Journal of Theoretical and Applied Information Technology
15th November 2017. Vol.95. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SPAM DETECTION ISSUES AND SPAM IDENTIFICATION OF FAKE PROFILES ON SOCIAL NETWORKS

[1]BALOGUN ABIODUN KAMORU, [2]AZMI BIN JAAFAR OMAR, [3]MARZANAH A.JABAR, [4]MASRAH AZRIFAH AZMI MURAD, [5]ABDULMAJID B. UMAR

[1,2,3,4,5]Dept of Software Engineering and information Systems, Faculty of Computer science and Information Technology, Universiti Putra Malaysia Serdang 43300.Selangor.D.E.Malaysia

E-mail: [1]balogun@consultant.com, [2]azmij@upm.edu.my, [3]marzanah@upm.edu.my, [4]masrah@upm.edu.my, [5]bumar05@yahoo.com

ABSTRACT

Spam has been a major and global threat, Social networks have become our daily live and everyday tools, while different social networks have different target groups. With the rapid growth of social networks, people tend to misuse them for unethical and illegal conducts, fraud and phishing. Creation of a fake profile becomes such adversary effect which is difficult to identify without appropriate research. The current solutions that have been practically developed and theorized to solve this issue of spam detection issue and spam identification of fake profiles, primarily considered the characteristics and the social network ties of the user' social profile. However, when it comes to social networks like Facebook, Twitter, SinaWeibo, Myspace, Tagged and LinkedIn such a behavioural observations are highly restrictive in publicly available profile data for the users by the privacy policies. The limited publicly available profile data of social networks makes it ineligible in applying the existing approaches and techniques in fake profile spam identification. Therefore, there is a need to conduct targeted research on identifying approaches for fake profile spam identification on selected and available data set of Facebook, Twitter and Sina weibo. In this research, we identify the minimal set of profile data that are necessary for identifying Fake profiles in Facebook, Twitter and Sina weibo and identifying the appropriate data mining approach and techniques for such task. We demonstrate that with limited profile data our approach can identify the fake profile with 84 % accuracy and 2.44 % false negative, which is comparable to the results obtained by other existing approaches based on the larger data set and more profile information.

Keywords: *Social Networks, Fake Profile, Spam Detection, Principle Component Analysis, Spam Identification*

1. INTRODUCTION

Spam is one of the diverse sorts of fraud unanswering these days. Social networks have been part of our daily activities. In recent past, social networks have made a drastic change in the social life and it changed the web into "social web" where users and their communities are the centres for online growth, commerce, and information sharing Rheingold,2000[1]. Social networks have a unique value chain which targets different user segments. To find an old friend, we used to peruse Facebook, but if it is to access micro blogging then we have Twitter. Sina weibo is mostly used by the Chinese people, Facebook is not popular in China due to government policy.

The surge of social networks' popularity and the availability of large amount of information from users' email addresses to their personal messages make them easy targets to the adversaries. Most of these targets focus on retrieving user information without user consent. For that, by intruding into the user profile or connecting with the user through fake profiles are considered as the mostly practised techniques Fire et al,2012 [2].

The advancement of social networks security it become tremendously difficult to infringing into social networks. Resultantly, now adversaries create fake profiles to get the access to other accounts. According to Statistics from Cloudmark around 20-40 % of the Facebook account could be fake profiles and this is fairly similar with Twitter and Sina weibo K.Lee, et al,2011 [3]. Due to high amount of users involvement and millions of daily transactions. It becomes hard to detect effort suspicious user behaviours in the network and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

separate them from the legitimate users. Conversely, effort is taken to find those malicious accounts and flag them as fake, yet it did not achieve the results as expected. This becomes more complex, attributable to, constricted user privacy policies, restriction for data collections and difficult to distinguish between the fake and the legitimate profiles. Even though most of the previous research scholars targeted on identifying profile cloning, spam information distribution, and intrusion detection Kontaxis et al,2011 [4]; Fire et al,2012 [2], now it becomes the right time to draw extra attention to finding solutions to differentiate legitimate and fake profiles in a sensible manner.

According to Andronicus Akinyelu et al,2014,[5] clinched that spam email and phishing has become a sombre threat to total sanctuary and frugality. Spam detection and phishing detection slant which has liked the contemporary fissure branded in the prose, their tactics conceded high classification correctness of 99.7% with trifling deceitful optimistic proportion of nearby 0.06 %. Owing to the hasty revolution in phishing attack outlines, up-to- date phishing concealment skills prerequisite to be momentously enriched to meriticoulsy warfare embryonic phishing attacks.

According to Nitin Jindal et al,2008 [6], they are generally three types of spam reviews:

Type 1 (Untruthful opinions): Those that deliberately mislead readers or opinion mining systems by giving undeserving positive reviews. It is also refers to as Fake reviews or bogus reviews.

Type 2 (reviews on brand only) : Those that do not comment on the products in reviews specifically for the products but only the brands.

Type 3 (non-reviews): Those that are non-reviews, which two main sub-types : (1) advertisement and (2) other irrelevant reviews containing no opinions example Questions, answers and random text.

1.1 Background

A fake profile is a social network of a person who maintains a false identity in the internet to pretend as someone else. It I found out by Krombholz et al,2012 [7] the fake user behaviour is different from the legitimate users. Therefore, the amount and the type of information that a fake user pass into their profile have a clear discrepancy from the legitimate user. Among the several ways of creating fake profile[7] which is used to increase the discernibility of niche content and manipulate the attraction towards the profile Cao et al, 2012 [8]; Bilge et al,2009 [9]. Next method is profile cloning Jin et al,2011 [10]; Kontaxis et al,2011 [11] where offender creates a similar profile of the legitimate users in the same or another social network by copying the victim's profile and adding victim's friend into new fake profile. The last method is creating a profile with a fake identity [7]. The trick on such profiles is perpetrators first attain the victims trust and confidence, and then cheat on them by collecting the confidential information. Due to the similarity of the features of legitimate and such fake identities, it is immensely difficult to distinguish them without ascertaining reliable data.

Facebook, Twitter and Sina weibo considered as the highly recommended social network site for everybody, irrespective of their continent and race. One of the latest scenario was, in the beginning of 2016 some hackers executed a botnet and Rat attack and created thousands of fake Facebook, Twitter and Sina weibo Fake profiles. The present process of identifying fake profiles in three social networks is limited to manual reporting of such profiles. When a user profile suspects a particular profile to be faked, he or she can use fake profile flagging option to notify the each platform about these.

1.2 Problem Statement

The profile data in social network consist of two main parts, static and dynamic. Former is about the information which is set by the user statically, while the latter is observed by the system and is the result of users' activity on the social network. The static data typically includes users' demographics and interests, and dynamic data relates to user activities and position in the social network [12]. Most of the existing research solutions depend on both static and dynamic data, which is inapplicable to other social networks, where it has merely a less number of visible static profiles and no dynamic profile details to the public. Due to its privacy policies and very restricted information visibility [13], none of the existing practical and theoretical means of fake profile detections are feasible to apply. Therefore, in this research our goal is to identify an approach to determine the legitimate profiles and fake profiles in Facebook, Twitter and Sina weibo. The focus of this research is recognizing and differentiating the legitimate profile , spam identification and fake profile in Facebook , Twitter and Sina weibo. Most of the solutions developed to address the above issue is based on Level of accuracy, Mail Ranking, web classification, intrusion detection, malware detection and spam detection on prevention but it has not resolve the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

issues and challenges. These social networks have rich and fully functional Application Programming Interfaces (API) to acquire relevant, real-time and up-to-date user information in analogous to the research requirement. Facebook API facilitates to access profile information like user activities, friends activities, friends of friends and most of the basic user details (age, birthday, profile status, relationship, status, likes, group details etc). Similarity, Twitter API provides twitter counts, followers, notifications, friends, basic user details etc. Sina weibo, API Interfaces, friends, Follows and Notifications.

1.3 Challenges and Issues

The key issue and challenges that we come across doing this research are data collection and fake profiles identification of Facebook, Twitter and Sina. Due to Limited privacy policies of the three Social networks, gathering data from it is highly restricted. We can access very limited profile characteristics and number of profiles via its API. Even to access certain basic Social networks user information such as education details, date of birth, suggestions, telephone number, total number of connections or skills and expertise we need the user permission. In addition, through the social networks API we are offered only to access first degree level user information yet with the normal web user interface, it provides access up to the third degree level of connections' information.

The access to actual fake profiles in Social networks context is greatly unattainable. However, we were able to find a list of web sources where certain Facebook, Twitter and Sina fake profiles have been manually identified and listed. So ,unlike previous research, in this research we have used only the authentic fake profile data for research instead of simulating the fake profiles.

1.4 Methodology

In this research, we considered six data mining techniques, Neural Network (NN), Support Vector Machine(SVM), Jrip, Naives Bayes, Decision tree (J48) and Principal Component Analysis (PCA). All these are well known and commonly used data mining techniques. In much social network research, Neural network and SVM are adopted as the principle mining techniques. Few such research areas are spam message identification, profile cloning and intruder detection. PCA is applied to reduce the number of dimensions of the data sets Jollife,2005 [14].

In summary, we have developed a technique for identifying the approaches to identify in Social networks by combining multiple data mining techniques. We have compared and discovered the appropriate data mining technique to identify fake profiles in social networks with minimal amounts of profile data. We have demonstrated that our approach performs with accuracy of 84% and False negative of 2.44 %, which is comparable to the results reported by existing research-that is based on other social networks data and much in-depth profile data.

The rest of the paper is organized as follows. Section Two Provides an over view of the research carried out in related to Facebook, Twitter and Sina weibo network and prior research on spam identification and fake profile identification. In section three, we describe the Facebook data set, Twitter data set and Sina weibo data set and the mechanism followed to collect data. Section four explains the methods used in the construction and evaluation of the each technique and their results. In section five we discuss the overall comparison of the accuracy rates. Section six identifies limitation of the study and future directions and finally, section seven we present our conclusion from this study.

2. RELATED WORK

In this section we provide some insight into the existing on Facebook Network, Twitter Network and Sina weibo Network with an example of cloning attack identification. Moreover, we describe existing work carried out in related to fake profile identification and in similar research background..

2.1 Research Based on Social Network data

So far, little research has been carried out accounting Facebook, Twitter and Sina weibo as the primary data source. Hsieh et al,2013[15] have conducted a research on different social networks to understand the probability of connections between two people based on their organizational overlap. Xiang et al,2010 [16] have used interaction activity and similarity of user profiles to develop an unsupervised model to estimate friendship strength. They evaluated the system on proprietary data from social networks.

2.2 Determine Fake Profile

In the midst of the different strategies that have developed to determine fake profile in social networks, many of them follow the similar portfolio of techniques, but they have been applied in different contexts (in different social networks or on different features set). Here we discuss only the selected unique solutions related to Facebook, Twitter and Sina weibo.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.2.1 In Facebook

The detecting cloned profile based on Facebook data. The approach consists of three components- distiller, profile hunter and profile verifier. The information distiller constructs test queries using the information extracted from the profile and run them in search engines and social networks. Then results returned against each query are taken into account by the distiller to create the user-record. The user record is a set of user identifying terms along with user's full profile name. The record is the input to the next component. The output of the information distiller is used by profile hunter to locate potential social network profiles belonging to the user. All the returned results are grouped as a profile-record. The profile record contains a link to the user's real profile and to all other returned profiles. The next component, profile verifier, examines the profile-record for similarity check with the user's original profile. Through the profile verification a similarity score is calculated based on the common values of information fields. Finally the profiles which have high probability to be cloned are presented with similarity scores. Fire et al,2012.[2] used topology anomalies to identify the spammers and fake profile. Apart from domain of graph theory and supervised learning, they exerted the parallel decision tree and Naïve Bayes classifier into their algorithm.

According to Boshmad et al,2011[17] adopted traditional web based Botnet design to build a group of adaptive social-bots as a a socialbot network and analysed its impact via Millions of the Facebook users. [10] analyzed the behaviour of identity clone attacks and proposed a detection framework.

Cao et al,2012 [8] ranked users in online services to detect fake accounts. Their ranking algorithm is supported by social graphs according to the degree-normalized probability of a short random walk which resides in non-sybil region. As a case study Krombholz et al,2012[7] have analyzed privacy related issues in social media contexts by creating desirable fake profiles and interacting with existing legitimate users of the network. Consequently they could discover how much information that can be harvested and analyzed from the users who interact with these fake profiles.

2.2.2 In Twitter

Identification of twitter fake profile is not far fetch, Twitter is classified into ; Mapping. Assembly and Classification. Twitter profile has the following Id, Name , Location ,Description, Profile image, Url. With the Profile model: Id, Nick Name, Current Location About me, Profile Image . they are preliminaries parameters that can be used to explain spam identification and fake profile identification on social networks especially Facebook, Twitter and Weibo. Spirin et al,2015 [18] They are as follows: Social Graph model, Algorithms based on labels propagation, Link Pruning and reweighing algorithms, algorithms with link-based features, algorithms with Link-based features, algorithms based on label refinement, Graph regularization algorithms and algorithms based on labels refinement.

Xiang et al,2010 [16] , have used interaction activity and similarity of users profile to develop and unsupervised model to estimate friendship strength. They evaluated the system on Proprietary data from Twitter.

2.2.3 In Sina Weibo

So far, little research has been carried out on weibo as primary data source; According to Zheng et al,2015,[19] weibo site users has reached 500 million. Statistics shows that weibo is consistently among the top 25 most frequent visited websites during the past few years. Weibo application is similar to Twitter, where user post messages, interact with friends, talk about news and share interesting topics via social network site. The features of Profile, Hash tag, mention. In weibo spam identification, fake profile, Fake identity of the weibo users. We identify the legitimate users of weibo and fake identity users. The Graph below fig1. Describe the simple following graph, in which user A is following user B, and user C are following each other. There are number of expressions in Sina Weibo allowing users to interact with others in a better way, including mention, repost and hashtag.

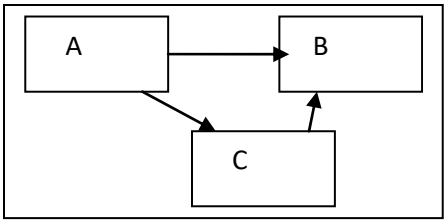

Figure 1: A simple following graph

[2] used malware anomalies to identify the spammers and Fake profiles, Apart from domain a simple following graph and the weibo exerted the parallel decision tree and Naïve classifiers into their algorithm.

Journal of Theoretical and Applied Information Technology
15th November 2017. Vol.95. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. Data Set

3.1 Facebook

As stated in the beginning of the paper Facebook is regarded as the primary data source for our research. Due to existing privacy policies and system API limitations of Facebook, we could get hold of few profile features only. To collect real fake profiles, we browsed the web and found several blogs and web sites where people have identified and listed the fake profiles. In some cases some profiles have been recognized by different people as fake profile. Conversely, profile which are specified as fake by only an individual is re-confirmed by manually checking whether they are fake. In the manual process of fake profile detection we followed the most commonly considered techniques by the social network community, such as, groups details are not matching with the users' other profile data, connected to other fake profiles, the information are not logical or reasonable , profile data is disharmonized and recommendation are made only among fake profiles, there are no credible connections in the connection chain and checking the legitimacy of the profile picture by searching on Google and TinEye. For example , Figure 2 shows two sample fake profiles. Through this process, we were able to confirm 34 fake profiles. Next, we have randomly identified 40 legitimate profiles from the Facebook Public profiles and confirmed their legitimacy through aforementioned manual techniques.

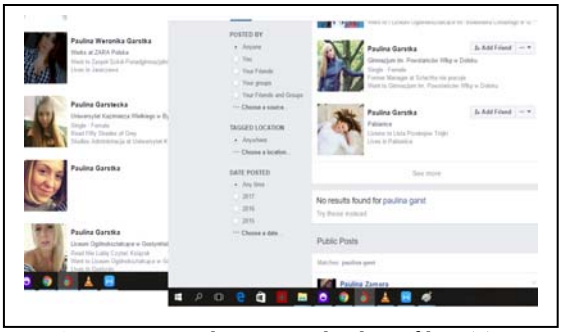

Figure 2: Example For Facebook Profiles: (1) Bot generated content (left side profile) (2) two profiles with same name, different pictures

List of all the profile features which were able to capture publicly in both legitimate and fake profiles along with the maximum and average values of each profile feature across all profiles in our dataset. Due to the Facebook restrictions, in variant of actual value, the highest number of connections is 500 and the highest number of features is 50. Therefore, rather than computing the normalized values via mean and standard deviation we utilized the maximum and minimum value of each feature. As each feature value doesn't present at least once in either a legitimate or a fake profile, the minimum value for all features is 0. The maximum and average values are shown in Table1

Table 1: Details of the profile features.

Profile Feature	Maximum Value	Average value	Description
No Language	5	0.347	Number of languages can speak
Profile Summary	1	0.52	Presence of profile summary
No Edubio data	7	1.467	Number of Education and Biodata
No of Connections	500	294.867	Number of connections
No News feed	37	2	Number of News feed made
Website URL	1	0.28	Presence of a URL for personal facebook site
No of friend list	50	10.213	Number of friend list on Facebook
No Place of work	16	3.08	Number of place of work
Profile Image	1	0.76	Presence of profile image
Number of Post posted	10	0.56	Number of post posted on facebook
Interest	1	0.267	Presence of any interests
No_Facebook Groups	51	8.907	Number of Facebook Groups and association added
School attended	16	0.613	Number of school attended
No of shared Photo	7	0.24	Number of shared photo
No of Tagged Photo Location	9	0.267	Number of Tagged photo and photo location

Finally, we divided all profiles into two equal groups randomly by three times, such that each of them contains the same number of fake and legitimate profiles. Moreover, each profile is distinctive from one another. Thus we have 3 data sets (Dataset 1, Dataset 2, Dataset 3), where each set contains two groups of 37 number of profiles-one of the group is marked as Training dataset and other is marked as Test dataset, Each of these groups of 37 profiles has 20 legitimate profiles and 17 fake profiles.

TABLE 2: The Number of fake and legitimate profiles used in training and test data sets

	Training dataset	Test dataset
Legitimacy Profiles	20	20
Fake Profiles	17	17

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.2 Twitter

In the midst of the different strategies that have been developed to determine fake profile in social networks, many of them follow the similar portfolio of techniques. Due to the existing privacy policies and system API limitation of Twitter. Twitter is a developer-friendly platform who provides Application Programming Interface(API) that allows us to crawl and collect data. In the past, Twitter provides API Whitelist which allow developers collect data from Twitter without query limitation.However, this feature has been revoked in February 11th 2016 [26]. After that, only few queries can be processed in a window of 15 minutes. To speed up our crawling process, we use multiple develop accounts to collect the data we need sort the fake profiles and determine the legitimate profiles.The API function we used is our crawler are:

Users/Lookup:

- ✓ This function can query up to 500 users in one request.
- ✓ Only 60 requests can be processed in 15-min window
- ✓ A requested user will not be returned if it is unknown, suspended, or
- ✓ deleted.
- ✓ If no users satisfies the condition, a HTTP 404 will be thrown

Due to the Twitter restrictions, in variant of the actual value, the highest number of connections is 500 and the highest number of skills is 50. Therefore, the computing the normalized values via mean and standard deviation. The maximum and average value are shown in Table 3.

TABLE 3: Details of cluster and profile features

Cluster or Profile feature	Maximum Value	Average Value	Description
No Languages	5	0.347	Number of languages can speak
No Id Name users,	1	0.52	Number of Identification and name
No Location users	7	0.52	Number of location of users
No Connections	500	294.867	Number of connections
No Recommendation	37	2	Number of recommendation made
Website URL	1	0.28	Presence of a URL for personal website
No Nickname Users	50	10.213	Number of Nickname
No Followers users	16	3.08	Number of Followers users
Profile Image	1	0.76	Presence of a profile image
No_Description	10	0.56	Number of description
Interests	1	0.267	Presence of any type interests
No Twitter Groups	51	8.907	Number of Twitter groups
No Tweets feed	16	0.613	Number of tweets
No Statuses	7	0.24	Number of statuses
No Biodata details	9	0.267	Number of Biodata

In twitter dataset, we divided all profiles into two groups such that each of them contains the same number of fake and legitimate profiles. They are training dataset and test dataset. Each of the groups of 37 profiles has 20 legitimate profiles and 17 fake profiles.

s stated in the beginning of the paper Facebook is regarded as the primary data source for our research. Due to existing privacy policies and system API limitations of Facebook, we could get hold of few profile features only. To collect real fake profiles, we browsed the web and found several blogs and web sites where people have identified and listed the fake profiles. In some cases some profiles have been recognized by different people as fake profile. Conversely, profile which are specified as fake by only an individual is re-confirmed by manually checking whether they are fake. In the manual process of fake profile detection we followed the most commonly considered techniques by the social network community, such as, groups details are not matching with the users' other profile data, connected to other fake profiles, the information are not logical or reasonable , profile data is disharmonized and recommendation are made only among fake profiles, there are no credible connections in the connection chain and checking the legitimacy of the profile picture by searching on Google and TinEye. For example , Figure 2 shows two sample fake profiles. Through this process, we were able to confirm 34 fake profiles. Next, we have randomly identified 40 legitimate profiles from the Facebook Public profiles and confirmed their legitimacy through aforementioned manual techniques.

TABLE 4: The Number of fake and legitimate profiles used in training and test data sets

	Training dataset
Legitimacy Profiles	20
Fake Profiles	17
Profile name	0

Journal of Theoretical and Applied Information Technology
15th November 2017. Vol.95. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3.3 Sina Weibo

Sina Weibo (Hereafter Weibo), a Chinese version of Twitter released by Sina corpo-ration in August 2009, has become the most popular Online Social Network platform in China. Weibo provides API service for developers to access their data. In Weibo,there is an up-limit of out-links, we browsed the web and found several blogs and web sites where people have identified and listed the fake profiles.

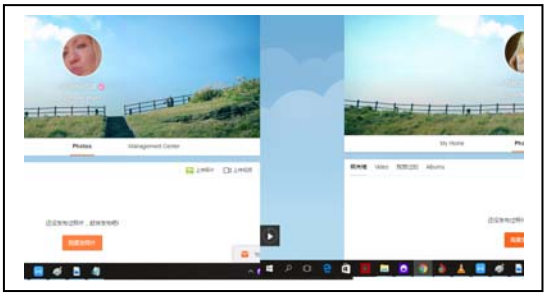

Figure 3: Example For Sina weibo fake profiles; Two profiles same name , different pictures, left side is legitimate and right side is fake

The profile data recommendations are made only among fake profiles, we confirm 34 fake profiles, Next, we have randomly identified 40 legitimate profiles the sina weibo public profiles and confirmed their legitimacy through aforementioned manual techniques. There are about 243 million active users daily of sina weibo [22]. Sina weibo is widely used in China, due to government restriction place on Facebook, twitter and other social networks. Table list all the profile features which we were able to capture publicly in both legitimate and fake profiles along with the maximum and average values of each profile feature across all profiles in our dataset.

TABLE 5: Details of the profile features Sina Weibo

Profile feature	Maximum Value	Average	Details
No Languages	5	0.347	Number of languages can speak
Profile Summary	1	0.5	Presence of profile summary
No Edu Qualification	7	1.467	Number of education qualifications
No Connections	500	294.867	Number of connections
No Recommendation	37	2	Number of recommendation made
Website URL	1	0.28	Presence of a URL for personal website
No Skills	50	10.213	Number of skills and expertise
No Professionals	16	3.08	Number of past and present professionals
Profile Image	1	0.76	Presence of a profile image
No_Awards	10	0.56	Number of award won
Interests	1	0.267	Presence of any type interests
No Sina weibo groups	51	8.907	Number of weibo groups
No of publications	16	0.613	Number of publications
No projects	7	0.24	Number of projects that work
No certificates	9	0.267	Number of certificates hold

In the end, we divided all Profiles into two equal groups randomly by three times, such that each of them contains the same number of fake and legitimate profiles.

TABLE 6: Number of the fake and legitimate profiles used in training and data sets

	Training dataset	Test dataset
Legitimacy Profiles	20	20
Fake Profiles	17	17

Thus we have 3 data sets (Dataset 1, Dataset 2, Dataset 3), where each set contains two groups of 37 numbers of profiles- one of the group is marked as Training dataset and other is marked as Test dataset. Each of these groups of 37 profiles has 20 legitimate profiles and 17 fake profiles.

4. METHOD AND RESULTS

In this section we explain how each techniques is used in the process of data mining to differentiate legitimate and fake profiles. The process has three levels, in the first level profile features are extracted by PCA and then second level NN and SVM are used to determine the fake and legitimate profiles. The third level, we calculate and compare the accuracy rates across the results of both techniques (see Figure 4)

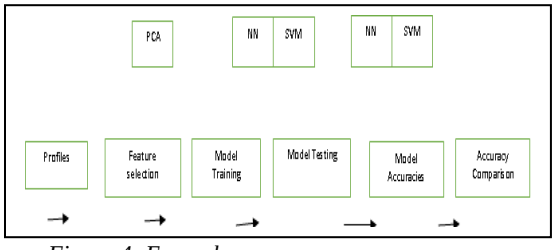

Figure 4: Example

4.1 Principal Component Analysis (PCA)

In this research PCA plays a major role by providing the support to take the decision on which profile features to be used in the data mining. PCA

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

is considered as the simplest but robust dimensionality reduction technique. Among the number of different mathematical ways of deriving PCA results we have selected the simplest case that is the variance maximization. In variance maximization first principal component has the highest projection variance which is the direction in feature space along and the second component defines the direction which has highest projection variance among all other orthogonal direction to the first component. In the process of calculating the scores on the features of Facebook, Twitter and Sina weibo, both fake and legitimate profiles are considered. We have used eigendcomposition which is the most commonly practiced calculation methodology for PCA to find the number of components. Initially, to ensure the sampling adequacy we have tested for Kaiser-Meyer-Olkin (KMO) and Bartlett's Test. The resulted KMO value is 0.724 which is higher than the acceptable level of 0.5 and the Bartlett's test is significant at $p<0.05$ Ashcroft and parker ,2009 [28]. This verifies the required numbers of samples to be adequate to precede the study.

Then we estimated the variation of the components and selected the components which have Eigen values more than 1 Kaiser ,1974[29]. Eingenvalues provide information of the variability in the data. There we found 5 Components which have the total variance of 64.82%. each component variations and their Eingenvalues are demonstrated in Table 7. Then we checked each component feature score which provides information about the structure of the observations and identified features that either load into several components with the scores value of more than 0.5 or not load into any component with more than 0. score value [29]. For these features to better understand the relationship between features and extracted principal components, we used Varimax rotatio to load the features into the components again. Still, we found some features unintendedly load to several components without

acceptable score values. To get clear features loading for the components we have removed such features step by step as mentioned in the following algorithm, finally, we could obtain the results as shown in Table 8 by removing Profile_Image, No_Facebook, Twitter, Sina Weibo_Groups. Due to removal of the features, remains are loaded int 4 components with the total variation of 66.15% and eveb the KMO value reduces to 0.655 which is still higher than the recommended boundary (>0.5) with the same significance value. The detail algorithm of PCA based features selection is given below shows the selected features and how they load the principal components at the end of running the feature selection algorithm.

TABLE 7: Total variance explained by PCA

Component	Total Initial Eigenvalues
1	4.375
2	1.801
3	1.290
4	1.195
5	1.061
6	.934
7	.840
8	.820
9	.618
10	.522
11	.440
12	.369
13	.284
14	.242
15	.208

Algorith for feature reduction through PCA

Initialize F with all features, where f is a feature in a feature set F
Initialize each Xf to zero, where Xf is an indicator variable associated to feature f

Do

Run PCA with Varimax rotation
If (Eigenvalue>1)
Select C, where C is selected principal components
Initialize L to empty, where L is a list
For each f E F
For each cE C
If Sf > 0.5, where Sf is feature scores for Feature f
Xf= Xf+ 1

End For
If Xf is not equal to I
Add f to L
End If
End For
End If
For Each f E L
Remove f from F
While (L length >0)
Output : F is the set of selected feature set

TABLE 8: Selected feature loading PCA

Feature No.	Profile feature	Comp. 1	Comp. 2	Comp. 3	Comp. 4
1	No language	0.614	0.098	0.218	-0.162
2	Profile summary	0.623	0.016	0.247	-0.097
3	No Edu biodata	0.827	0.139	-0.157	0.266
4	No news feeds	0.702	0.171	0.153	0.311
5	Website	0.195	0.860	0.106	-0.046

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

	URL				
6	Interests	0.079	0.913	0.025	0.040
7	No connections	0.208	-0.177	0.684	0.157
8	No recomm.	-0.007	0.161	0.800	0.076
9	School attended	0.426	0.335	0.695	0.122
10	No shared photo	0.164	0.072	0.154	0.685
11	No. tagged photo	-0.075	-0.082	0.077	0.843

TABLE 9: Selected feature correlation matrix

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]	[10]	[11]
[1]	1.000										
[2]	0.261	1.000									
[3]	0.311	0.379	1.000								
[4]	0.287	0.296	0.624	1.000							
[5]	0.161	0.236	0.259	0.237	1.000						
[6]	0.155	0.061	0.177	0.266	0.659	1.000					
[7]	0.095	0.169	0.105	0.340	0.036	-0.053	1.000				
[8]	0.120	0.251	0.021	0.205	0.161	0.141	0.278	1.000			
[9]	0.502	0.317	0.319	0.448	0.406	0.331	0.465	0.516	1.000		
[10]	0.132	0.153	0.206	0.207	0.075	0.091	0.188	0.132	0.270	1.000	
[11]	-0.065	-0.034	0.123	0.156	-0.900	-0.054	0.101	0.136	0.120	0.319	1.000

As per the Table 8 we can see that all the correlations are less than 0.6, only in two combinations (feature [3] and [4], feature [5] and [6] the values are marginally higher than 0.6. Nevertheless, those features are loading the same component. For example, both feature [3] and [4] are loading the component1 Table 5, similarly feature [5] and [6] are loading component 2 Table 5. Therefore, we can state that the selected features are not highly correlated to each other Chuang,2011 [29]. As a conclusion of this PCA based feature selection step we summarize the selected features in Table 9.

TABLE 10: List of selected and all profile features

S/No.	Feature Name	Is Selected PCA ? (Y-Yes, N-No)	Selected Feature Number
1	No_Languages	Y	[1]
2	Profile_Summary	Y	[2]
3	No_Edu_Biodata	Y	[3]
4	No_Connections	Y	[4]
5	No_Recommendation	Y	[5]
6	Web_Site_URL	Y	[6]
7	No_Shared_Photo	Y	[7]
8	No_Tagged_Photo	Y	[8]
9	Profile_Image	N	
10	No_Awards	N	
11	Interests	Y	[9]
12	No_Groups	N	
13	No_Publications	N	
14	No_Projects	Y	[10]

4.2 Neural Network (NN)

Currently there are many neural network (NN) algorithms that are used to train models either through supervised learning or unsupervised learning. In this research our focus is on the supervised learning where we have the legitimacy as response variable and selected profile features Table 9 as the input. We selected the Resilient back propagation (Rprop) algorithm as the base algorithm. Rprop does not account for the magnitude of the partial derivatives (only the sign) of the patterns and work out independently on each weight Riedmiller and Braun,1992[30]. Rprop is considered as one of the fastest algorithm in data mining Kumar and Zhang,2006,[31]. we selected neuralnet package in R Profiles for statistical computing Gunther and Fritsch,2010,[32].

Neuralnet is flexible to include the custom-choice of error-function, number of covariates with response variables and the number of hidden layers with hidden neurons. Since response variable (legitimacy of the profile) is considered as binary (if legitimate, then the value I "1" and if fake, then value is "0"), logistic function (default) is chosen as the activation function of the training and cross-entropy (err.fct= "ce") is selected as the error function. To ensure that the output is mapped by the activation function to the interval [0, 1]; we defined linear. Output as FALSE Gunther and Fritsch,2010 [32]. With this preparation, we trained the model by determining the number of hidden neurons and layers in relation to the optimized results. After several iterations, the best result (I.e highest accuracy) is achieved with one hidden layer with two neurons.

irst, we trained the model for all three datasets with all the features and saved their models in different variables. Then again we selected same datasets, remove the features which are not selected by the PCA and saved the models to different variables.

The "compute" function of the library is used to predict results for new data based on the stored NN models. Since the compute function automatically redefine the NN structure only to calculate the output for arbitrary covariates, we could easily figure out the predictions for the legitimacy of each related test datasets with all and selected features. Then the results are compared with the actual legitimate values (i.e. whether the profile is fake or legitimate) and calculated the accuracy for each dataset with all and selected features in Table 10.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

% Accuracy = *Total number of correctly identified profiles, both fake or legitimate X 100*
Total number of profiles

We can see from Table 11, that the accuracy result is higher in the case of selected features than when all features are used in the NN. In the case of all features, the model deterorate due to unnecessary data points leading to the over-fitting problem of NN. This definitely the importance of the PCA step in our approach if NN is used for detection legitimacy of Facebook, Twitter and Sina weibo profile.

TABLE 11: Accuracy of the results obtained through neural network training

	Dataset	Training error	Accuracy(%)
All Features	Dataset1	0.043	84.85
	Dataset 2	0.083	68.29
	Dataset 3	0.064	86.11
Selected Features	Datset 1	0.025	87.88
	Dataset 2	0.089	70.73
	Dataset 3	0.012	89.89

4.3 Support Vector Machine (SVM)

In this section we apply support vector machine (SVM) based approach to identify the fake profile for the SVM training we applied C-support vector classification (C-svc) which is a Quadractical Programming (QP). C-svc can find the best possible hyperplane by measuring the margin between two classes using 2-norm of the normal vector and norm-1 is used for the feature selection Zhang et al,2013 [33] according to the Mercer's theorem Cortes and Vapnik, 1995 [34] the Kernel function K can be considered as equal to a dot product in input space and due to the nonlinearity of the profile features, SVM is able to create a random decision functions in the input space on the kernel function. Both the Radial Basis function Kernel (rdfdot) and polynomial Kernel (Polydot) are used as Kernel functions for better understanding of SVM performance on the dataset.

The Radial Basis Kernel is, selected because it uses the heuristics in sigest to calculate better sigma value, and we did not need to assign values to the Kernel parameters. Radial basis function Kernel K can be written as:

$K(Xi, Xj) = exp (y/ Xi-Xj/\^2)$

Polynomial Kernel is selected as it uses a combination of features of the input sample instead of determining similarity of those independently. The polynomial Kernel function can be written as:

$K (Xi,Xj) = (-y/Xi.Xj + C)\^d$

When C=0 Kernel is called homogenous

For both the Kernels $K (Xi, Xj) = o (Xi).o(Xj); y= - 1/\%2$

The transformation function o maps a dot product of input data points into higher dimensional feature space where the non-linear patterns would demonstrate linearity is an adjustable parameter and y>0

Since we intended to use C-svc classifier, we use KSVM (function of R,Kernlab package) to train the SVM model. KSVM facilitates the sequential Minimal Optimization (SMO)algorithm for solving SVM quadratic programming (QP) optimization problem Joachims 1999 [34] we have performed training with two proposed kernel functions (Radial Basis and Polynomial) to create SVM models for all three training datasets with all features and PCA based selected features. Then the models are tested using the test dataset of the respective group. In this way we have total 12 models (2 Kernel functions, 3 datasets - each with both all features and selected features) to test and compare. Each model is tested with the pertinent test dataset and we calculated the accuracy rate. The Consolidated results and presented in the Table 12.

TABLE 12: Accuracy rate of the results based on kernel dataset

	Dataset Kernel Type	Radial Basis Kernel (%)	Polynomial Kernel (%)
All Features	Dataset 1	78.79	84.85
	Dataset 2	73.17	73.17
	Dataset 3	88.89	91.67
Selected Features	Dataset 1	75.76	84.85
	Dataset 2	78.05	75.61
	Dataset 3	91.67	91.67

In each of the scenarios polynomial kernel derived the optimized results with less number of vectors in comparison to the Radial Basis Kernel. Since we need to compute the dot product of each support vector with the test point, the computational complexity of the model is linear to the number of support vectors. We observe from table 11 that other than Dataset 2 with selected features, Polynomial Kernel Performs better or equal to the Radial Basis Kernel. Also, the Polynomial Kernel Performs better when is applied on selected features by PCA than when all features are considered.

Journal of Theoretical and Applied Information Technology
15th November 2017. Vol.95. No 21

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TABLE 13: False positive and false negative values based on the kernel and feature selection

Feature Selection		Radial basis kernel (%)	Polynomial kernel (%)
All feature	False positive	14.84	9.84
	False negative	6.73	6.93
	False positive	10.94	13.52
Selected features	False negative	7.24	2.44

Additionally, in Table 12, we present average false negative and false positive values across all three datasets for both the kernels. We can see from Table 12 that the false negative value of Polynomial Kernel with selected features is the lowest. It is very important to note that in this case false negative has higher risk value in business than false positives. For example, due to false identification of a fake facebook, twitter and sina weibo profile as legitimate profile .

Thus the above discussion concludes for identification of legitimacy of Facebook, Twitter , Sina weibo with SVM, SVM with polynomial kernel applied on PCA based selected profile features gives the highest accuracy with the lowest false negative

5. METHOD AND RESULTS

In this research, we have compared the results of two data mining techniques to determine the most appropriate approach to differentiate the legitimate profiles from fake profiles in 3 social media. Table… Summarizes the final accuracy values akin to each technique by calulating the average across all three datasets. In addition, it shows the average false positive rate and false negative rate for each technique.

Although the RBF Kernel is the mostly used kernel in the data mining context, in our scenario polynomial kernel gives us the higher accuracy compared to RBF kernel. Additionally polynomial kernel flase negative value is reduced when PCA based selected features are used Tables 12 , therefore we can conclude in case of SVM, polynomial kernel with selected feature is the right choice.

TABLE 14: Accuracy comparison of two techniques NN & SVM

Algorithm	Feature selection	Accuracy rate (%)	False positive (%)	False negative (%)
Neural Network	All features	79.75	15.17	5.08
Neural Network	Selected features	82.83	13.21	4.29
Support Vector Machine(Polynomial Kernel)	All features	83.23	9.84	6.93
Support Vector Machine	Selected features	84.04	13.52	2.44

In accordance to the final accuracy rates, we can see that SVM has the highest accuracy rate between the two techniques regardless of the number of features used. However the difference between NN and SVM is 2.48 % when all features are selected and 1.21 % when only the extracted features are selected. As per the theoretical rationale SVM vector machine is more preferred data mining technique for the data set like this, because SVM can compute results even with less number of training data points and it does not suffer from local extrema.

False positive and false negative columns exhibits the percentage of the number of legitimate profiles detected as fake and number of fake profiles detected as legitimate respectively. Compare to the false positive, false negative has a higher risk, because if a fake profile is identified as legitimate, then the impairment can be occured is much higher whilst a legitimate profile detected as fake as shown in Table 12, SVM with selected feature has the lowest false negative value (2.44%). Thus, between the approaches (NN and SVM), SVM with polynomial kernel gives the most accurate result with low false negative for the task of identification of fake profile in Facebook, Twitter and Sina weibo.

NN and SVM provide higher accuracy when the features are selected through the PCA. For both dataset 1 and 2 the accuracy values with the selected features are higher than when all the features are selected. In addition the false negative value is less for both techniques when only the selected features are used for legitimate determination. Thus, PCA based feature selection is an important step in the process of identification of fake profile in Facebook, Twitter and Sina weibo.

NN and SVM provide higher accuracy when the features are selected through the PCA. For both dataset 1 and 2 the accuracy values with the selected features are higher than when all the features are selected. In addition the false negative value is less for both techniques when only the selected features are used for legitimate

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

determination. Thus, PCA based feature selection is an important step in the process of identification of fake profiles in Facebook, Twitter and Sina weibo.

So, from the above discussion, we conclude PCA based feature selection and subsequently SVM with polynomial kernel based modelling for determining legitimacy of profile is the right approach for identification of fake profile from Facebook, Twitter and Sina weibo, where limited number of profile features are public.

SVM accuracy can be advanced by further analysing the kernel, fine-tunning the kernel parameters and tolerance level Cristianini and shawe-taylor,2000 [35]. NN is more accurate when there are higher number of data points, we can expect more optimized results while the numbers of profiles are increased. In the facebook, twitter and sina, it is quite difficult to increase the number of data points as Facebook, Twitter impose limitation on accessing its data and it is particularly challenging to increase the number of fake profiles, while the sina weibo its limited on the url website.

Next we show how our result compare with the results of previously proposed approaches. It is difficult to implement and run previous approaches on our data set, because neither of the previous approach is based on limited Facebook, Twitter and Sina weibo data. So, in Table 13, we present the accuracy of the results of previous research as reported by them along with social network on which it was applied and the dataset requirement of the approach.

In summary, prior research which focused on fake profile identification, has similar accuracy rate compared to what we accomplished in our research. In all these prior research, researchers have used the user activities as a criterion to decide the legitimacy of a profile. A user activity of a profile includes all the dynamic information of a user (number of posts, information about friends and their behaviour). Such dynamic data of a user are impossible to access in Facebook, Twitter and Sina weibo, due to Facebook, Twitter and Sina weibo data accessibility restriction which is elucidated in section 1.3. Though, the prior studies listed in Table 14 have analyzed more than thousands of profiles to accomplish the shown accuracy, all the fake profiles exploited are simulated. On the contrary, our approach considered actual fake profiles in Facebook, Twitter and Sina. Additionally, with the consideration of practicality of the approach, our approach is based on limited static profile data and does not include any profile data that is hard to access or mostly restricted by all the 3 social media. Considering these significant differences, compared to results of prior research our results of 84% accuracy with 2.44% false negative can be considered as an excellent improvement.

TABLE 15: Details of prior research on fake profile identification

Technique used	Accuracy (%)	Feature type	Social network	Source
Support Vector Machine	78	Dynamic and static e.g.: profile age, presence of profile image, followers and friend count, post/messages, details of tweets	Twitter	(Chakraborty et al. 2012)
Naives Bayes	67	Static e.g. profile's content such as age, gender, location	Twitter	Feizy et al.2009
Decision Tree	69.25	Static e.g. profile's content such as Age, gender, location	My Space	Feizy et al.2009
Nearest Neighborhood	67.05	Statitc	My Space	Feizy et al.2009
Decision Tree	86.10	Dynamic e.g. profiles	My space	
Nearest Neighborhood	84.59	Dynamic e.g.profile's connectivity , the amounts and types of interactions	Twitter	Feizy et al.2009
Weka Classifier:	Random Forest algorithm 94.5			97 Dynamic e.g number of friends, friend requests, details of short text messages.

6. LIMITATION AND FUTURE WORK

The main limitation is the verification of the sources and the published fake profiles. There can

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

be situations where the source classifies a profile as a fake profile without proper evidence. Second, when a cloning attack occurs on certain profile we cannot actually identify which profile is the fake. One similar setup is shown in Figure 1. Between two of these profiles one can be legitimate. Our future intention of this study is to follow a similar approach and analyze other social networks to check the status of the accuracy level of differentiating fake and legitimate profiles exclusively based on the limited factual data. Also, we can improvise the data mining by considering other important information such as characters of user name, including length, lower case, and so on,location information including size of address and geographical connection.

7. CONCLUSION

In this paper, we propose an approach identify the fake profile in Facebook, Twitter and Sina weibo with limited profile data. As we concluded in our discussion SVM with Polynomial Kernel on PCA based selected features has the capability to train a model to achieve higher accuracy with low false negative on differentiating the legitimate profiles and fake profiles in Facebook, Twitter and Sina weibo. Many of the past research on fake profile are based on where both dynamic and static behavioral data on social network and in most cases tested only on the simulated fake dataset. Even though there is a research conducted on Twitter data, Facebook, Sina weibo data in related to spam detection [36, 37, 38], to the best of our knowledge this is the first research to identify 3 social media together with an approach of fake profile identification and spam identification. Our approach is based on static profile feature data and not dynamic data, which is not accessible in all the three social networks. We demonstrate that with limited profile data our approach can identify the fake profile with 84% accuracy and only 2.44% false negative, which is comparable to the results obtained by other existing approaches based on the larger dataset and more profile information.

At present, social network users strongly contemplate on data privacy, in parallel social network communities have advance their security and authenticated frameworks to provide better information hiding capabilities to users with new restriction on accessing the information in the network Fang et al,2010 [39]; Chen et al,2009 [40]. Along with that this research can be a motivation to work on limited social network information and find solutions to make better decision through authentic data. Additionally, we can attempt similar approaches in other domains to find successful solutions to the problem where the least of information is available.

REFRENCES:

[1] H. Rheingold, "The Virtual Community:Homesteading on the Electronic Frontier" 2000. Conference on ICJA'2000. MIT press.

[2] M. Fire, et al, " Strangers intrusion detection-detecting spammers and :fake profiles in social networks based on topology anomalies" 2012. International Human Journal1(1): 26-39.2012.

[3] K. Lee., et al,. "Seven Months with Devils: A long-Term Study of Content Polluters on twitter ". ICWSM'11. In Proceeding ICWSM 2011.

[4] G. Kontaxis, et al; " Detecting Social Network Profile Cloning. Pervasive Computing and Communications Workshops "(PERCOM Workshops), 2011 IEEE International Conference'11.

[5] A. Akinyelu, et al., "Classification of Phising Email Using Random Forest Machine Learning Technique" Journal of Applied Mathematics, 2014. Hindawi Publishing Corp. Vol.2014.pp6.

[6] N.Jindal., " Mining Comparative Sentences and Relations" . In Proceeding of AAAI'06 conference 2006. pp 1331-1336.

[7] K. Krombholz., et al,. " Fake identities in social media: A case study on the sustainability of the Facebook business model." 2012. International Journal of service science research '12(4)175-212.

[8] Q. Cao,.et al. " Aiding the detection of fake accounts in large scale social online services". In proceeding of NSDI'12.(2012)

[9] L. Bilge, et al. " All your contacts are belong to us: automated identity theft attacks on social networks. Proceeding of the 18th International conference on world wide web,ACM. 2009.

[10] L. Jin, et al,;" Towards active detection of identity clone attacks on online social networks. Proceeding of the first ACM conference on Data and application security and privacy,ACM. 2011.

[11] G. Kontaxis, et al; " Detecting Social Network Profile Clone attack. Pervasive Computing and Communications Workshops "(PERCOM Workshops), 2012 IEEE International Conference'12

[12] P. Kazienko; and K. Musial. "Social Capital in Online Social Networks. Knowledge-Based

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Intelligent Information and Engineering Systems. Springer. 2006.

[13] D. Bradbury. " Data Mining with Linkedin." Computer Fraud & Security .2011 (10)5-8.

[14] I. Jollife. " Principal Component Analysis, Wiley Online Library.(2005).

[15] C.J. Hsieh,. et al, " Organizational overlap on social networks and its applications. Proceedings of the 22nd International Conference on world wide web, International world wide web conference steering committee.2013.

[16] R.Xiang, et al, " Modelling relationship strength in online social networks. Proceedings of the 19th International conference on world wide web,ACM. 2010.

[17] Y. Boshmad, et al. " The Socialbot network : when bots socialize for fame and money. Proceeding of the 27th Annual Computer Security Applications Conference.ACM. 2011.

[18] N.Spirin and J. Han " Spammers in Sina weibo" ACM. 2015.

[19] C.Zheng. " The optimality of naives bayes". 2015 .In FLAIRS conference'15.

[20] F. Benevenuto,. et al, " Identifying video spammers in online social networks" In Proceeding 8th ACM/IEEE-CS Joint conference on digital library. 2008

[21] J.Caverlee., et al., " Socialtrust: tamper-resilient trust establishment in online communities". In Proceeding 8th ACM.2008.

[22] Fan. Yang,. et al., "Automatic detection of rumor on sina weibo"MDS'12 Proceeding of the ACM SIGKDD workshop on Mining data semantic. 2012.

[23] Xinjiang Lu, Zhiwen Yu, Bin Guo, Jiafan Zhang, Alvin Chin, Jilei Tian, Yang Cao; "Trending Words Based Event Detection in Sina Weibo" . July 2014 BigDataScience '14: Proceedings of the 2014 International Conference on Big Data Science and Computing.2014

[24] Hao Chen, Jun Liu, Jianhong Mi "SpamDia: Spammer Diagnosis in Sina Weibo Microblog"June 2016 MobiMedia '16: Proceedings of the 9th EAI International Conference on Mobile Multimedia Communications.2016.

[25] Qian Zhang, Bruno Goncalves.;" Topical differences between Chinese language Twitter and Sina Weibo

"April 2016 WWW '16 Companion: Proceedings of the 25th International Conference Companion on World Wide Web.ACM.2016.

[26] Supraja Gurajala, Joshua S. White, Brian Hudson, Jeanna N. Matthews; .2015 "Fake Twitter accounts: profile characteristics obtained using an activity-based pattern detection approach"July 2015 SMSociety '15: Proceedings of the 2015 International Conference on Social Media & Society.2015.

[27] Monika Singh, Divya Bansal, Sanjeev Sofat; " A Novel Technique to Characterize Social Network Users: Comparative Study" .November 2016 ICCNS '16: Proceedings of the 6th International Conference on Communication and Network Security.2016.

[28] D. Aschcroft and D.Parker,. " Development of the pharmacy safety climate Questionnaire: a principal component analysis." Quality and Safety in Health care 18 (1): 28-31. 2009.

[29] H.F. Kaiser., " An index of factorial simplicity" Psychometrika 39 (1): 31-36. 1974

[30] M. Reidmiller and H, Braun. " A fast adaptive learning algorithm " Proceeding of ISCIS VII Universitat ,Citeseer.

[31] A.Kumar and D.Zhang,. " Personal recognition using hand shape and texture." Image Processing , IEEE Transactions on 15 (8): 2454-2461.

[32] F. Gunther and S.Fritsch. " Neuralnet: Training of neural networks." The R Journal 2(1): 30-38.2010.

[33] C. Zhang., et al. " Knowledge-based Support Vector Classification Based on C-SVC" Procedia Computer Science 17: 1083-1090. 2013.

[34] C. Cortes and V. Vapnik. " Support-Vector networks." Machine Learning 20 (3): 273-297. 1995.

[35] N. Cristianni, and J. Shawe-Taylor. " An introduction to support vector machines and other kernel-based learning methods, Cambridge University press. 2000.

[36] Diego Saez-Trumper." Fake tweet buster: a webtool to identify users promoting fake news ontwitter"2014 HT '14: Proceedings of the 25th ACM conference on Hypertext and social media.2014.

[37] Prudhvi Ratna Badri Satya, Kyumin Lee, Dongwon Lee, Thanh Tran, Jason (Jiasheng) Zhang; "Uncovering Fake Likers in Online Social Networks"October 2016 CIKM '16: Proceedings of the 25th ACM International on Conference on Information and Knowledge Management.2016.

[38] Wanqiu Guana, Haoyu Gaob, Mingmin Yangb, Yuan Lia,, Haixin Mac, Weining Qianc, Zhigang Caob , Xiaoguang Yangb; "Analyzing

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

user behavior of the micro-blogging website Sina Weibo during hot social events"Physica A: Statistical Mechanics and its ApplicationsVolume 395, 1 February 2014, Pages 340–351.2014.

[39] L. Fang and K. Lefevre. " Privacy wizards for social networking sites. Proceeding of the 19th international conference on world wide web, ACM. 2010.

[40] X.Chen and S.Shi; " A literature review of privacy research on social network sites" Multimedia Information Networking and Security, 2009. MINES'09. International Conference on IEEE.2009.

[41] G. Stringhini., et al. " Detecting spammers on social networks. Proceedings of the 26th Annual Computer Security Application Conference, ACM.2010.

[42] R.Taylor. " Interpretation of the correlation co-efficient : a basic review." Journal of diagnostic medical sonography 6(1): 35-39. 1990.

[43] W.Shu and Y.H.Chuang., " Why people share knowledge in virtual communities." Social behavior and personality: an international journal 39 (5): 671-690. 2011.

[44] A. Chakraborty., et al. " Spam: a framework for social profile abuse monitoring". 2012.

[45] S.Li., et al., " Fusing images with different focuses using support vector machine." Neural Networks, IEEE Transactions on 15(6): 1555-1561.

[46] V.M. Prieto, M. Alvarez, & F.Cacheda. " Detecting Linkedin spammers and its spam Nets". International Journal of Advanced Computer Science and Applications. 4(9). 2013.

[47] K. Krombholz., et al., " Fake identities in social media: A case study on the sustainability of the Facebook Business model." Journal of service science research 4 (2): 175-212.2012.

[48] T. Joachims, " Making large Scale SVM learning practical" Proceeding of ICCSE '99. 1999.

[49] Akshay J. Sarode, Arun Mishra; "Audit and Analysis of Impostors: An experimental approach to detect fake profile in online social network"September 2015 ICCCT '15: Proceedings of the Sixth International Conference on Computer and Communication Technology 2015.

[50] Yuanyuan Bao, Chengqi Yi, Yibo Xue, Yingfei Dong "A new rumor propagation model and control strategy on social networks"August 2013 ASONAM '13: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis.