Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2022)
IEEE Xplore Part Number: CFP22N67-ART; ISBN: 978-1-6654-9707-7

Profile Similarity Recognition in Online Social Network using Machine Learning Approach

Soumya.T.R[1]
Assistant Professor, Dept. of CSE
Prathyusha Engineering College
Thiruvallur, TamilNadu
amburgps@gmail.com

S.Solai Manohar[2],
Professor, Deptartment of EEE,
CMR Institute Of Technology,
Bengaluru, Karnataka
ssmanohar76@gmail.com

N.Bala Sundara Ganapathy[3]
Professor, Department of IT,
Panimalar Engineering College,
Chennai, TamilNadu
balabsg@gmail.com

Leema Nelson[4]
Department of Computer Science & Engineering,
Chitkara University Institute of Engineering & technology
Chitkara University, Punjab, India.
leema.nelson@chitkara.edu.in

A. Mohan[5]
Department of Information Security,
SIMATS Schools Of Engineering,
Saveetha University, Tamilnadu.
annamalaimohan@gmail.com

M.Thurai Pandian[6*]
School of CSE,
Vellore Institute of Technology,
Vellore, India.
Thuraipandianm@gmail.com

2022 4th International Conference on Inventive Research in Computing Applications (ICIRCA) | 978-1-6654-9707-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIRCA54612.2022.9985683

Abstract— **Now-a-days, On-line Social Networks (OSN) plays a major role. It helps to maintain a social bond across world. People can share their information like personal, political and other career related work within minutes. High number of people started indulging in social networks because of its time complexity and easiest way of sharing details. On the other hand, the threats on social networks increased in a rapid way to gain sensitive information of a people sharing important details like passwords, bank details, personal details etc. The online privacy and security threats reports gone to top. It made researchers to involve more on this area of threats in order to protect individual privacy and security regarding social networks. Attackers started to exploit the vulnerabilities of OSNs which highly attracted malicious entities. Each OSNs had ten to hundred million users collectively generate billions of personal sensitive data content. To protect users from exploitation is one of the major challenges. The current research are mainly focused to protect the privacy of an individual's online profile in OSN. The creator constructs a fake profile to pretend himself as a genuine person in these sites. The aim in this paper is to check and detect the trusted and fake profiles created in OSN through machine learning based techniques. By using algorithms like (SVM)Support vector machine, (DNN) Deep neural network, Random Forest to overcome processes such as treating missing value, identification of variable, cleaning and validation of data will be done on the entire dataset. Here, python language is used for improving the efficiency of code. With help of this, millions of fake profiles can be detected automatically. Additionally, various machine learning algorithms are compared and discussed along with the classification report. By identifying the confusion matrix, the results show the effectiveness of the proposed machine learning algorithm, which has best accuracy, precision, Recall and F1 score.**

Keywords— Dataset, Pre-processing, Classification, Deep neural network.

*Corresponding author

I. INTRODUCTION

In this century, online social network are playing an essential role and became an important part of communication through internet. They provide all type of information which involves real-time information. So larger number of people got influenced and started depending to the entire social networking sites according to their needs. Even aged people too started creating their social networking sites to meet their instant needs regarding health and for also important contacts. The entire dependency towards social networks leads to a major occurring of threats across internet. These threats leads to a great disadvantage on people's side who is not that much aware of what happening behind the web. It may cause serious fears towards the individual and creates lack of confidence while sharing the information. This will cause lack of privacy of an individual. Hence fake accounts are brought into picture in these cases by various researchers and analysts. They started working on this serious problem in order to find fake account across social networks. Various privacy setting features are added for each post in most of the social sites. However, still some private information can be leaked. For example, consider when a particular person posting their picture in social network. Even though they adjusted the feature of privacy settings made to be viewed only by his familiar friend. There is no guarantee that it is perfectly safe or it may be shared by his friend to another person by a fake profile without user's knowledge. It is a very simple and instant process to create a fake profile which may influence the people maliciously and create harm to their privacy by simply knowing their activities through social engineering.

This paper creates a system based model helps to recognize the illegal account in online social media networks. Our first task is to get a dataset for our model. Based on the requirements, the dataset will be collected from various resources available on the internet. The collected dataset is not consistent and structured. To make the data more consistent and structured, data pre-processing can be used. This method helps to make the data more perfect to train the proposed model. It involves in performing data cleaning which allow us to found out the useful data for our system. The upcoming sections are ordered as follows: The existing work is showed in section II. Explanation of the steps regarding the cleaned dataset from raw unstructured data through pre-processing is given in section III. We provide our algorithms which is used to detect whether the profile is fake or true one in section IV. The fake profile recognition model by comparison report is shown in section V. Finally, we conclude our study in section VI.

II. LITERATURE SURVEY

The unconventional model by machine learning and NLP (Natural Language Processing) techniques done by an author to identify fake profiles increasing rate in online social networks. By the help of some important information such as profile photo of an individual, name for authentication purposes which mainly relayed in social networks with weak user [1]. The suspicious behavioral activities are detected using daily social media user mobility data. Two characteristics are taken from mobility of a user. DBSCAN algorithm, is given in this model for this purpose. The results show that proposed algorithm could learn normal daily activities of social media users and detect anomalous activitiess [2] the study is relayed on real datasets which has both genuine and spam profiles in Facebook and Twitter. They found set of 14 generic statistical characteristics to found spam profiles [3]. This paper takes behavior, reactions, connection of an individual and they get validated on OSN. The data gives back the entire individual usage and connectivity on particular sites [4]. In this paper cloned attacks are examined and done by making initially in form of graphs based on similarities [5]. This paper takes behavior, reactions, connection of an individual and they get validated on OSN. The data gives back the entire individual usage and connectivity on particular site [6]. To identify fake and attackers who fetch user's private social information in order to gain their trust they used profile cloning detection technique which generate more possibilities. Two novel techniques are suggested by an author which are related to profile cloning detection technique. First one focuses on coincidence in attributes concerning the two profiles but in the other method focuses on same related relationship networks [7]. This paper done with datamining techniques of social network with the logistic generation algorithm to find suspicious users by limit of threshold parameters [8]. This paper reviews the Sybil attack in social networks, which has the potential to compromise the whole distributed network. In the Sybil attack, the malicious user claims multiple identities to compromise the network. This attacks which majorly create impact on the full voting rank and which access. Various defense mechanisms are used to minimize these attacks [9]. In this study, they worked to make clear problems which arise on OSN. It helps both organizations and employee from attacks. [10]. In this paper data hiding technique is used which hide in profile photo that identifies fake profile in various forms like cryptography and watermarking [12]. Piece of work is done by an author by checking each and every individual in a unique way by writing style or about their specific social behavioural pattern [13].

III. PROPOSED WORK

The proposed system model finds the fake profiles in OSN by using machine learning technologies. In order to find out the fake profiles anybody may get suffered by the suspicious account. The dataset is taken from different sources which is inconsistent. The dataset includes various attributes like number of friends count, followers count, listed count, statuses count, favorites count, and sex. Generalized dataset is applied to extract patterns and to acquire results with best accuracy. The report will get loaded in the data and will check for cleanliness, and it go for trimming and cleaning of dataset for analysis. Notice our document steps carefully for cleaning processes. Prediction data is mainly classified as training and testing set. Basically, 8:2 ratios are applied to divide the two sets. Then we will apply machine learning algorithms to gain accuracy. In Facebook, with the help of the user followers and with other attributes we can find whether the profile is fake or good one. Classification algorithm results get verified and feedback is put back into the classifier. When the count of training data increases the classifier started to look in an more accurate way which helps in predicting the fake profiles. We compare our algorithms to increase our accuracy result. Our main objective is to get higher accuracy of predicting fake profile in online social networks.

Process:

1. Collecting raw datasets and pre-processing the datasets.
2. Data Validation to predict the fake and genuine profiles.
3. Splitting dataset
   a) Training dataset
   b) Testing dataset.
4. Applying algorithms
   a) Deep neural networks
   b) Random forest
   c) Support vector machine (SVM).
5. Comparing accuracy results, roc curve and confusion matrix parameters.

These steps are implemented for predicting fake accounts. The pros are: The social networking sites are making our lives better but on the other side it creates various serious threats. The issues arise for privacy, online phishing, scamming, potential to misuse etc. These are mostly relayed on fake profiles. In this project, we implemented a model through which we can detect a fake profile using machine learning algorithms which provide security to our social lives. Table 1 denotes attributes and distributions of social network data.

Table 1. Attribute and descriptions

Attribute	**Explanation**
Status Count	The status average number created by an individual are likely to be in low when number the account is fake.
Friends Count	The friends count in individual account.
Followers Count	Unusual links, advertisements are shared and posted which creates low count of followers.
Favorites Count	The count of favorite denotes the user interest in the entire lifetime of the account.
Listed Count	The specified individual member counts in public lists.
Sex Code	The gender of the account holder.
Language code	The language of the account holder. The creation date, and the usage of the timeline for less period of time.

PROPOSED METHODOLOGY:
Figure 1 is represents as our proposed architecture, and it depicts the framework to recognize the account profile which is fake in social networks.

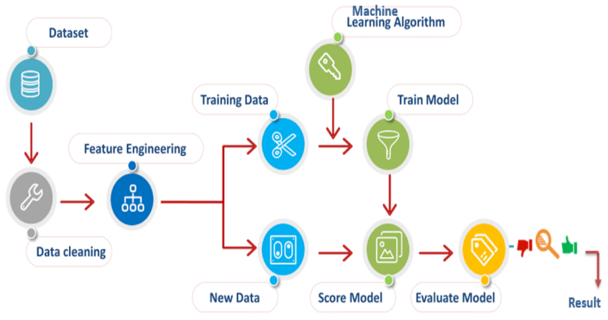

Figure 1.Framework of account profile

Initially we gather the fake and genuine user's datasets and loading process is made for data cleaning process. Next process we will start trimming and cleaning processes of given datasets. The array values get converted into numerical values and collected dataset for predicting given data is divided into Training and Test set. 8:2 ratios are applied to split these two sets. Then we create a data model with the trained data. Then machine learning techniques such as, Support Vector Machine, Random forest and Deep Neural Network are applied. Based on various attributes that we can find the profile is fake or genuine. The classification algorithm result is then verified and feedback is given back into the classifier. When the count of training data increases the classifier becomes more accurate while predicting the result and the evaluation of the parameters is made by confusion matrix, accuracy, ROC graph and classification report is done.

DEEP NEURAL NETWORK: DNN has large number of neurons arranged in a sequence of multiple layers, where input is received to the neurons from the previous layer and simple computation is performed named (ANN) Artificial Neural Network. It can be done by supervised, semi supervised or unsupervised learning methods. It get some simple inputs, each shows a weight of each important thing, and generate an output decision is answered as "0" or "1".

SUPPORT VECTOR MACHINE: It helps in two-group classification problems. It is widely used in classification problems. At first we find data points. The hyperplane differentiates two classes very well based on its nature. Then we can find the efficient right efficient hyperplane to process our model. It is helps to find the maximum margin (the maximum distance between data points of two classes we defined).

RANDOM FOREST: It is one of the flexible algorithm used for both classification and regression tasks. It groups various types of algorithms or same algorithm for larger number of times. It is processed by constructing a certain section of decision trees which results in a forest of trees, hence it is named as "Random forest". It also overcomes limitations of decision trees. Steps: N random records are taken from the dataset which is cleaned. A decision tree is build based on N records. And the number of trees choosed. Steps 1 and 2 are repeated.

IV. RESULTS

EVALUATION PARAMETERS:

Percent Error = (1- Accuracy)*100

Accuracy= Count of total correct predictions/total count of predictions.

Confusion matrix: The performance of a classification model on a test data whose true values are known is described by confusion matrix.

I. TPR = TP / (TP + FN)
II. FPR = FP / (FP + TN)
III. TNR= TN / (FP + TN)
IV. FNR = 1-TPR

Precision = TP / (TP + FP)
Precision: The positive predictive value that is really correct. The model predicts the fraction of pertinent instances among retrieved instances.

Recall = TP / (TP + FN)
Recall: The proportion of positive observed values which are correctly predicted. The model predicts the fraction of pertinent instances that are retrieved.

The weighted average of Precision and Recall is F1 score. Both FP and FN are taken in these scores.

Roc curve: It is a graph shows the performance of a model that is classified at all threshold classification. It has two features:

1. True Positive Rate
2. False positive rate.

RANDOM FOREST:

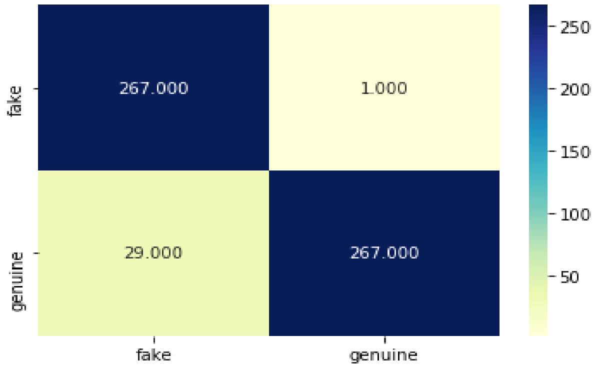

Figure 2 - Confusion matrix

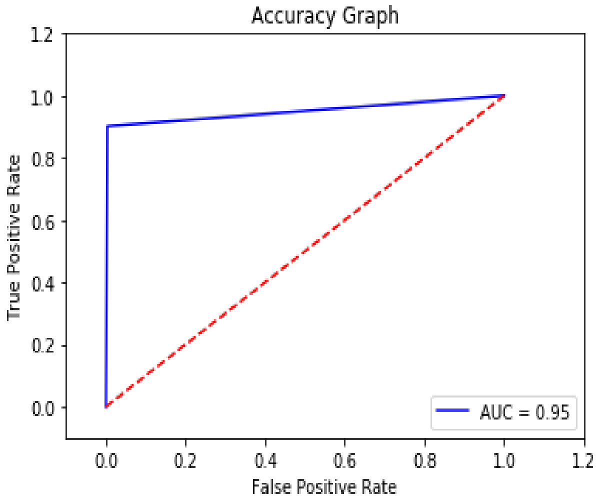

Figure 3 - Roc Curve

96% is acquired while classifying the efficiency of data in random forest. For training 80%of data is taken in Random forest classifier and 20% is for classification.

SUPPORT VECTOR MACHINE:

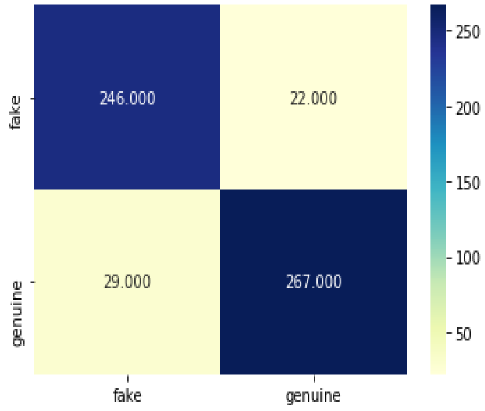

Figure 4 - Confusion matrix

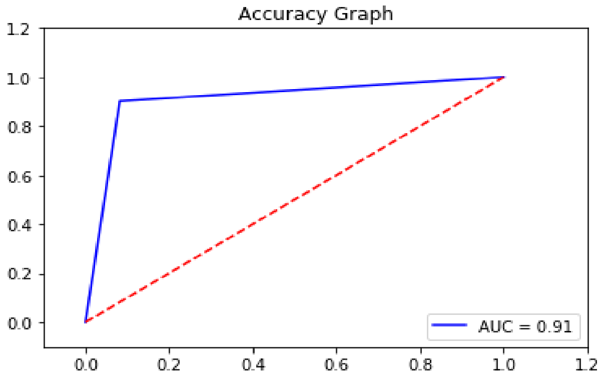

Figure 5 - Roc Curve

SVM efficiency in classifying data is about 91%. 80% of data is taken for training SVM classifier and classification is for 20%.

DEEP NEURAL NETWORK:

Table 2: Accuracy Results Comparison

PARAMETERS	RANDOM FOREST	SUPPORT VECTOR MACHINE(SVM)	DEEP NEURAL NETWORK(DNN)
Precision	0.90	0.89	0.99
Recall	0.99	0.92	0.93
F1-Score	0.95	0.91	0.96
Accuracy percent	95	91.57	96.09

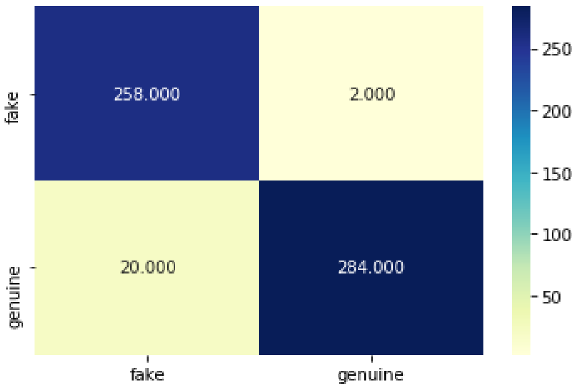

Figure 6- Confusion Matrix

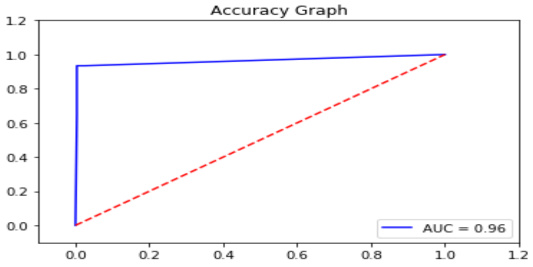

Figure 7 - Roc Curve

DNN efficiency in classifying data is about 96%. 80% of data is taken for training DNN classifier and classification is for 20%.

V. CONCLUSION

The analytical process of our work started from data collecting, pre-processing, identification of missing value, analysis and final model building and evaluation is done. With help of Deep neural network (DNN), Random forest and Support vector machine (SVM) algorithms of machine learning approach helps in detecting various accuracy results. By comparing the results. The best result is deep neural network (DNN) algorithm (96%). The desire is to add new features which will further detect and identify processing techniques more in future.

VI. FUTURE ENHANCEMENT

Fake accounts can be easily identified when Facebook updates new features. One person may have various number of accounts in Facebook which leads to a creation fake accounts. To overcome this we can add Aadhar card number while getting into a user's account will create unique identity for each individual in order to protect their true identity. By this we can easily restrict the creation of a single account rather than multiple accounts for each individual and creation of fake accounts can be easily deployed.

VII. REFERENCES

1. Pulluri, S.R., Gyani, J. and Gugulothu, N., 2017. A comprehensive model for detecting fake profiles in online social networks. International Journal of Advanced Research in Computer and Communication Engineering, 6(6).
2. Zheng, X., Zeng, Z., Chen, Z., Yu, Y. and Rong, C., 2015. Detecting spammers on social networks. *Neurocomputing*, *159*, pp.27-34.
3. Ahmed, F., & Abulaish, M. (2013). "A generic statistical approach for spam detection in online social network" Computer Communications. Vol. 36 (10–11), pp. 1120–1129.
4. Kharaji, M. Y., & Rizi, F. S. (2014). "An IAC Approach for Detecting Profile Cloning in Online Social Networks",https://arxiv.org /abs/1403.2006.
5. G. Narasimha Murthy, M.Eranna, (2017)" Detection of Behavioral Abnormality of Compromised Accounts in OSN'S", International Journal of Innovative Research in Science, Engineering and Technology", Vol. 6, Issue 10.
6. Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini,A. (2016). The rise of social bots. Communications of the ACM, 59(7), 96-104.
7. Piotr Bródka, Mateusz Sobas, Henric Johnson, "Profile Cloning Detection In Social Networks", European Network Intelligence Conference, 2014, pp. 63 – 68.
8. Girisha Khurana, Mr Marish Kumar, "Review: Efficient Spam Detection on Social Network", IJRASET, Vol. 3, Issue. 6, 2015, pp. 76 – 80.
9. Gunturu. R, "Survey of Sybil Attacks in Social Networks," https://arxiv .org/abs/1504.05522, 2015.
10. Mahmood.S, "Online Social Networks: Privacy Threats And Defenses," Security and Privacy Preserving in Social Networks, 2013, pp. 47–71,Springer.
11. G. Parthasarathy and DC.Tomar, "Trends in citation analysis", Intelligent Computing, Communication and Devices, Springer, New Delhi, pp.813-821, 2015.
12. N, SaiSupriya, Rashmi S, Parthasarathy G and Priyanka, "Face Mask Detection Using CNN." Smart Intelligent Computing and Communication Technology 38,pp.118, 2021.
13. Parthasarathy G, Soumya T.R.,Ramanathan L and Ramesh P, "Improvised Approach for Real Time Patient Health Monitoring System Using IoT." In Intelligent Systems and Computer Technology, pp. 78-83. IOS Press, 2020.
14. R. Majji, P. G. O. Prakash, R. Cristin, and G. Parthasarathy, "Social bat optimisation dependent deep stacked auto-encoder for skin cancer detection," Iet Image Processing, vol. 14, no. 16, pp. 4122-4131, Dec 2020.
15. Lazar, A.J.P., Sengan, S., Cavaliere, L.P.L. et al. Analysing the User Actions and Location for Identifying Online Scam in Internet Banking on Cloud. Wireless Pers Commun (2021). https://doi.org/10.1007/s11277-021-08585-y
16. Raj, Jennifer S., and Mr C. Vijesh Joe. "Wi-Fi Network Profiling and QoS Assessment for Real Time Video Streaming." IRO Journal on Sustainable Wireless Systems 3, no. 1 (2021): 21-30.
17. Hamdan, Yasir Babiker. "Faultless Decision Making for False Information in Online:A Systematic Approach." Journal of Soft Computing Paradigm (JSCP) 2, no. 04 (2020): 226-235.
18. Navaneethakrishnan, M., S, V., Parthasarathy, G. & Cristin, R. (2021). Atom search-jaya-based deep recurrent neural network for liver cancer detection. IET Image Processing, 15(2):337–349. doi: 10.1049/ipr2.12019.