IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol.-7, Issue-7, July 2019 E-ISSN: 2347-2693

Automatic Detection of Fake Profiles in Online Social Networks

R.V. Kotawadekar[1*], A.S. Kamble[2], S.A. Surve[3]

[1,2,3] Department of MCA, Finolex Academy of Management and Technology, Maharashtra, India

[*] *Corresponding Author: rahul.kotawadekar@famt.ac.in, Tel.: +91-99757 26126*

DOI: https://doi.org/10.26438/ijcse/v7i7.4045 | Available online at: www.ijcseonline.org

Accepted: 16/Jul/2019, Published: 31/Jul/2019

***Abstract*—** In the present generation, the social life of everyone has become associated with the online social networks. These sites have made a drastic change in the way we pursue our social life. Making friends and keeping in contact with them and their updates has become easier. But with their rapid growth, many problems like fake profiles, online impersonation have also grown. There are no feasible solution exist to control these problems. In this project, we came up with a framework with which automatic detection of fake profiles is possible and is efficient. This framework uses classification techniques like Support Vector Machine, Naive Bayes and Decision trees to classify the profiles into fake or genuine classes. As, this is an automatic detection method, it can be applied easily by online social networks which has millions of profile whose profiles cannot be examined manually.

***Keywords*—**Threats, Facebook Immune System, Classification, Training Datasets, Profile Attributes

I. Introduction

A social networking site is a website where each user has a profile and can keep in contact with friends, share their updates, meet new people who have the same interests. These Online Social Networks (OSN) uses web 2.0 technology, which allows users to interact with each other. These social networking sites are growing rapidly and changing the way people keep in contact with each other. The online communities bring people with same interests together which makes users easier to make new friends.

A. History

These social networking sites starting with http://www.sixdegrees.com in 1997, then came http://www.makeoutclub.com in 2000. The www. sixdegrees.com couldn't survive much and closed very soon but new sites like MySpace, LinkedIn, Bebo became successful and Facebook was launched in 2004 and presently it is the largest social networking site in the world.

1.2 Social Impact

In the present generation, the social life of everyone has become associated with the online social networks. These sites have made a drastic change in the way we pursue our social life. Adding new friends and keeping in contact with them and their up- dates has become easier.

The online social networks have impact on the science, education, grassroots organizing, employment, business, etc. Researchers have been studying these online social networks to see the impact they make on the people. Teachers can reach the students easily through this making a friendly environment for the students to study, teachers now-a-days teachers are getting themselves familiar to these sites bringing online classroom pages, giving homework, making discussions, etc. which improves education a lot. The employers can use these social networking sites to employ the people who are talented and interested in the work, their background check can be done easily using this. Most of the OSN are free but some charge the membership fee and uses this for business purposes and the rest of them raise money by using the advertising. This can be used by the government to get the opinions of the public quickly.

The examples of these social networking sites are sixdegrees.com, The Sphere, Nexopia which is used in Canada, Bebo, Hi5, Facebook, MySpace, Twitter, LinkedIn, Google+, Orkut, Tuenti used in Spain, Nasza-Klasa in Poland, Cyworld mostly used in Asia, etc. are some of the popular social networking sites.

II. THREATS

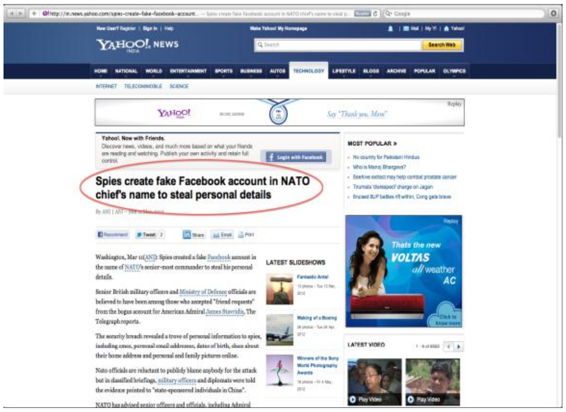
A. Cloud Confidentiality

Fake profiles are the profiles which are not genuine i.e. they are profiles of persons who claim to be someone they are not, doing some malicious and undesirable activity, causing problems to the social network and fellow users.
Why do people create fake profiles?

- Social Engineering
- Online impersonation to defame a person
- Advertising and campaigning a person, etc

1) *Social Engineering:* Social Engineering in terms of security means the art of stealing confidential information from people or gaining access to some computer system mostly not by using technical skills but by manipulating people themselves in divulging information. The hacker doesn't need to come face to face with the user to do this. The social engineering techniques are like Pretexting, Diversion theft, phishing, baiting, quid pro quo, tailgating, etc

Eg: Creating a profile of some person X not in some online social networking site like Facebook. Adding the friends of the X in Facebook and making them believe that it's the profile of X. They can get the private information meant for only X by communicating with X's friends in Facebook.



Spies create fake Facebook account in NATO chief's name to steal personal details

Washington, Mar 11(ANI): Spies created a fake Facebook account in the name of NATO's senior-most commander to steal his personal details.

Senior British military officers and Ministry of Defence officials are believed to have been among those who accepted "friend requests" from the bogus account for American Admiral James Stavridis, The Telegraph reports.

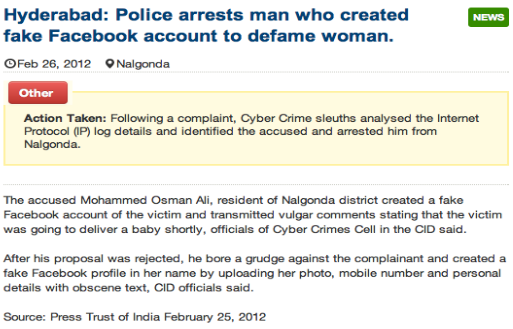
The security breach revealed a trove of personal information to spies, including emails, personal email addresses, dates of birth, clues about their home address and personal and family pictures online.

Nato officials are reluctant to publicly blame anybody for the attack but in classified briefings, military officers and diplomats were told the evidence pointed to "state-sponsored individuals in China".

NATO has advised senior officers and officials, including Admiral

Fig. 2.1 Example of social engineering

2) *Online impersonation to defame a person:* The other reason why people create fake profiles is to defame the persons they do not like. People create profiles in the name of the people they don't like and post abusive posts and pictures on their profiles misleading everyone to think that the person is bad and thus defaming the person.



Hyderabad: Police arrests man who created fake Facebook account to defame woman.

NEWS

Feb 26, 2012 Nalgonda

Other

Action Taken: Following a complaint, Cyber Crime sleuths analysed the Internet Protocol (IP) log details and identified the accused and arrested him from Nalgonda.

The accused Mohammed Osman Ali, resident of Nalgonda district created a fake Facebook account of the victim and transmitted vulgar comments stating that the victim was going to deliver a baby shortly, officials of Cyber Crimes Cell in the CID said.

After his proposal was rejected, he bore a grudge against the complainant and created a fake Facebook profile in her name by uploading her photo, mobile number and personal details with obscene text, CID officials said.

Source: Press Trust of India February 25, 2012

Figure 2.2: Example of online impersonation

Fig 2.2 shows the screenshot from a website which shows that a man named Mohammad Osman Ali has created a fake profile of a woman in Facebook and tried to defame her. The police finally caught and arrested him. This shows a very serious problem existing now-a-days.

3) *Advertising and Campaigning:* Imagine a scenario where a movie is released and one of your friends in Facebook posted that the movie was awesome. This makes a first impression on you that the movie is good and you would want to watch it. This is how advertising and campaigning works through OSN. The review posted by a genuine user is always desirable but these reviews when posted by fake profiles and completely undesirable.

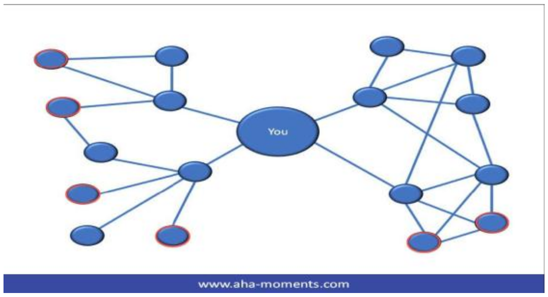
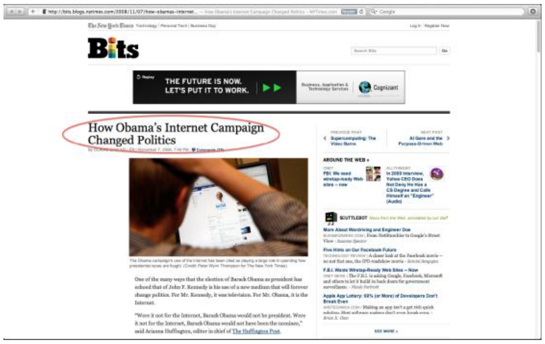


Figure 2.3: Social influence via online social network

Assume that Fig 2.3 shows a social graph where the blue nodes shown are real profiles, the red circled profiles show fake profiles and the edges show the connections between them. If the fake profiles start advertising a brand or campaigning for some politician then the users connected to the fake profiles are misled in believing them. In turn the profiles who didn't add the fake profiles are affected using the mutual connections.



How Obama's Internet Campaign Changed Politics

One of the many ways that the election of Barack Obama as president has echoed that of John F. Kennedy is his use of a new medium that will forever change politics. For Mr. Kennedy, it was television. For Mr. Obama, it is the Internet.

"Were it not for the Internet, Barack Obama would not be president. Were it not for the Internet, Barack Obama would not have been the nominee," said Arianna Huffington, editor in chief of The Huffington Post.

The New York Times
Bits
THE FUTURE IS NOW. LET'S PUT IT TO WORK.
Cognizant
AROUND THE WEB »
FBI: We need wiretap-ready Web sites – now
In 2009 Interview, Yahoo CEO Does Not Deny He Has a CS Degree and Calls Himself an "Engineer" (Audio)
SCUTTLEBOT
More About Wardriving and Engineer Doe
Five Hints on Our Facebook Future
F.B.I. Wants Wiretap-Ready Web Sites – Now
Apple App Lottery: 60% (or More) of Developers Don't Break Even

Figure 2.4: Example of advertising and campaigning

Fig 2.4 shows a screenshot, which shows the post in Newyork Times showing the most successful internet campaigning done by Obama which collected around 500 million dollars of election fund for him. This shows the power of internet campaigning.

B. Social Bots

Social bots are semi-automatic or automatic computer programs that replicate the human behavior in OSN. These are used mostly by hackers now-a-days to attack online social networks. These are mostly used for advertising, campaigning purposes and to steal user's personal data in a large scale. These social bots communicate with each other and are controlled by a program called botmaster. The botmaster may or may not have inputs from a human attacker. The social bots look like human profiles with a randomly chosen human name, randomly chosen human profile picture and the profile information posted randomly from a list prepared from before by the attacker. These social bots send requests to random users from a list. When someone accepts the request, they send requests to the friends of the user who accepted the request, which increases the acceptance rate due to existence of mutual friends.

Recently a researcher from University of British Columbia made a social botnet of103 bots in Facebook and added 3000 friends in just 8 weeks. He was able to extract around 250 GB of personal data of users. This shows the extent of the applications of social bots by the attackers.

C. Facebook Imune System (FIS)

When we consider Facebook, it has its own security system to protect its users from spamming, phishing, etc. and this is called Facebook immune system. FIS does real time checks on every single click and every read and write operation done by it. This is around 25 Billion checks per day and as high as 620,000 checks per minute at peak as of May, 2011.

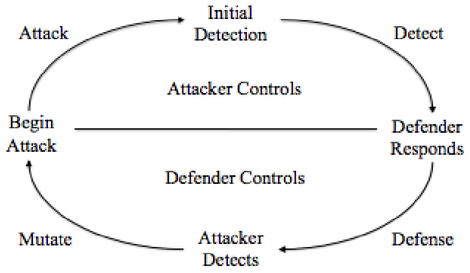


Figure 2.5: The adversarial cycle

Fig 2.5 shows the adversarial cycle in which the top part is controlled by the attacker and the bottom part shows the response by the FIS to control the attack, which when detected by the attacker, he/she mutates the attack and attacks it again. This goes on like a cycle and is never ending. FIS is able to detect the spam, malware and phishing produced by the compromised ad fake accounts. They are actually able to reduce the spam to less than 0.4. FIS is not successful in detecting the social bots and the fake accounts created by humans. This can be seen by the example mentioned above where a researcher created 103 social bots to collect a lot of personal data of users and Facebook couldn't detect this attack.

III. PROPOSED WORK

A. Overview

Each profile (or account) in a social network contain lots of information such as gender, no. of friends, no. of comments, education, work etc. Some of these information are private and some are public. Since private information is not accessible so, we have used only the information that are public to determine the fake profiles in social network. However, if our proposed scheme is used by the social networking companies itself then they can use the private information of the profiles for detection without violating any privacy issues. We have considered this information as features of a profile for classification of fake and real profiles.

The steps that we have followed for detection of fake profiles are as follows.

1. First all the features are selected on which the classification algorithm is applied. Proper care should be taken while choosing the features such as features should not be dependent on other features and those features should be chosen which can increase the efficiency of the classification.
2. After proper selection of attributes, the dataset of previously identified fake and real profiles are needed for the training purpose of the classification algorithm. We have made the real profile dataset whereas the fake profile dataset is provided by the Barracuda Labs, a privately held company providing security, networking and storage solutions based on network appliances and cloud services.
3. The attributes selected in step 1 are needed to be extracted from the profiles (fake and real). For the social networking companies which want to implement our scheme don't need to follow the scrapping process, they can easily extract the features from their database. We applied scrapping of the profiles since no social network dataset is available publicly for the research purpose of detecting the fake profiles.
4. After this the dataset of fake and real profiles are prepared. From this dataset, 80% of both profiles (real and fake) are used to prepare a training dataset and 20% of both profiles are used to prepare a testing dataset. We find the efficiency of the classification algorithm using the training dataset containing 922 profiles and testing dataset containing 240 profiles.
5. After preparation of the training and the testing dataset, the training dataset is feed to the classification algorithm. It learns from the training algorithm and is expected to give correct class levels for the testing dataset.
6. The levels from the testing dataset are removed and are left for determination by the trained classifier. The efficiency of the classifier is calculated by calculating the no. of correct prediction divided by total no. of predictions. We have used

three classification algorithms and have compared the efficiency of classification of these algorithms.

B. Proposed framework

The proposed framework in the figure 3.1 shows the sequence of processes that need to be followed for continues detection of fake profiles with active leaning from the feedback of the result given by the classification algorithm. This framework can easily be implemented by the social networking companies.

1. The detection process starts with the selection of the profile that needs to be tested.
2. After selection of the profile, the suitable attributes (i.e. features) are selected on which the classification algorithm is implemented.
3. The attributes extracted is passed to the trained classifier. The classifier gets trained regularly as new training data is feed into the classifier.
4. The classifier determines the whether the profile is fake or real.
5. The classifier may not be 100% accurate in classifying the profile so; the feedback of the result is given back to the classifier.
6. This process repeats and as the time proceeds, the no. of training data increases and the classifier becomes more and more accurate in predicting the fake profiles.

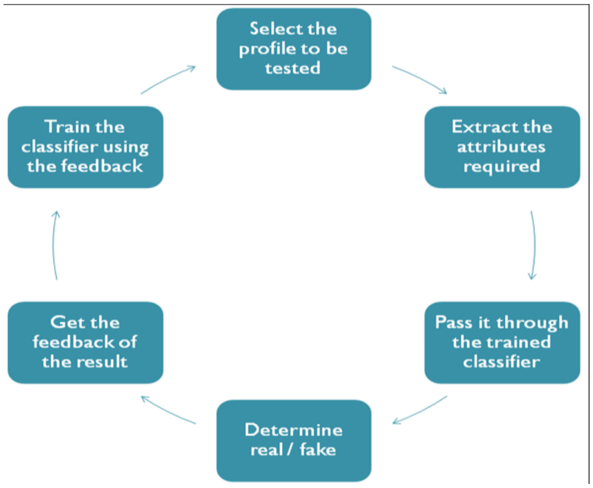

Select the profile to be tested
Extract the attributes required
Pass it through the trained classifier
Determine real / fake
Get the feedback of the result
Train the classifier using the feedback

Figure 3.1: Framework for detection of fake profiles and learning

C. Classification

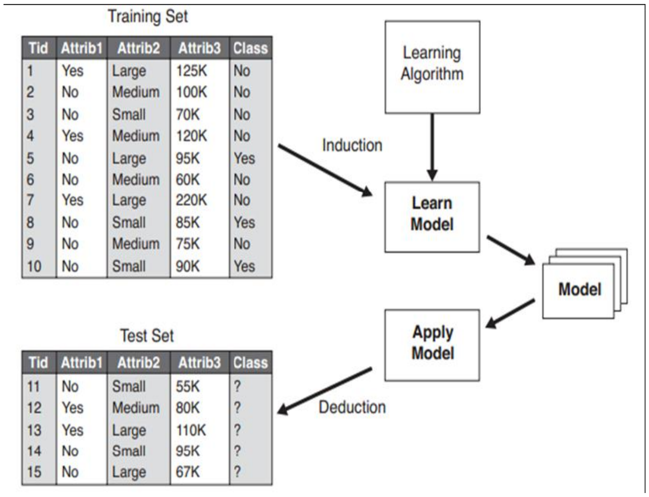
Classification is the process of learning a target function f that maps each records, X consisting of set of attributes to one of the predefined class labels, y.

A Classification technique is an approach of building Classification models from an input data set. This technique uses a learning algorithm to identify a model that best fits the relationship between the attribute set and class label of the training set. The model generated by the learning algorithm should both fit the input data correctly and correctly predict the class labels of the test set with as high accuracy as possible. The key objective of the learning algorithm is to build the model with good generality capability.

The classifiers that we have implemented for classifying the profiles are:

_ Naive Bayes Classification
_ Decision Tree Classification
_ Support Vector Machine

1) *Naive Bayes Classification:* In Bayesian classification we have a hypothesis that the given data belongs to a particular class. We then calculate the probability for the hypothesis of being true. This is among the most practical approaches for certain types of problems. The approach requires only one scan of the whole data. Also, if at some stage additional training data is added then each training example can incrementally increase or decrease the probability that the hypothesis is correct.



Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Figure 3.2: General approach for building a classification model

2) *Decision Tree:* A decision tree is a popular Classification method that generates tree structure where each node denotes a test on an attribute value and each branch represents an outcome of the test. The tree leaves represent the classes. The figure 3.3 shows the decision tree evaluated from our training dataset used in the project. It displays the relationships found in the training dataset. This technique is fast unless the training data is very large. It does not make any assumptions about the probability distribution of the attributes value. The process of building the tree is called induction.

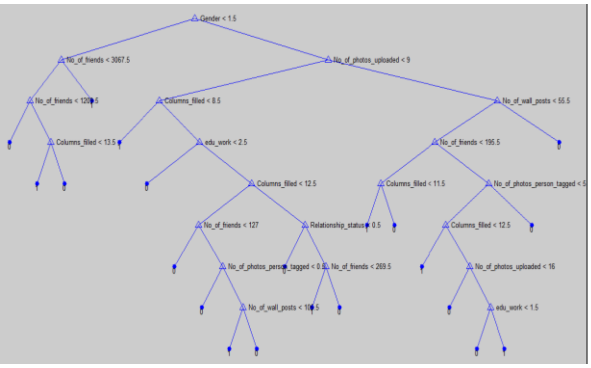


Figure 3.3: The decision tree produced from the training dataset

3) *Support Vector Machine:* An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The best hyperplane for an SVM means the one with the largest margin between the two classes. An SVM classifies data by finding the best hyperplane that separates all data points of one class from those of the other class. The support vectors are the data points that are closest to the separating hyperplane.

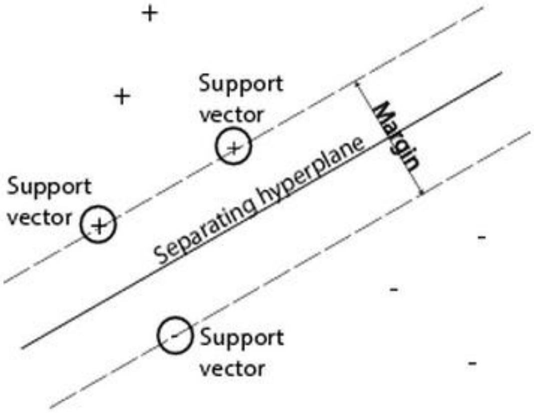


Figure 3.4: Support Vector Machine classification for 2 dimensional data

IV. Implementation

A. *Datasets needed*

We need dataset with a mixture of real and fake profiles labeled accordingly. The algorithms need to be trained using the training dataset and should be evaluated using the testing dataset. But there are no such datasets available because of privacy issues. As there is no standard dataset present, we need to prepare the dataset by scrapping the profiles from Facebook. To scrap the data from the profiles, we need to be friends with the profiles which are being scrapped. We used the profile Facebook/Nitreddy with 957 friends to scrap the real profiles.

B. *Scrapping data*

Scripts written in python language were used which logs into Facebook automatically and scraps required data. Facebook Graph API is also used along with python to extract some required data. Anti scrap detection techniques were implemented to prevent Facebook immune system from detecting. 957 profiles were scrapped out of which some profiles were hiding data from friends also which were removed from the dataset which left 872 real profiles in the dataset. Barracuda labs is presently working on Facebook spam detection making applications for them. They detected and scrapped 350 fake profiles and analyzed the data. We collected the data from them, filtered the profiles in which data is hidden, leaving 290 fake profiles in the dataset

C. *Attributes that we have considered*

- No. of friends
- Education and work
- Gender
- No. of columns filled in about me
- Relationship status
- No. of wall posts posted by the person
- No. of photos uploaded by the person

1) Why only these attributes? In the fake profiles dataset given by Barracuda labs, these were the only attributes we were able to extract.

Some other attributes which can be used in these classification algorithms are:

- Ratio of same gender friends and total friends.
- Ratio of the no. of friend requests sent and accepted
- No. of groups
- No. of likes, etc.

D. *Evaluation parameters*

$$\text{Efficiency} = \frac{\text{No. of correct predictions}}{\text{Total No. of Predictions}}$$

$$\text{False Positive rate} = \frac{\text{No. of real profiles detected fake}}{\text{Total No. of fake profiles to be Detected}}$$

$$\text{False Negative rate} = \frac{\text{No. of fake profiles detected real}}{\text{Total No. of real profiles}}$$

V. Results

1) The efficiency of the SVM is highest when the data is well trained.
2) The efficiency of the Nave Bayes is lowest which don't change much when the training dataset increases.
3) As the no. of attributes increases for the training dataset the efficiency of all the algorithms increases.
4) The false positive rate of the SVM is least whereas Nave Bayes shows high false positive rate.

VI. CONCLUSION AND FUTURE SCOPE

We have given a framework using which we can detect fake profiles in any online social network with a very high efficiency as high as around 95%. Fake profile detection can be improved by applying NLP techniques to process the posts and the profile.

ACKNOWLEDGMENT

I am grateful to numerous people who have contributed towards shaping this research paper. At the outset, I would like to express my sincere thanks to Prof. W. R. Parulekar for his advice during our research paper work. As my supervisor, he has constantly encouraged me to remain focused on achieving my goal. His observation and comments helped me to establish the overall direction of the research and to move forward with investigation in depth. He has helped me greatly and been a source of knowledge. I sincerely thank everyone who has provided us with new ideas, constructive criticism, and their invaluable time. Last but not the least; I would like to thank all my colleagues and friends for their help and cooperation

REFERENCES

[1] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In Proceedings of the 4th Workshop on Social Network Systems, SNS, volume 11, page 8, 2011.

[2] Y. Boshmaf, I. Muslukhov, K. Beznosov, and M. Ripeanu. The socialbot network: when bots socialize for fame and money. In Proceedings of the 27th Annual Computer Security Applications Conference, pages 93{102. ACM, 2011.

[3] C. Wagner, S. Mitter, C. Korner, and M. Strohmaier. When social bots attack: Modeling susceptibility of users in online social networks. In Proceedings of the WWW, volume 12, 2012.

[4] G. Kontaxis, I. Polakis, S. Ioannidis, and E.P. Markatos. Detecting social network profile cloning. In Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on, pages 295{300. IEEE, 2011.

[5] A. Wang. Detecting spam bots in online social networking sites: a machine learning approach. Data and Applications Security and Privacy XXIV, pages 335{342, 2010.

[6] G.K. Gupta. Introduction to Data Mining with Case Studies. Prentice Hall India, 2008.

[7] Spies create fake facebook account in nato chief's name to steal personal details, http://in. news.yahoo.com/spies-create-fake-facebook-account-nato-chiefs-name-114824955.html.

[8] Man arrested for uploading obscene images of woman colleague, http://www.ndtv.com/ article/andhra-pradesh/man-arrested-for-uploading-obscene-images-of-woman-colleague-173266.

[9] How obamas internet campaign changed politics, /bits.blogs.nytimes.com/2008/11/07/how-obamas internet-campaign-changed-politics.

[10] M. Huber, M. Mulazzani, and E. Weippl. Who on earth is mr. cypher: Automated friend injection attacks on social networking sites. Security and Privacy{Silver Linings in the Cloud, pages 80{89, 2010.

Authors Profile

Mr. R. V. Kotawadekar pursed Bachelor of Engineering in Information Technology from University of Mumbai, in 2008 and Master of Computer Applications from University of Mumbai in year 2015. He is currently pursuing Ph.D. and currently working as Assistant Professor in Department of MCA, Finloex Academy of Management and Technology, Ratnagiri since 2016. His main research work focuses on Machine Learning, Artificial Intelligence, Cloud Computing, Cloud Security and Privacy, Big Data Analytics and IOT. He has 3 years of teaching experience and 1 year of Research Experience.

Mr A. S. Kamble pursed Bachelor of Science (Computer Science) and is currently pursuing Master in Computer Applications from Uinversity of Mumbai at Finolex Academy of Management and Technology, Ratnagiri.

Ms S. A. Surve pursed Bachelor of Science (Information Technology) and is currently pursuing Master in Computer Applications from Uinversity of Mumbai at Finolex Academy of Management and Technology, Ratnagiri.