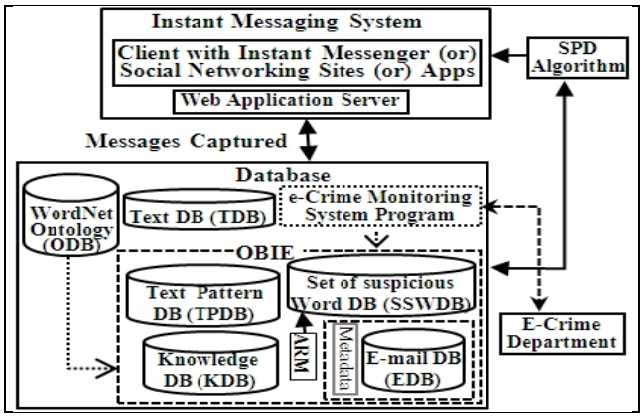

Fig. 1. Proposed Framework to detect suspicious messages from Instant Messaging Systems (IMS).

This Framework makes use of databases (to store dynamic messages) and Ontology Based Information Extraction techniques (to detect suspicious words from messages) guided with Pre-defined Knowledge based rules checked with ARM. Following are the major tasks performed are as follows: 1) *Word Extraction from unstructured text* 2) *E-crime monitoring system program* and 3) *SPD algorithm* (a generic approach). The working pseudo code of our Framework is demonstrated using Schematic-cum-algorithmic representation is shown in Fig. 2. Steps of algorithm are illustrated as follows:

1. In this step, filtering of unnecessary words from messages (TDB) is done; during this process, the suspicious words are identified using algorithms discussed in Fig. 3 and Fig. 6. The detected suspicious words are stored in *TPDB* for further processing.
2. Once suspicious words are found, the message(s) are marked as suspicious in *SSWDB*, based on rules given in Table I which is guided with pre-defined knowledge based rules i.e., *rule 1* monitored with WordNet (ODB), for appropriate meaning. In turn, it is checked with ARM i.e., *rule 2*. For undetected words *rule 3* is applied. The *KDB* maintains the detected stem words along with the domain (i.e. type of cyber threat activity). The *metadata* is checked, for identifying from and to which Email-id the suspicious words belong and other relevant information.

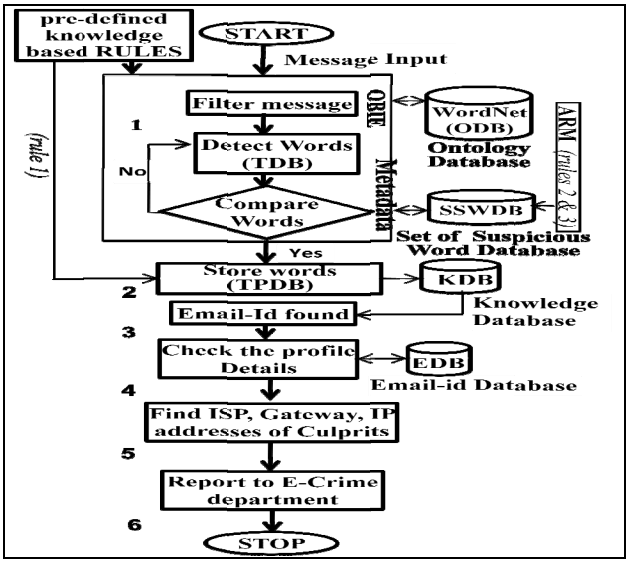

Fig. 2. Schematic cum algorithmic representation for proposed Framework for tracing criminals from instant text messages named as SPD algorithm.

3. Email-id details like Phone Number, Contact details, Company details, Age and other relevant information are traced by browsing their profiles from *EDB*, which are provided during the creation of email account, with the aid of Relational Wrapper Algorithm [12] discussed in Section III.B.
4. The suspicious messages that are sent through email-id account using which computer (IP-address), ISP address [13] and location details [14] are tracked by e-Crime monitoring system program and a report is generated which is shown in Fig. 8.
5. The generated report consists of type of cyber threat activity performed by the culprits in cyberspace, email profile details and other relevant information.
6. The E-Crime department will take action to enforce an inquiry on the report under E-Crime act [2].

The OBIE plays a crucial role that predicts and maps the domain (topic) to which these suspicious words belong. For this, we are using the databases like *(TDB, TPDB, ODB, SSWDB, KDB and Metadata)*. In this Framework, *TDB* (Transaction database) is used to store the dynamic messages that are sent between the users or clients. *ODB* (Ontology Database) is a lexical database, used to identify terms, Synonyms, Concepts, Taxonomy (concept hierarchy), relations and Axioms. *TPDB* (Text pattern database) stores the extracted stem words after filtering out unnecessary words from the messages *(TDB)*. OBIE utilizes WordNet database *(ODB)* as and when required. In *SSWDB* (set of suspicious words database) filtered suspicious words are stored based on pre-defined knowledge based rules 2 & 3, of Table I. The *SSWDB* words are dynamically compared with *TPDB* words, which are again guided by *ODB* and ARM. *ODB* is used twice to avoid ambiguity of words between *TPDB & SSWDB*. If the

 978-1-4799-2608-4/14/$31.00 ©2014 IEEE

detected words are suspicious, the e-crime monitoring system program is initiated by SPD (Suspicious Pattern Detection) algorithm, shown in Fig. 8. *EDB* (E-mail database), maintains the e-mail details of the users that shows the username, father's name, study details, job & location details, phone number and other relevant information.

Metadata is the essential component, that maintains information of all databases used, users information to whom the message belongs and other relevant information pertaining to Framework (time, date, receivers and senders details, etc.). Just like a log of history maintained by most of IMS. The pre-defined rules of Table I, specifically *rule 1* is given to OBIE, using associative rules [6], are framed carefully by analyzing brainstorming session of real time datasets that are taken from FBI and CBI investigations of solved cases [1] and GTD [25].

Table I. Shows *3 rules* to be satisfied by OBIE Model while extracting and mapping the stem words to Domain.

RULE 1 (Pre-defined Knowledge based rules)	
Type of threat activity (Domain)	**Stem words to be detected in a given context**
Murder →	kill, assault, assassinate, eliminate, gun dagger, knife, stab, location, money
Kidnap →	Hijack, capture, seize, abduct, usurp, grab, gun, take_hostage, location, amount, kill,property
Terrorist attack →	Bomb, vehicle, location, suicide_attack, bag, holy_place, laptop, demolish, payment , cash
Drug supply & Smuggling →	Packet, brown_sugar, cash, cocaine, hashish, M.tabs, Methoquoline, opium, charas, location, injection,Morphine, LSD STR/ECA, dibucaine
Match Fixation→	Location, luxurious_flat, cash, hotel, bet, gifts (car), virgin_girl,bank, payment, loose_game
Corruption charges →	Luxurious_flat, money, bank, cheque, deposit, diamond, avoid_tax, laptop, offshore_account
Robbery & theft→	Jewelry_shop, place, bank, night, gun, knife, location, vehicle, break, night, keys, locker
sexual harassment→	Phone_messages, beautiful, come, payment, spend_night, location, jewelry_items, park hotel, car_gift, body_parts, property, help,Job
RULE 2 (threshold value)	
Check the user-defined threshold value for the stem words that may belong to multiple domains, using association rule (Support and Confidence) [6]. [In OBIE, both *rules 1 & 2* are applied to Information Extraction Module (TPDB) and Ontology Editor (SSWDB) respectively, before sending to knowledge database (KDB)]	
RULE 3 (undetected words)	
Ambiguous and undetected words to be checked and corrected automatically based on nearness of stem words using ontology taxonomy constructed [5] by *rule 1*. [In OBIE this *rule 3* is applied to Information Extraction Module (TPDB) before sending to knowledge database (KDB)]	

A. Word Extraction From Unstructured Text Using Ontology

1) Information Theory principle using probability

Ontology makes use of probabilistic topic models, to organize words under a topic (domain) into a meaningful hierarchy. The *GSHL algorithm* [15], builds the concept hierarchy based on the principle of information theory [16]. Kullback-Leibler, divergence ($D_{KL}(P||Q)$)(also known as relative Entropy) is given first to underpin the principle for establishing relationships between topics [16]. Following the Gibbs inequality [16], KL divergence value is to be: $D_{KL} > 0$.

2) Topic (Domain) Hierarchy Construction Algorithm

Having, defined the information principle, for establishing a relationship between topics, the next step is to organize topics into hierarchies. Wang Wei and et al. [15] developed Global Similarity Hierarchy Learning (GSHL) algorithm. This *GSHL*, recursively searches for most similar topics of the current "root" topic and removes those that do not satisfy the condition on difference of K_L divergence. GSHL, start with an initial topic as the root node and look for top n most similar topics according to (*dis*)similarity measures. The parameters used in algorithm are as follows:

- N — The total number of topics (domains i.e. root words).
- M_c — The maximum number of sub-nodes (words) for a particular node (root node, i.e. domain/topic).
- TH_s and TH_d — The thresholds for similarity and divergence measures.
- TH_n — The noise factor, defined by the difference between two K_L divergence measures $D_{KL}(P||Q)$ and $D_{KL}(Q||P)$.
- I — Maximum number of iterations.

The parameters TH_s, TH_d, TH_n are user-specified constants, which are tuned to obtain desirable precision and accuracy values. Specifically, in our experiment, we have found that setting TH_s, TH_d, and TH_n within some narrow range results in only a slight variation of precision values. The pairwise measures of Cosine similarity, JS divergence, and KL divergence are collectively denoted as the $\boldsymbol{M_s}$ matrix. The algorithm will terminate according to the conditions specified in the while loop. The pseudo code for *GSHL* algorithm is shown in Fig. 3.

Algorithm 1. *GSHL (root)*
Require: Initialize V, Ms, I, TH_s, TH_d, TH_n, and M_c.
Ensure: A terminological ontology with "broader" and "related" relations.

```
1  Initialize V, Ms, I, THs, THd, THn, and Mc;
2  while (i < I and V is not empty) do
3    Add current root into V ;
4    Select most similar Mc nodes (words) of
5       root word (topic)
6    from Ms;
7    Add similar nodes into Vtemp;
8    Remove nodes in Vtemp against//similarity and divergence
9                                   //difference condition
10     for (all nodes n ,in Vtemp) do
11       if (Sim(ni, root) > Sim(ni, Sibling(root))) then
12         Assert broader relations between root and
13           topic ni;  //relationship between topics is broader
14       else default
15         Assert related relation between root and
16           topic ni;  //relationship between topics is nearer
17       end if
18      Move topic ni  from Vtemp to V ;
19      Increment i by 1;
20    end for
21   Remove current root from V ;
22 end while
```

Fig. 3. Shows the terminological ontology algorithm to find the root word from domain (topics) using threshold value.

3) OBIE Model for Root word (Domain) Extraction

Ontology represents knowledge as a set of concepts within a domain and finds the relationships among those concepts. It is used to reason about the words within that domain (topic) and describe the domain. The hidden suspicious words are explored from these messages and domain *(Murder, Kidnap, Terrorist attack, Drug supply and smuggling, Match fixation,*

 978-1-4799-2608-4/14/$31.00 ©2014 IEEE

corruption charges, Robbery & theft, and *Sexual harassment)* is found using ontology in our Framework. The OBIE model using ontology is shown in Fig 4.

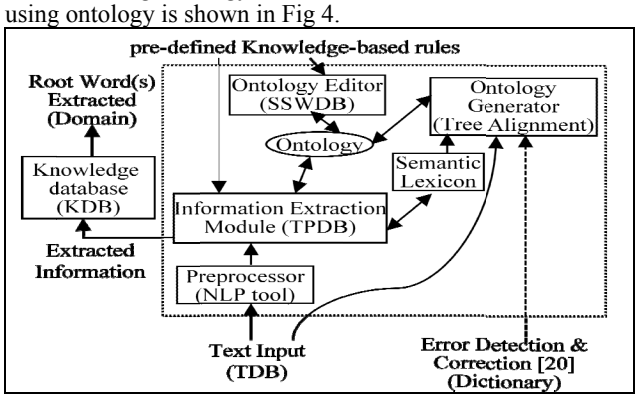

Fig. 4. General Architecture of OBIE Model for Root word extraction from unstructured text messages using NLP & Ontology.

In OBIE, these messages from Text Database (TDB) are given as input to the preprocessor component which converts the text to pure textual format. The preprocessor, uses shallow NLP (Natural Language Processing) tools. These tools perform functions such as Part-Of-Speech (POS) tagging, sentence splitting and identifying occurrences of regular expressions. The NLP tool like GATE [17], sProUT [18] and others tools are discussed in [5].

Information Extraction component task includes: Identifying processing tokens, apply stop words, characterize tokens, apply stemming algorithms and finally create searchable data structure. Filtering unnecessary words from unstructured text is done using information extraction techniques [17]. Stop list algorithm eliminates searchable processing items that are unimportant (Preposition and articles, for example "the", "of", "is", etc.) [19]. Porter Stemmer algorithm consists of a set of condition and action rules which eliminates the prefix-suffixes to find stem words, for the root (domain) from *TDB* and finally store these stem words in *TPDB*. These stem words are guided by ontology, to do this semantic lexicon for the language in concern is often used which is supported with WordNet.

The ontology is used internally by the Ontology Generator component using *TreeAlignment* algorithm which is discussed in Fig. 6. During this process it make use of semantic lexicon (WordNet), for stem words extracted and builds a Tree with empty root node as explained in next section. Subsequently, the inputs given from knowledge-based rules to Ontology Editor (SSWDB) is mapped with the Ontology Generator (Tree alignment) which uses WordNet again, to avoid ambiguity of the pre-defined words and proper mapping is done by identifying specified Domain, for the empty root node. Erroneous messages may be suspicious, which consist of errors in the words are checked and corrected using string matching algorithms [20]. The root word(s) or domain(s) which are identified from unstructured text (TDB) are finally stored in knowledge database (KDB) along with stem words.

4) Tree Alignment algorithm to find the Root word (Domain)

The suspicious messages that are communicated between the users are captured dynamically using OBIE; the output obtained is shown in Fig. 5. During the above process these words are monitored at the backend using Ontology, Ontology Editor and the *TreeAlignment* algorithm which is shown in Fig. 6, for aligning the stem words to specified root (domain).

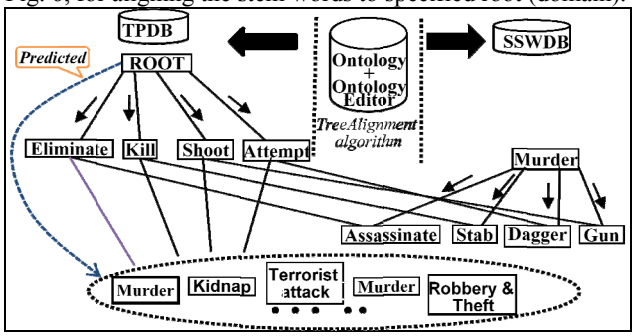

Fig. 5. Taxonomy of Suspicious words from Text Pattern Database (TPDB) mapping with set of suspicious words (SSWDB) in OBIE.

When the "Eliminate" (VERB), stem word is searched using WordNet (Ontology database), the synonyms found are "Excretion", "Killing", "Evacuation", "Demolition", "Assassinate", "Murder" and other equivalent words. Among them "*Eliminate"* is mapped to "*Assassinate"*, as shown in Fig. 5. Similarly other words that are present in *TPDB* are mapped with the *SSWDB* database. The Ontology actively guide in the construction of partial tree (restricted to 2-levels, i.e. *"parent-child"*) using algorithms in [21][22]. During this Tree building phase, the stem words *(TPDB)*, participate for mapping of words with the Domain words *(SSWDB)*, at the same time threshold value for each Topic is evaluated using *GSHL* algorithm which is shown in Fig. 3.

Algorithm 2. *TreeAlignment(RootNode,node)* for Domain (Topic) extraction using Ontology concepts
Require: Initialize *RootNode, Node(s), MinThreshold* value
Ensure: Domain (topic) extraction from stem words.

```
1  for (int i:1 to numDomainTopic {
2   //traverse tree for DomainTopic i
3    NumLevels=2;   //size of tree is 2(parent & child)
4    for (int j:1 to NumLevels){
5      RootNode[ ]Nodes=Empty // top most level 1
6      Node[ ]nodes=getNodesAtlevel(nodes);
7      //All the stem words from the TCDB at level 2
8      RootNode[]Nodes=Node[]nodes
9      //stem words assigned to RootNode
10       checkForRepetitiveData(Nodes);
11       { call GSHL(Root) }//check the stem words(TCDB)
12          //matched in other domain i.e SSWDB using WordNet
13     checkFordisjunctiveData (nodes);
14       { call GSHL(Root, MinThreshold) } //check for stem
15         //words that belongs to more than one domain (topic)
16    }//end for
17 }//end for
```

Fig. 6. Tree Alignment for classification of Domain(s) *(SSWDB)* with stem words *(TPDB)*.

In the above algorithm, *checkForRepetitiveData* function performs the task as follows: The stem words stored in *TPDB* are matched with words stored in *SSWDB*, from existing domain (topics) types. Similarly, the *checkFordisjunctivedata* function does the task of finding stem words that may exist in some other domain (topics) but, not in all the domains for this minimum threshold value is considered because some of the stem words may also fall in multiple domain (topics) that may satisfy the above user-defined threshold value.

 978-1-4799-2608-4/14/$31.00 ©2014 IEEE

B. *E-crime Monitoring System Program*

Evidence extraction and correct path discovery in IMS, is challenging task. In our Framework, the E-Crime monitoring system program, monitors these messages dynamically that are sent, if any suspicious words are embedded in these messages, such messages are traced and a report is generated with the details of words, cyber threat category type and culprit details which includes E-mail id, Phone No, ISP details, IP address, Location, etc. Finally the generated report is sent to E-Crime department, for necessary action.

For finding the complete information of the cyber attackers from *EDB*, the R2D (RDF-to-Database) relational wrapper [23] is used by e-crime monitoring system program which determines complete domain-specific, Entity-Relationship Diagram using the RDF-to-Relational Schema transformation process and then SQL queries are fired, the results obtained are depicted in Fig. 7.

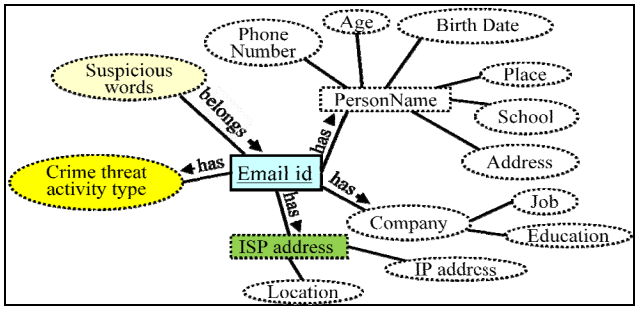

Fig. 7. Mapping details derived of culprits in cyberspace when suspicious words detected using R2D wrapper when applied to KDB.

C. *Role of SPD Algorithm*

The SPD algorithm is the backbone of our Framework as discussed earlier, in previous section that it has initiated the overall progress starting from storing text messages in *TDB* till finding the culprits by providing a detailed report from *KDB* and *EDB* databases to E-crime department when suspicious messages are found. SPD algorithmic steps are revisited again as shown in Fig. 8.

Algorithm 3. *SPD (TDB)*

Input: Instant messages stored in Text Database (TDB) (day to day) from instant messaging system (IMS)/Framework.

Output: Report to E-crime department if suspicious messages are detected.

```
1 Do{ //Apply Ontology based IE technique for filtering unnecessary
      //words and pick suspicious words (if found) and push to
      //(TPDB) which include stem words mapped with pre-defined
      //knowledge-based rules stored in (SSWDB) rule 2 discussed
      //in section III.
    Push Messages to TDB //instant messages stored in TDB
2:  Do{//Scan TDB for relevant suspicious words patterns if found
      //store it in TPDB and perform mapping with SSWDB using
      // Ontology (OBIE) building taxonomy of stem words
3     Call TreeAlignment algorithm{ // algorithm discussed in Fig 6
        // initially all stem words (TPDB) mapped to empty root
        //node using OBIE model forming a tree as shown in Fig. 5
4       Call GSHL algorithm {//Check threshold values of stem
          //words (TPDB) with root nodes stored in SSWDB using
          //algorithm shown in Fig. 3 and find the type of threat
          //activity i.e. domain topic(s); consider rule 1, 2, & 3
5         Scan TDB
6         Push patterns to TPDB //stem words
        } //end of call GSHL
7       Compare TPDB with SSWDB{
8         If TPDB==SSWDB Then Push patterns to KDB
            //stem words with Domain(s) stored
          else
            Do Nothing
9         End If
        } //end compare
      } //end  call TreeAlignment
10   }until TDB!=NULL //end of do
11  If TPDB==KDB Then //suspicious words found  with Domain
12   Check KDB {//check the knowledge database using E-crime
        // monitoring system program for type of cyber threat
        // activity (i.e suspicious stem words along with cyber
        // threat category i.e domain) using R2D relational wrapper
13      if KDB=='TRUE' then {// if suspicious words match then
14        Check EDB {//trace the profile details (emailid, phone
                    //number, ISP IP address and location details
15          Report to E-crime Department //detailed report
              //including threat activity details traced from KDB/EDB
16        }//end of check EDB
17      }//end if
18   }//end of check KDB
19 } //end of do
```

Fig. 8. Shows Overall working of our proposed Framework for identifying suspicious messages and reporting to E-crime department.

IV. Experimental Results

A. *Evaluation method for datasets*

We used Precision metric [24] to evaluate our Framework. The extracted suspicious words efficacy are based on two factors, the number of actual words available in the pre-defined database i.e. *SSWDB* with respect to domain, to that of the number of extracted words from user generated testbeds.

$$\text{Precision } (P) = \frac{\text{Correctly Extracted}}{\text{Total Extracted Correctly}}$$

$$\text{Recall}(R) = \frac{\text{Correctly Extracted}}{\text{Total No. of Possible Words}}$$

B. *Preparation of datasets and results obtained*

The Terrorist Attack (Domain) dataset is taken from Global Terrorism Database (GTD) which has recorded information on terrorist events around the world since 1970 to till date. The complete representation related to terrorist attacks is found using *CODEBOOK* [25]. We obtained dataset using brainstorming session from domain experts using GTD that consists of 59787 rows, size of 30MB, and 7 columns taken out of 99 columns, named it as User Generated Content (UGC) i.e. UGC-testbed-1 and tested with our Framework. The outputs obtained are shown in Table II.

Table II. Outputs Obtained from UGC-testbed-1 Dataset

Terms	Framework Output
Total Extracted Correctly	**1779**
Correctly Extracted	**1703**
Total Possible words extracted	**1732**
Precision	**95.72**
Recall	**98.32**

Dataset used are manually created by brain storming session from domain experts using GTD, as we could not able to get real suspicious contents that are stored in history from IM [26] and SNS [9], due to authorization restriction.

C. *Comparison of our framework with existing IMS*

978-1-4799-2608-4/14/$31.00 ©2014 IEEE

Currently none of Instant Messengers, Social Networking Sites and Mobile Phones (Apps) has the ability to detect suspicious messages during online chat. The features based on which our Framework is compared with IM/SNS/Apps (IMS) are shown in Table III.

Table III. Comparison of our Framework with (IM, SNS & Apps)

Features	IM, SNS & Apps	Proposed Framework
Cyber threat Activity Detection	Static Detection (time consumed)	Dynamic Detection
Report Generation for E-crime department	No Report	Report with details (Email-id, Phone No., etc.)
Ontology support	No	Yes
Dynamic Location Mapping based on ISP and IP address	No	Yes (using R2D Wrapper)
Efficiency	Very Good	Moderate (as online messages are monitored & stored)
Database & Data Mining support	No	Yes
System Architecture	Easy to Design	Complex to design

V. Challenges and Future Work

Framework aids the E-crime department to identify suspicious words from cyber messages and trace the suspected culprits. Currently existing Instant Messengers and Social Networking Sites lack these features of capturing significant suspicious patterns of threat activity from dynamic messages and find relationships among people, places and things during online chat, as criminals have adapted to it. The User Generated Content (UGC) testbed is proven to be useful, for monitoring terror and suspicious crimes in cyberspace which provides national and international security. We used simple English terms like *kill, murder, etc*. But, in practical scenarios these words are in specific coding language, for example "*picnic*" is used instead of *"kill"*.

Issues and challenges of our Framework are:

- If the suspicious messages are encrypted, we need to detect using decryption techniques [27].
- The suspicious words sent in short-form, code words and Steganography techniques are not detected and hence neglected as ignore words.
- Support for Multilingual languages to be included [28].
- The media may also actively participate in transmitting messages to terrorists and criminals indirectly, via news papers and TV channels (Text, Audio and Video) unknowingly [29].
- Integration with HADOOP to solve Big Data problems.

If the proposed Framework integrated with existing IM and SNS at Server-side, for surveillance will change the world of cyberspace to rest in peace without cyber crime.

References

[1] (2012). [Online]. Available: http://www.fbi.gov/sandiego/press-releases/2012/ic3-2011-internet-crime-report-released.

[2] 3GPP2 partners, "Short Message Service over IMS: 3rd Generation Partnership Project 2," developed under 3GPP2, published in 2007.

[3] (2012). [Online]. Available: [Online]. https://wikis.oracle.com/display/CommSuite7RR92909/Developing+an+Instant+Messaging+Architecture.

[4] (2012). [Online]. Available: http://www.ontologyportal.org/

[5] Daya C. Wimalasuriya, and Dejing Dou,"Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches," Journal of Information Science,Volume 36, No. 3, pp. 306-323, 2010.

[6] David W. Cheung, and et al., "Maintenance of discovered association rules in largedatabases: an incremental updating technique," published by IEEE in 1996.

[7] M. Mahmood Ali, and L. Rajamani, "Framework for surveillance of instant messages, " published by inderscience in IJITST, vol. 5, 2013.

[8] Michael Robertson, Yin Pan, and Bo Yuan, "A Social Approach to Security: Using Social Networks to Help Detect Malicious Web Content," published by IEEE in 2010.

[9] (2012). [Online]. Available: http://www.digitaltrends.com/social-media/facebook -scans-chats-and-comments-looking-for-criminal-behavior/

[10] Appavu, and et al.,"Data mining based intelligent analysis of threatening e-mail," published by Elsevier in knowledge-based systems in 2009.

[11] M. Mahmood Ali, and L. Rajamani, "Phishing Detection in Instant Messengers using Data Mining Approach," proceedings of ObCom 2011, published by Springer-Verlag Berlin Heidelberg 2012, part I, CCIS 269, pp. 490–502, 2012.

[12] Sunitha Ramanujam, and et al., "A Relational Wrapper for RDF Reification," E. Ferrari et al. (Eds.): TM 2009, IFIP AICT 300, pp. 196–214, IFIP International Federation for Information Processing 2009.

[13] (2012). [Online]. Available: http://www.iplocation.net/index.php

[14] Satyen Abrol, Latifur Khan, and Tahseen Al-khateeb, "MapIt: Smarter Searches using Location Driven Knowledge Discovery and Mining," Proc. of 1st SIGSPATIAL ACM GIS 2009 International Workshop on Querying and Mining Uncertain Spatio-Temporal Data, in conjunction with ACM GIS 2009, Seattle, Washington, November, 2009.

[15] Wang Wei, Payam Barnaghi, and Andrzej Bargiela, "Probabilistic Topic Models for Learning terminological ontologies," published by IEEE Tran., on Knowledge and data engineering, vol 22, no. 7 in july, 2010.

[16] D.J. MacKay, "Information Theory, Inference, and Learning Algorithms," Cambridge University Press, 2003.

[17] H. Cunningham, Information Extraction, Automatic, Encyclopedia of Language and Linguistics, second edition Elsevier Science, 2005.

[18] Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorskiand, and Ulrich Schafer, "SProUT–a general-purpose NLP framework integrating finite-state and unification-based grammar formalisms," published by springer in Finite-State Methods and Natural Language Processing (LNCS) Volume 4002, pp. 302-303, 2006,.

[19] (2012). [Online]. Available: http://www.webconfs.com/stop-words.php

[20] M.W.Du, and S.C.Chang, "An Approach to Designing Very Fast Approximate String Matching algorithms," IEEE journal, 1994.

[21] Y. Zhai and B. Liu, "Web data extraction based on partial tree alignment," in Prococeeding of ACM, 2005.

[22] Jer Lang Hong, "Data Extraction for Deep Web Using WordNet," published by IEEE Transactions on systems, man and cybernetics, 2011.

[23] Sunitha Ramanujam, and et al., "A Relational Wrapper for RDF Reification," E. Ferrari et al. (Eds.): TM 2009, IFIP AICT 300, pp. 196–214, IFIP International Federation for Information Processing 2009.

[24] C.D. Manning, P. Raghavan, and H. Schutze, Introduction to Information Retrieval, Cambridge Univ. Press, 2008.

[25] (2013). [Online].http://www.start.umd.edu/gtd/dowloads/codebook.pdf.

[26] (2012). [Online]. Available: http://dir.yahoo.com/Society_and_Culture/Crime/Types_ of_Crime/

[27] E. Thambiraja,G. Ramesh, and Uma Rani, "A Survey on Various Most Common Encryption Techniques," published by IJARCSSE Journal volume 2 issue 7, pp. 226-233, 2012.

[28] M. Mahmood Ali, and Lakshmi Rajamani, "Framework for Surveillance of Emails to Detect Multilingual Spam and Suspicious Messages," IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions , IIT Kanpur, India, pp. 42-56, 2013.

[29] Chaditsa Poulatova, "The Media: A Terrorist Tool or a Silent Ally," published by IEEE in 2011.