ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

An Upgraded Identity Deception Detection Technique based on Non-Verbal Behavior Using PSO Approach

M. Preensta Ebenazer[1], Dr. P. Sumathi[2]

[1]Research Scholar, PG & Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India

[2]Assistant Professor, PG & Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, India

ABSTRACT: Identity deception plays a significant role in real world applications, especially in online social applications. A number of users attempt to create multiple accounts and may involve in performing malicious activities. In this work, Wikipedia home pages are considered for the detection of identity deception where the concerns regarding fake users matters a lot. In the existing work, non-verbal behaviour based deception detection was introduced which attempted to detect fake identities based on time window. However this work lacks from efficient detection of deception accounts due to the absence of selecting an optimal time window process. Larger time window resulted in more time consumption for detecting an identity deception. This problem is resolved in this approach by introducing an optimal time window selection process through which efficient detection of identity deception can be achieved with limited time period. An optimal result is achieved using the Particle Swarm Optimization approach which attempts to select the most optimal set of variables based on the time window selection process. The main objective of the proposed research is to achieve reduced time consumption by selecting the most optimal time window. Experimental tests were conducted which proves that the proposed approach provides better result than the existing methodology in terms of increased accuracy and reduced time complexity.

KEYWORDS: Fake users, Wikipedia, identity deception, time window and optimal time window variable

I. Introduction

Online social media application is one of the emerging fields in real world applications. Numbers of users who make use of online web sites have increased enormously. Sock puppetry is one of the main issues in online social media application where the users maintaining multiple accounts and involving in malicious activities have increased tremendously. Wikipedia is an online social media web site in which users are allowed to write their articles. Wikipedia comprises of namespaces which can be updated by various users. Users are allowed to write new articles, edit the articles, include or remove the contents from the web site and so on [4]. However fake users detection is the main issue in Wikipedia where multiple users attempt to alter the contents that are stored in the Wikipedia pages.

Various approaches were introduced and discussed in the previous research works in terms of finding the malicious behavior involved in the Wikipedia pages, so that identity deception can be identified well. There are two main ways that attempt to identify the deception activities that are present in online applications. They are

- Verbal behavior based detection
- Non-verbal behavior based detection

Most of the research works focused only on the verbal behavior of the deception detection process. Verbal based identity deception makes use of the online user’s verbal behavior like their way of writing in the articles, their profile names, grammatical uses, and way of article concept explanation and so on. Non-verbal based behavior detection is the process of finding identity deception by considering the non verbal behaviors like time duration, amount of contents modified, number of users accessing at the same time and so on [4]. Verbal based behavior leads to efficient prediction of the identity deception activities involved in the system in a considerable manner. However immense improvements

DOI: 10.15680/IJIRCCE.2015. 0311191

in technologies in the environment made users to escape from the deception detection process that focused on verbal behavior. This approach overcomes the flaws by considering the non verbal behavior of the users.

The main contribution of this research work is given as follows:

- Identity deception activities involved in the Wikipedia pages are identified in terms of considering more number of parameter values
- Optimal time window selection is introduced by using a methodology called particle swarm optimization which is used to select the time window at run time based on user requirements.

The organization of this work is given as follows: In section 2, an overview of the previous research works is given. In section 3, proposed methodology of this work is unraveled with the detailed flow. In section 4, performance evaluation conducted using the java simulation environment to prove the improvement of the proposed research work is disclosed. In section 5, overall conclusion of this research work is elucidated.

II. RELATED WORK

In this section, various previous research methodologies that have been conducted to achieve the efficient detection of identity deception present in network environments is illustrated. They are as follows:

Alan Wang et al [1] proposed a novel mechanism for automatically detecting deception detection activities in the criminal records. This is done by proposing a novel adaptive analysis method that handled the missing values present in the data base. Missing values are found by analysing and finding the statistical relationship present between the various statistical values. The main goal of this research work is to find the deception detection present in the criminal data base in terms of various functional parameter values. It is done in terms of a performance measure called the accuracy to find the improvement.

Racha Ajami et al [6] expounded various privacy issues which occur in the mobile social networks. Mobile social networks constituted an improved growth in the industry sector which ensured ease of use and more flexibility to the users. These social networking web sites made use of mobile contacts, photos and videos to improve the ease of access. This information access permission leads to a privacy violation of the user in terms of releasing their personal information. These issues were addressed for providing assured framework to the users, so that they can enjoy the social networking feature with assured privacy level.

Mauro Conti et al [2] introduced a novel approach for detecting fake profiles that resided in the face book media networking web sites in terms of malicious behaviours of the fake profile users. This mechanism ensured that fake profiles can be detected in a considerable manner by considering different factors. This work proposed a methodology to find the security threats by constructing the graph which depicted the link between different factors of attributes.

Sadia Afroz et al [7] described a new technique that can find the most optimal set of deception behaviours that are present in the environment. This approach finds the deception activities in terms of the writing styles of the users who are involved in the system. The writing styles of user are different in different terminologies and this is analysed in this approach for detecting the unusual behaviour of users in an efficient manner.

Melissa et al [3] introduced a technique called improved optimization methodology which attempted to detect the optimal behaviour by using fusion technology. Fusion technology is the aggregation of the concepts that are involved in the system in terms of different identity parameter values. This aggregation is done, so that the values that are produced by invalid users can be found easily. This aggregation concept is based on generating a valid proof of valid users present in the environment.

Michail Tsikerdekis [5] introduced a new methodology for detecting online deception activities present in the social network in terms of verbal behaviour of users. This methodology attempts to predict the online deception behaviour by using the parameter like their identity, age, address, ip details and so on. This approach is used to identify the online deception behaviour present in the environment in terms of different valuable parameters.

III. AN OPTIMIZED MULTIPLE ACCOUNT IDENTITY DECEPTION

This proposed work concentrates on detection of deception activities that resides in various social networking media web sites in terms of user’s malicious behaviour. This work mainly focuses on finding the malicious behaviour from Wikipedia web sites in terms of malicious revisions that are made by users. Identity deception process of the proposed methodology is explained as follows:

- Initially data is gathered from Wikipedia related pages and pre-processed to improve the performance level

- Non verbal behaviour present in the pages are gathered from the Wikipedia related articles
- Optimal time window selection is done based on PSO approach
- Improved detection of deception activities are carried out using random forest approach

The overall system flow of this research work is given as follows

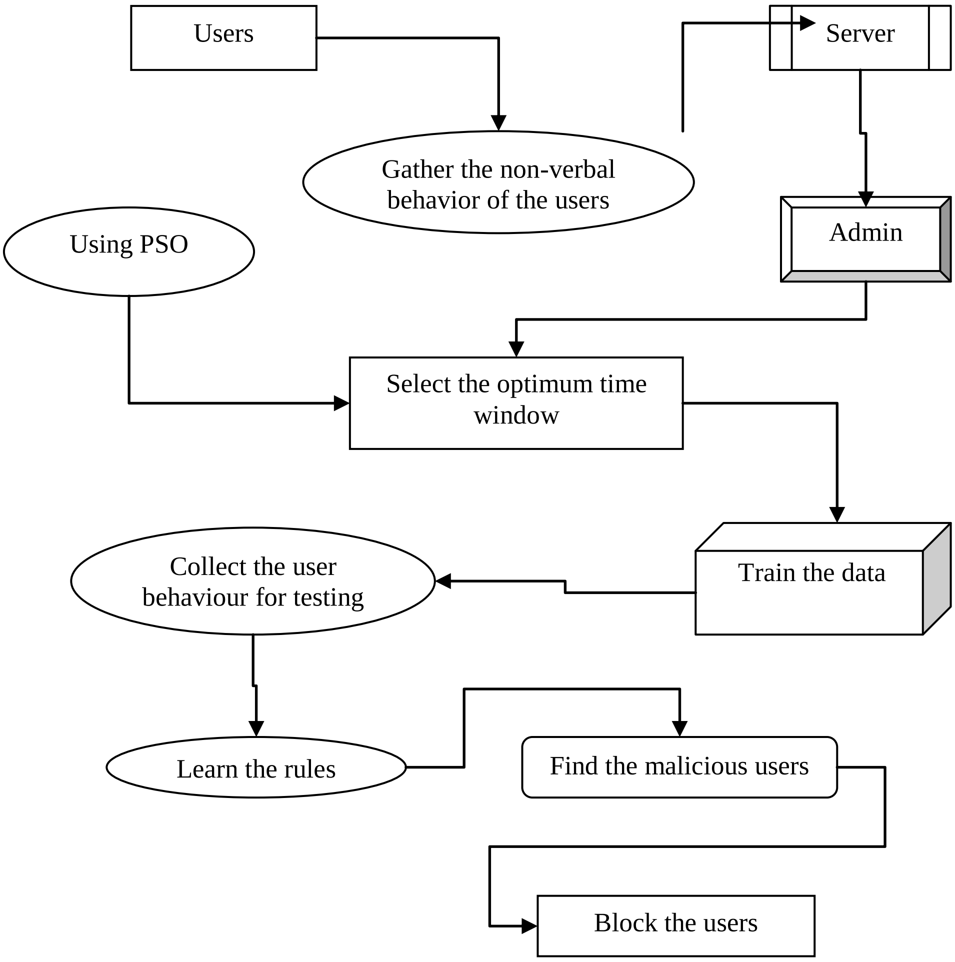
Users
Server
Gather the non-verbal behavior of the users
Admin
Using PSO
Select the optimum time window
Train the data
Collect the user behaviour for testing
Learn the rules
Find the malicious users
Block the users

Figure 1 Overall flow of the identity deception process

The above diagram provides a flexible and efficient detection of the deception behaviours present in the network environment in terms of improved system performance and security level. In the following sub section detailed explanation of the proposed approach is given

A. *PSO Based Optimized Identity Deception Detection:*

In this approach, Wikipedia data set is considered for efficient processing and finding the malicious users involving in some malicious activities. The proposed work focuses on non verbal behaviour of the users who are accessing the Wikipedia web pages. The gathered data from the Wikipedia web pages consist of many unwanted data's and missing values in the data set. These are initially pre-processed before proceeding with the detection of deception behaviours involved in the system. Pre-processing comprises of the two steps. Those are

- Filtering

DOI: 10.15680/IJIRCCE.2015. 0311191

- Normalization

Filtering is the process of finding more similar data that are present in the data set and avoids the dissimilar data. Filtering is done by finding the most similar data in terms of calculating the Euclidean distance between different data items present in the data set. After grouping the similar kind of data, normalization is done to fill the missing values present in the data set. Missing values are found by calculating the average value of the remaining similar data items which are grouped in the previous step called filtering process. After pre-processing, the complete representation of the data set is obtained without any missing or corrupted data in the data set.

From the pre-processed data set non verbal behaviour of the users are then found to predict whether the corresponding user is involved in the deception activities or not. The non verbal behaviour of the users is classified into two categories based on timing parameter. They are time dependant parameter and time independent parameter values. Time dependant parameter variable values are the one in which the user behaviour is predicted in a particular time period window. Time independent parameter values are the one in which general user behaviour is predicted independent of timing values. Initially, this process starts with the initial registration of users in terms of the time at which they enter and the number of articles revisioned by them in a particular time period. These revisions are calculated for various number of namespaces such as article(R_{at}), article discussion(R_{dt}), user page(R_{ut}), and user discussion page (Rt_t) Wikipedia-related pages and Wikipedia-related discussion pages are combined under one variable(RW_t). The divergence between these coefficients is calculated to know the variability between the different namespace attribute values by using a co-efficient called Gini index. The formula that uses gini index for calculating the divergence between the different names spaces are given as follows:

$$GR_t = 100[\frac{2\sum_{x=1}^{6}(w_x Rx_t \sum_{j=1}^{x} w_j) - \sum_{x=1}^{6} w_x^2 Rx_t}{(\sum_{x=1}^{6} w_x)\sum_{x=1}^{6}(w_x rx_t)} - 1$$

Where

x = set of revision data for each name space

w = relevance weight value assigned for each name space

Another non verbal based behaviour used in this approach is the addition and deletion of the contents that are present in the Wikipedia articles. The calculation of total number of bytes that is added and deleted in each revision in a particular time period by users is calculated as follows:

$$\widehat{RB}_t = \frac{\sum_{i=1}^{R_t} RB_i}{R_t}$$

Another parameter that is considered is the time duration spent between every revision approximately by the users. This average duration taken between every time revisions are calculated as follows:

$$AD_t = \frac{\sum_{i=2}^{R_t} T_i - T_{i-1}}{R_t}$$

Where

n = total number of revisions

T = set of all unit time for each revision which was made

After extracting the time duration spent by every individual user in a particular time period in all revision, PSO approach is introduced for selecting the optimal time window. Particle swarm optimization is an evolutionary algorithm which is used to optimize the problem by iteratively improving the candidate solution with improved fitness values. PSO approach provides flexible environment for this process and is used to select the most optimal particle from the set of particles. The best value is found by updating the position and velocity of each particle in terms of updation in the fitness value of the particles i.e time duration.

Let N be the number of particles in the swarm, each having a position in the search-space and velocity. In every iteration process, each candidate solution is calculated by the objective function being optimized, deciding the fitness of that solution. Fitness is computed by,

Fitness 1= Min (False Positive rate)
Fitness 2 =Min (Detection time)
Fitness= {Fitness 1, Fitness 2}

Fitness exists if there is less false positive rate with less time window. The time window values are taken randomly. Then, the existing method is used to detect malicious users in less time and also with high accuracy. These fitness values are expected to move the swarm towards the best solutions. Based on this algorithm, the optimal time window is used

Algorithm
Input: Wikipedia data set
Output: Optimal time window

```
1.  Pre-process the data set
        a.  Filter the data
        b.  Normalize the data
2.  Extract the non verbal user behavior parameter values
3.  Initialize N number of particles in the swarm, each particle having a position $X_i$ and velocity$V_i$.
4.  Let pbest be the best known position of particle i and gbest is the best known position of the entire swarm
5.  Initialize the particle's position $X_i$
6.  For each particle i=1, 2… N
        a.  Calculate fitness value for every particle
                i.  Fitness 1= Min (False Positive rate)
               ii.  Fitness 2 =Min (Detection time)
              iii.  Fitness={ Fitness1, Fitness2}
7.  End for
8.  If fitness value is better than the best fitness value (pBest)
        a.  Set current value as the new pBest
9.  End if
10. Until a termination criterion is met
        a.  Select the particle with best fitness value of all particles as the gbest
        b.  For every particle
                i.  // Calculation of particle velocity
                    $V_i(t+1) = wv_i(t) + c_1 r_1[\hat{x}_i(t) - x_i(t)] + c_2 r_2[g(t) - x_i(t)]$
        c.  End for
        d.  Update particle position
                $x_i(t+1) = x_i(t) + v_i(t+1).$
11. Until some stopping condition is met
```

Where
i = index of the particle
$v_i(t)$ = the velocity of particle i at time t
$x_i(t)$ = the position of particle i at time t
w, c1, and c2 = coefficients

The above algorithm predicts the optimal time window period that can be taken to identify the deception activities that are involved in the system. By using this optimal time window period, better identification of deception behaviour is done by using random forest classification approach. From this work, it can be concluded that the proposed approach can find the deception activities that are present in the Wikipedia application that are enforced by the malicious users involved in the system. The malicious behaviour is found by analyzing and comparing the divergence between the multiple revisions that are made by users in different time period. The effectiveness of the proposed research work is evaluated by comparing it with the existing approach results in terms of various performance measures.

IV. SIMULATION RESULTS

The existing and proposed methodologies performances are analyzed and implemented by using efficient techniques. The performance metrics that are considered are precision, recall, accuracy and time. By using the proposed algorithm of PSO, it bestows an optimal method to provide efficient results

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

The dataset for evaluating the proposed research scenario is the Wikipedia dataset. This data set consists of attributes like user id, number of revisions made by time, time period consumed by every user for making changes in the articles, and number of bytes added or removed by the users in every revision. This data set is processed by both existing and the proposed system to evaluate their performance improvements. In the following sub sections, detailed description of the performance measures and its comparisons are mentioned.

A. Recall

Recall is defined as the correct predictions that are made among the total number of predictions done by the users. Recall of the proposed research work should be more than the existing scenario for better performance. The following Table 1 depicts the values that are obtained while evaluating the proposed and existing mechanism.

Number of users	Recall	
	Multiple account identity deception detection	PSO based optimal identity deception detection
5	0.49	0.7
10	0.5	0.7
15	0.5	0.7
20	0.5	0.7
25	0.51	0.7
30	0.51	0.7
35	0.55	0.71
40	0.59	0.75
45	0.6	0.72
50	0.61	0.7

Table 1 Recall Comparison Value

The graphical representation of the above mentioned values are represented as follows:

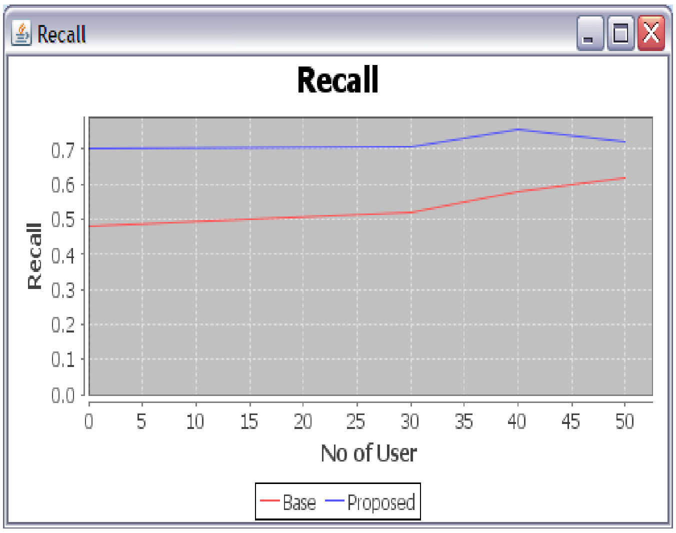

Figure 2 Recall comparison

From Figure 2, it is observed that the proposed work provides higher recall values using PSO algorithm. In x axis the methods are plotted and in y axis the recall values are plotted. It can be observed that the existing system's recall values are less than the proposed system's recall values. Hence it can be concluded that the proposed method shows an improvement in terms of recall values.

B. Precision

Precision is defined as the correct prediction of the deception activities from the set of all available Wikipedia revision related data. Precision of the proposed research approach should be higher than the existing approach for better system performance. The following Table 2 depicts the values that are obtained while evaluating the proposed and existing mechanism.

Number of users	Precision	
	Multiple account identity deception detection	PSO cased optimal identity deception detection
5	0.54	0.9
10	0.56	0.88
15	0.61	0.84
20	0.60	0.8
25	0.62	0.79
30	0.62	0.78
35	0.68	0.76
40	0.75	0.75
45	0.75	0.79
50	0.75	0.78

Table 2.Precision Comparison Values

The graphical representation of the above mentioned values are represented as follows:

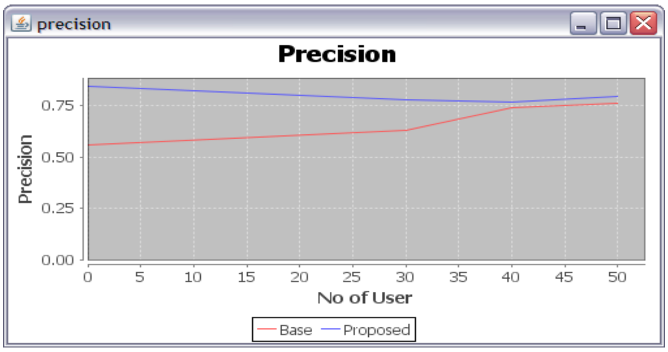

Figure 3 Comparison of Precision

From Figure 3, it is observed that the proposed system provides higher precision values using PSO algorithm. In x axis the methods are plotted and in y axis the precision values are plotted It can be observed that the existing system's precision values are less than the proposed system's precision values. Hence it can be concluded that the proposed method shows an improvement in terms of precision.

C. Accuracy

Accuracy is defined as the amount of correct predictions of deception detection in Wikipedia related articles. The accuracy of the proposed research work should be more than the existing work for the better system performance improvement. The following Table 3 depicts the values that are obtained while evaluating the proposed and existing mechanism.

Method	Accuracy
Multiple account identity deception detection	77
PSO cased optimal identity deception detection	88

Table 3 Accuracy Comparison Values

The graphical representation of the comparison is represented as follows:

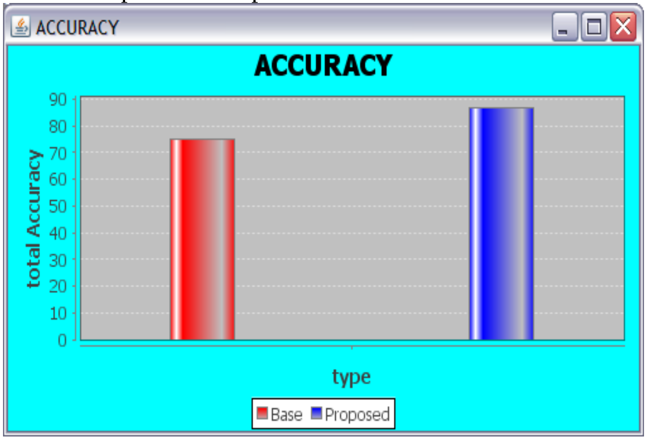

Figure 4 Comparison of Accurac
y

From Figure 4, it is observed that the proposed method provides higher accuracy values using PSO algorithm. In x axis the methods are plotted and in y axis the accuracy values are plotted. It can be observed that the existing system's accuracy values are less than the proposed system's accuracy values.. Hence it can be concluded that the proposed method shows an improvement over the existing method in terms of accuracy.

D. Time Complexity

Execution time is defined as the total time consumed for detection of the identity deception activities. Identity deception detection time should be less in the proposed work than the existing scenario. The following Table 4 depicts the values that are obtained while evaluating the proposed and existing mechanism.

Method	Time (Milliseconds)
Multiple account identity deception detection	2700
PSO cased optimal identity deception detection	2100

Table 4 Time Complexity Comparison Values

The graphical representation of the evaluation is represented as follows:

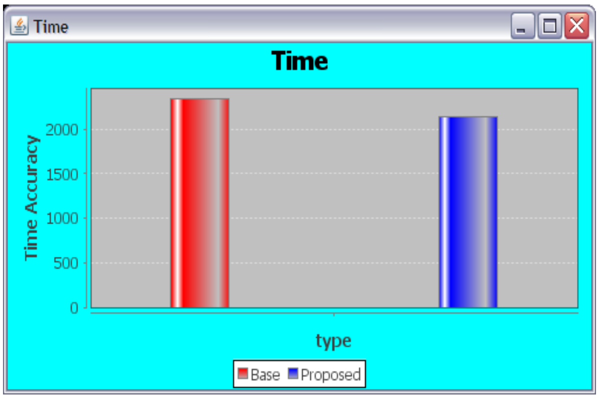

Figure 5 Comparison of Time Complexity

From Figure 5, it is observed that the proposed work provides lower time complexity values using PSO algorithm. In x axis the methods are plotted and in y axis the time values are plotted. It can be deduced that the existing method's time complexity is higher than the proposed method's time complexity values. Hence it can be concluded that the proposed method shows an improvement in terms of time complexity

V. Conclusion and Future Work

Identity deception plays a major role in the social media applications which attempt to provide a secured environment for the users. In this work, a novel optimized PSO based approach is used for finding identity deception activities present in the Wikipedia application. This approach attempts to select an optimal window size for detecting the deception activities in an accurate manner. This approach provides a flexible mechanism for detecting deception activities from a fake user's behavior in a secured manner. The experimental tests conducted prove that the proposed mechanism provides better results than the existing methodology.

References

1. Alan Wang, Hsinchun Chen, Jennifer J. Xu, and Homa Atabakhsh, "Automatically Detecting Criminal Identity Deception: An Adaptive Detection Algorithm", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans, Vol. 36, Issue No. 5, September 2006
2. Mauro Conti, Radha Poovendran, Marco Secchiero, "FakeBook: Detecting Fake Profiles in On-line Social Networks", IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, May 2012
3. Melissa C. Zoepfl and Harold J. Korves, "Improving Identity Discovery through Fusion", IEEE Computer Society, May 2009
4. Michail Tsikerdekis and Sherali Zeadally, "Multiple Account Identity Deception Detection in Social Media Using Nonverbal Behaviour", IEEE Transactions on Information Forensics and Security, Vol. 9, Issue No. 8, August 2014
5. Michail Tsikerdekis and Sherali Zeadally, "Detecting and Preventing Online Identity Deception in Social Networking Services", IEEE Computer Society, Jul 2015
6. Racha Ajami, , Nabeel Al Qirim , Noha Ramadan, "Privacy Issues in Mobile Social Networks", Procedia Computer Science, Volume 10, Oct 2012
7. Sadia Afroz, Michael Brennan and Rachel Greensta, "Detecting Hoaxes, Frauds, and Deception in Writing Style Online",IEEE Symposium on Security and Privacy, Feb 2012

BIOGRAPHY

Dr.P.Sumathi is working as an Assistant Professor in the Department of Computer Science, Government Arts College at Coimbatore. She completed PhD in the area of Grid Computing at Bharathiar University. She completed M.Phil in the area of Software Engineering at Mother Teresa Women's University. She completed MCA at Kongu Engineering College at Perundurai. She has published many national and International journals. She has about seventeen years of teaching and research experience. Her research interests include Data Mining, Distributed Computing and Software Engineering.

Ms. M Preensta Ebenazer received MCA degree from V.L.B Janakiammal Collegr of Engineering, Coimbatore, India. She is pursuing Full Time M. Phil in the department of Computer Science under the guidance of Dr. P. Sumathi at Government Arts College (Autonomous), affiliated to Bharathiar University, Coimbatore, India. Her area of research is Information Security. Her research interests include Computer Networks, Information Security and Software Engineering. She has presented 2 papers in National Conferences