International Journal of Research (IJR)
e-ISSN: 2348-6848
p-ISSN: 2348-795X
Vol. 9 Issue 06
June 2022

Identifying Of Fake Profiles Across Online Social Networks By Using Neural Network

Tamtam Sunil Goud[1], Gorla Yaswanth[2], Shaik Sohail Ahammed[3] , Sirisha Kamsali[4], M. Sri Lakshmi[5]
[1,2,3] U.G. Scholar, [4]Guide Assistant Professor, [5]Head of the Department
[1,2,3,4,5] Computer Science And Engineering
[1,2,3,4,5] G. Pullaiah College Of Engineering And Technology
Email : [1]sunilgoudmjss@gmail.com , [2]vkrish783@gmail.com , [3]sohailahammed@gmail.com
[4]sirisha@gpcet.ac.in

Abstract

In seeing the present condition, online social networks are engaging with the majority of the people. From child to adult, all are spending a considerable time on these platforms either by exchanging information or making efficient communication with others. But nowadays, these social networking sites are suffering from a lot of fake accounts in taking advantage of vulnerabilities, either taking the benefits or targeting accounts attempting cybercrimes.

Keywords : K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, Support Vector Machine(SVM)

Introduction

Malware is a universal concept for all types of software attacks. With the fast growth of technology, malware is one of the most significant security threats [1]. Any program designed to infiltrate or harm a computer system with infecting a legitimate user's computer, such as information stealing or spying is considered a malware. Malicious software can be categorized into different classes depending how they attempt to harm or behave such as Trojan, Virus, Rootkit, Worm and Spyware [13]. While the family of malware is developing, anti-virus detectors seem unable to satisfy critical requirements, appearing in millions of computer software being threatened. According to Kaspersky Labs in the year of 2018 1, there was about 5.638.828 different hosts that were attacked. A further article by Juniper Research 2 noticed that more than 33 billion records will be stolen by cybercriminals in 2023 alone. Nowadays, detection of malware is challenging due to the high accessibility of attacking methods on the Internet. Additionally, the high availability of anti-detection methods which gives everyone the chance to originate an attack or malicious software without a certain level of experience. Moreover, attackers are using methods to immediately upgrade to a newer version in a short period of time to avoid detection methods Therefore, malware protection of computer systems is one of the most fundamental tasks for users and organizations, since even a single attack can result in serious damages to data and severe losses. Huge losses and frequent attacks impose the need for reliable and accurate techniques for detection. Malware detection is divided into static analysis, which means the analysis of a compiled file or program and dynamic analysis that means analyzing the run time behavior such as battery consumption, memory reads and writes, and network utilization of the device [2]. Static analysis concerns about reading the source code of malware without the execution of the program file, it tries to find the behavioral attributes of the file such as file format

inspection, string extraction, AV scanning, fingerprinting, and disassembling of the binary format. The dynamic analysis depends on actual-time monitoring of the file while it is being executed, running in a virtual environment. Malware detection techniques can also be divided into signaturebased and heuristics-based [6]. However, the accuracy in this method is not always adequate for detection, resulting in a lot of false-positives and false-negatives. The need for a new detection method is becoming urgent. For this reason, machine learning-based techniques can be invaluable to the safety of the system. Machine learning models can be trained to classify the malware into infected execution and legitimate execution. Malicious software examination requires powerful detection abilities for deciding whether a suspicious file is malicious or not. It is also used in searching for what family a malware is likely to belong to. Machine learning methods can be used to discover what is a normal act in the beginning, and search for anything that might be far of it. Therefore, it can provide protection for users by keeping systems safe and stop attacks much faster than they had in the past. Numerous machine learning methods are observed to be helpful in malware detection and classification some of these techniques are Random Forest (RF), Support Vector Machine (SVM), Na¨ıve Bayes (NB), Logistic Regression (LR) and AdaBoost [3]. In our study, we follow the methodology proposed in reference paper [4]. In this paper, the authors experimentations are based on 306 features extracted from actual behavior of the files (Heuristic) observed by a sandbox. They applied machine learning algorithms on 984 malicious files and 172 benign files to perform binary and multi-class classification. The highest accuracy achieved with the Random Forest model 95.69% for multi-class classification and 96.8% for binary classification. In our work, we are using Random Forest classifier to select the most important features and ignore irrelevant features. Thus, we obtained better accuracy than the referenced paper in [4]. Our experimentations show that the highest accuracy that are achieved by Decision Tree is 98.2% for binary classification and the Random Forest achieved 95.8% for multi-classification, respectively. The rest of this paper is organized as follows. Section 2 presents our related work. Section 3 presents our methodology and datasets followed by the results in section 4. Finally, Section 5 presents our conclusion and highlights our future works.

Research Problem

The concern about fake profile is protecting personal data or information from cyber attacks known as phishing attacks. The cyber attackers are often use this in stealing of information. In detecting of passwords, sharing of irrelevant contents, raising awareness this type of profiles are involved in all unlawful activity. In managing and taking the advantages of the critical situation this can be lead to the anonymity through a longer way. For reducing the incidents like trolling, hacking and cyber bullying this is need to be identified

Research Rationale

In securing the all types of social accounts and keeping the users away from the cyber hackers this is necessary to identify those and using of ANNs model this can improved in more better way.

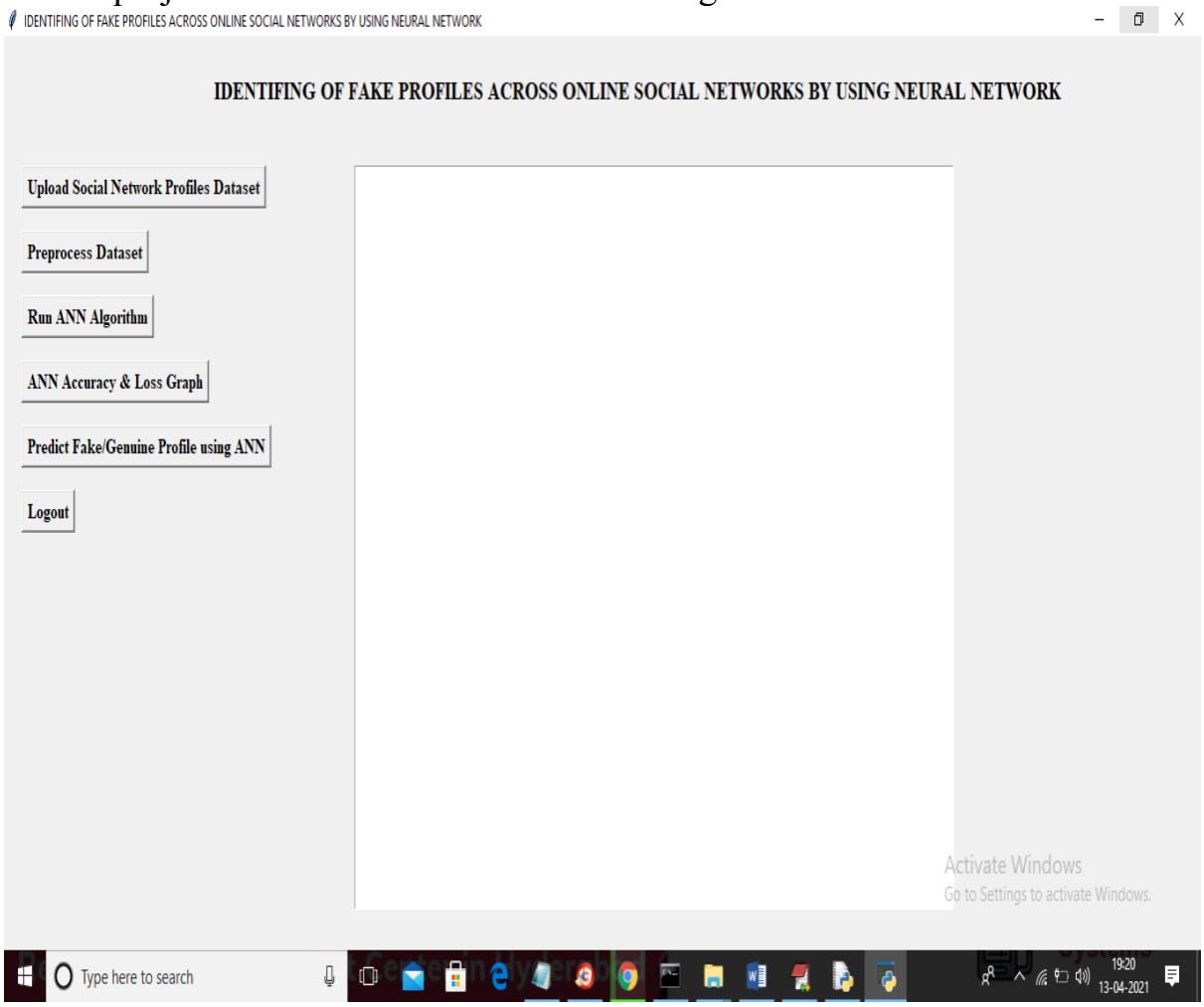
Research Approach

In regards to this, an "artificial neural network" system has been introduced as a part of the computer system. It is designed for simulating in a way in which the human brain possesses and analyses information. The inductive research approach can be considered for this type. In viewing the existing process and situations this can be observed through the patterns and

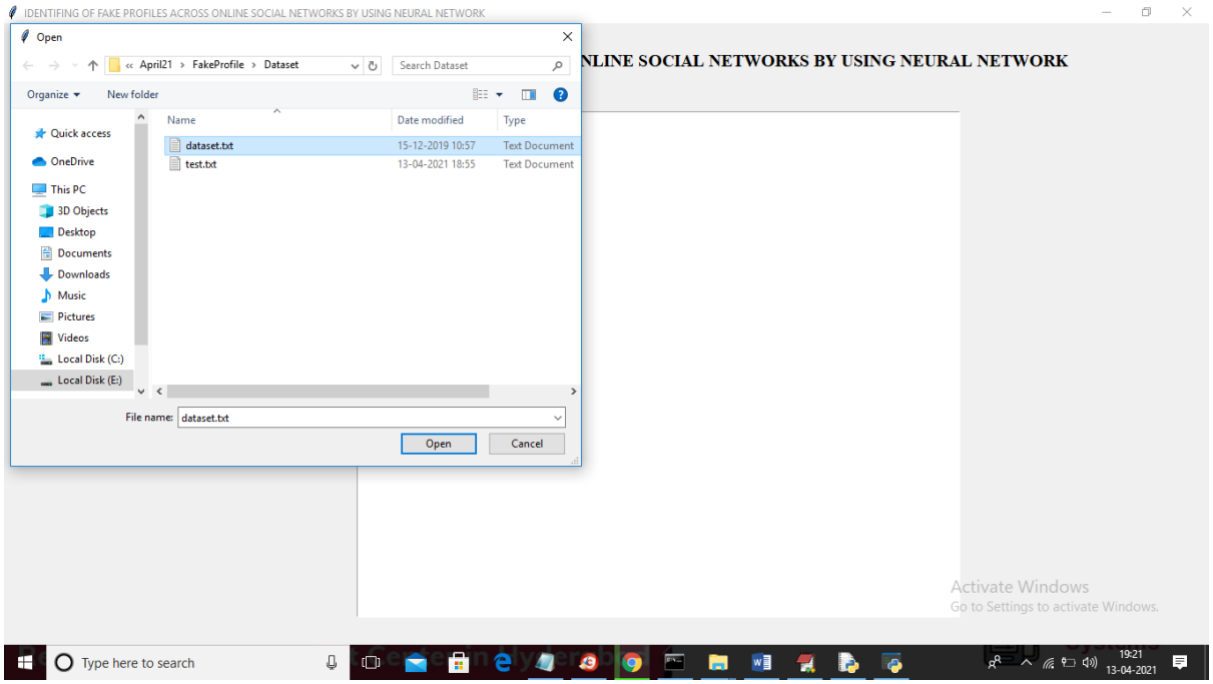
system regularities. In taking the technical advantage ANN model need to be used effectively. It can be described as a foundation of artificial intelligence which will solve the problem in proving the difficulty according to human standards. Therefore "artificial neural networks" (ANNs) are introduced as a process of modeling, allowing the human nervous system through learning technique. By depending on the prediction, this detection process is revealing about the "user-level activities' '. User influence is also vital in reporting about the abnormalities. The social influence upon users can be assessed with the two types of factors. One is to find the user's impact upon others, and the other is to give the user importance. The evaluation is also based on the "fine-grained feature'

SCREENSHOTS

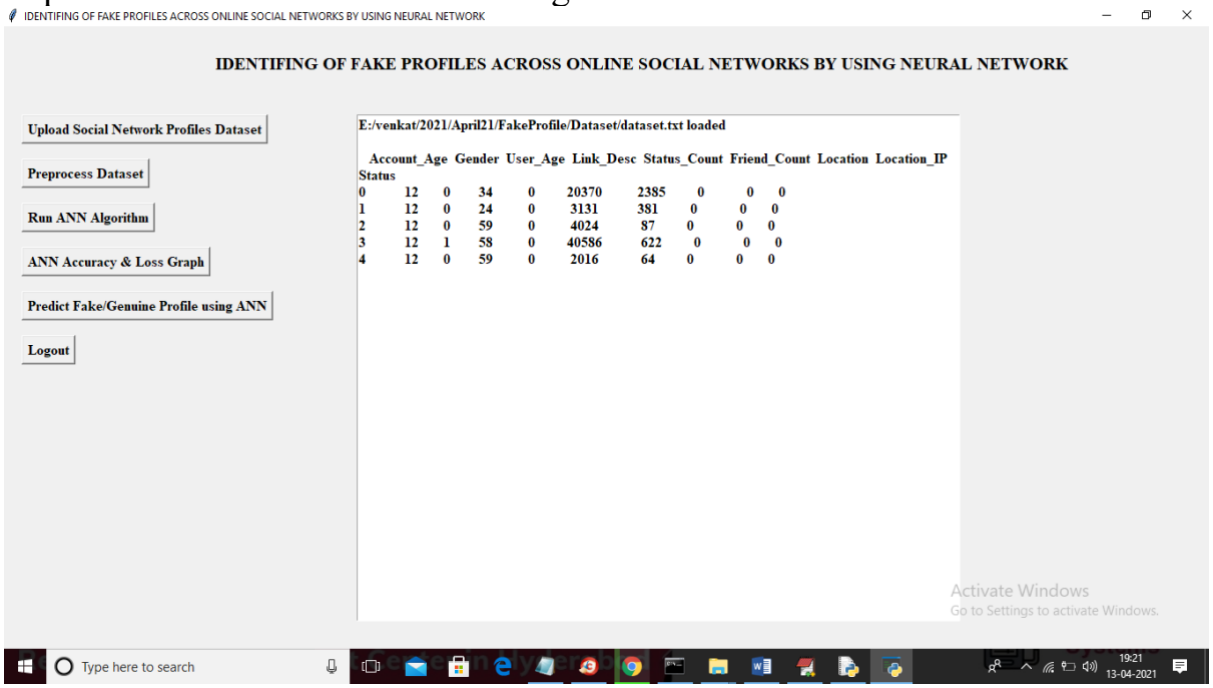
To run project double click on 'run.bat' file to get below screen



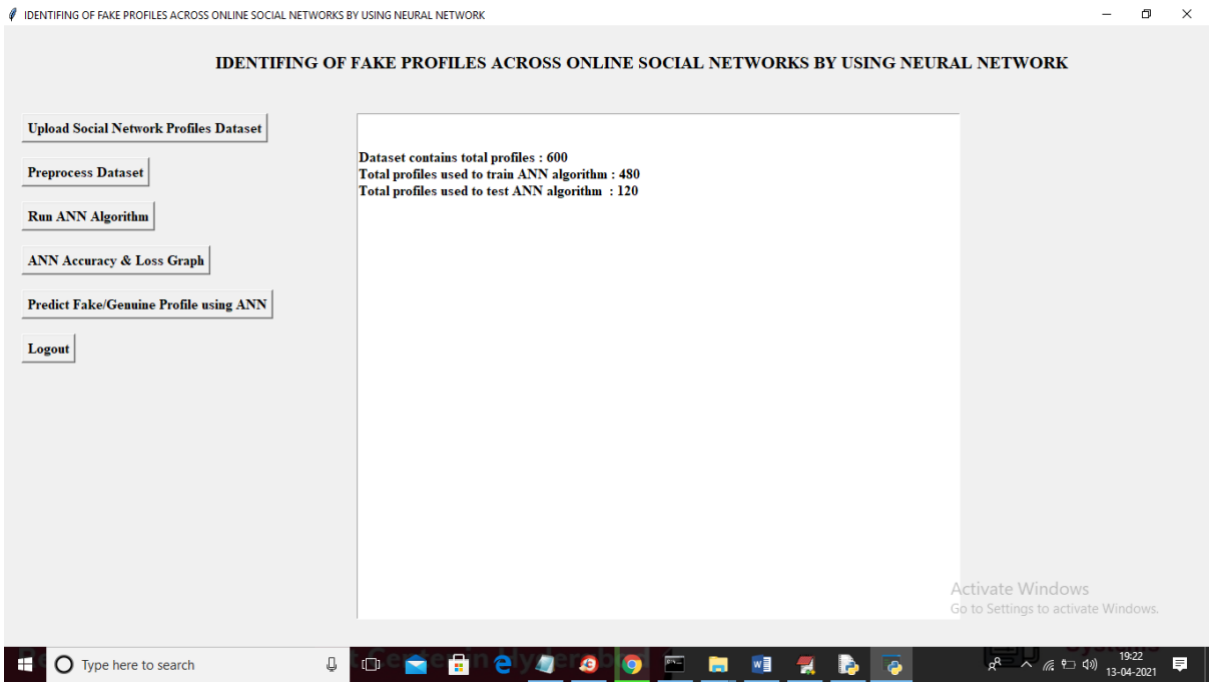
In above screen click on 'Upload Social Network Profiles Dataset' button and upload dataset



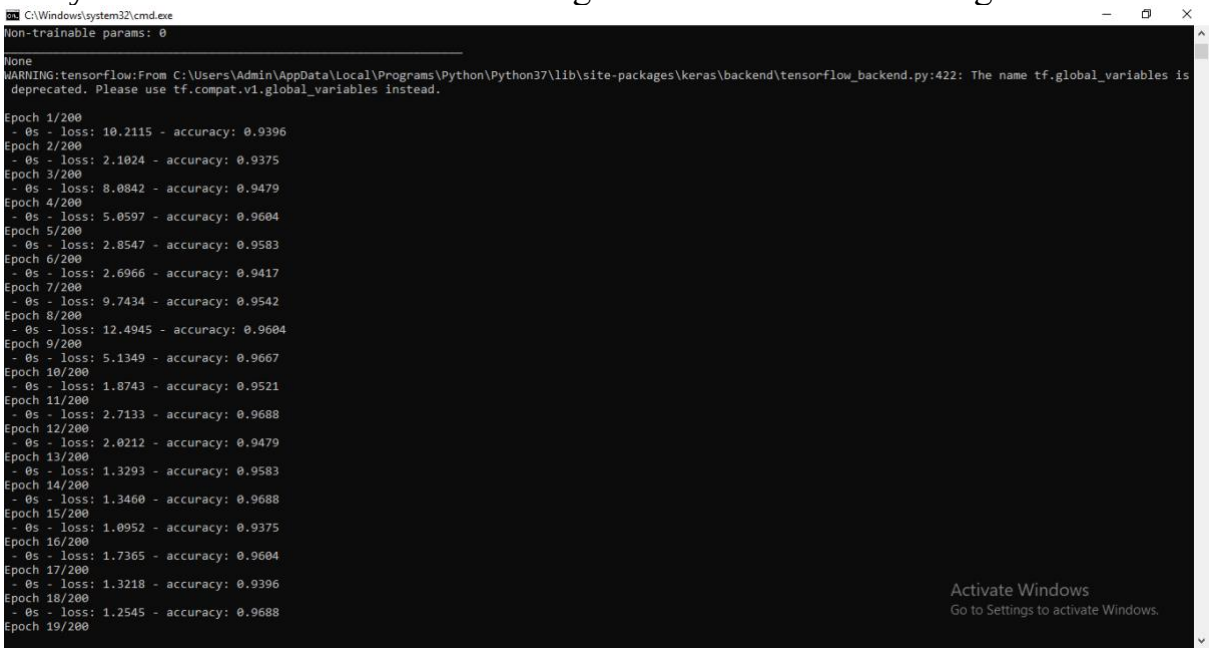
In above screen selecting and uploading 'dataset.txt' file and then click on 'Open' button to load dataset and to get below screen



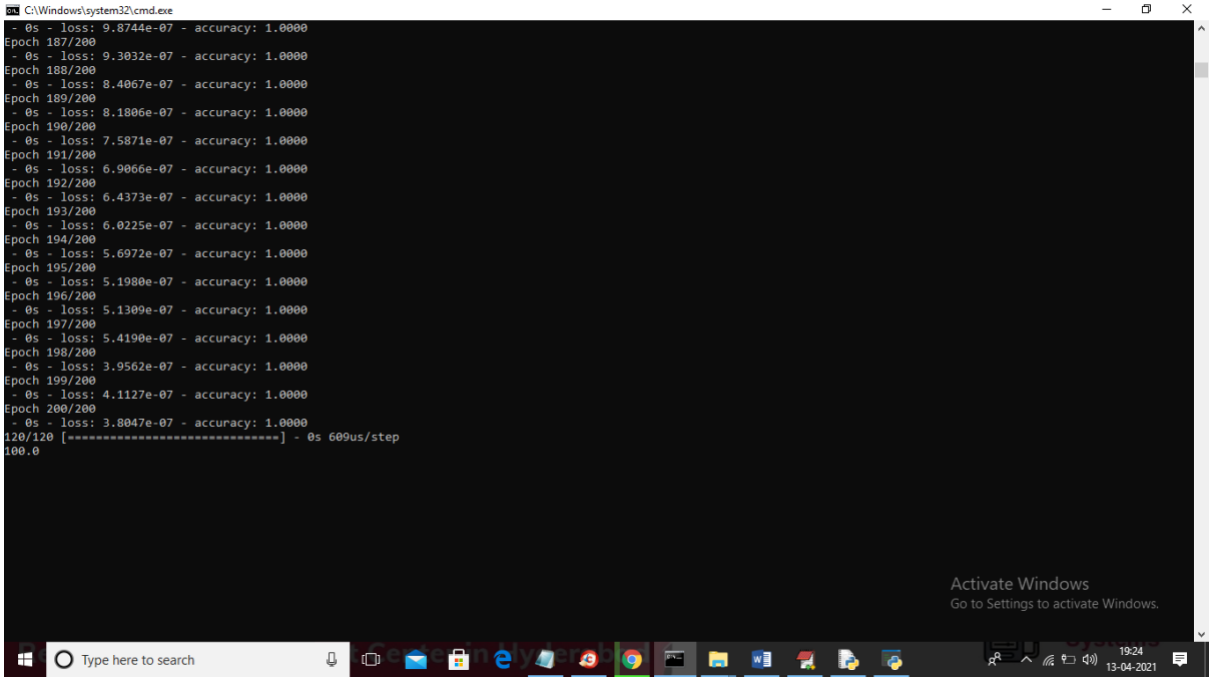
In above screen dataset loaded and displaying few records from dataset and now click on 'Preprocess Dataset' button to remove missing values and to split dataset into train and test part



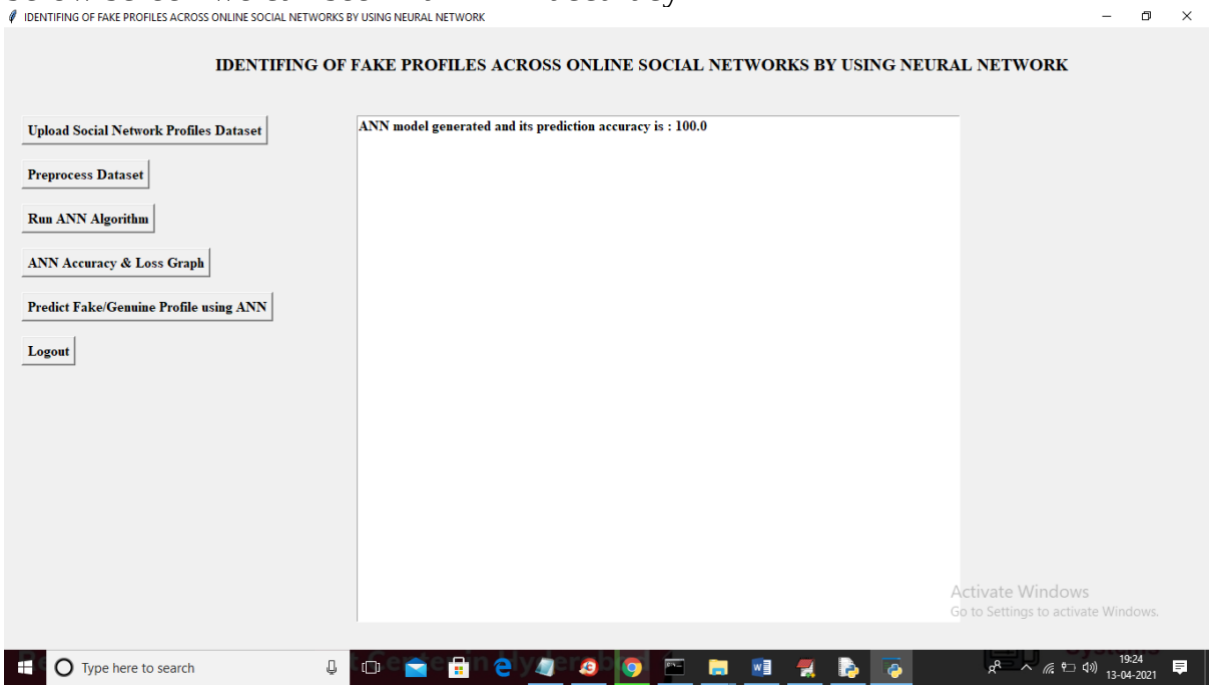
In above screen we can see dataset contains total 600 records and application using 480 records for training and 120 records to test ANN and now dataset is ready and now click on 'Run ANN Algorithm' button to ANN algorithm



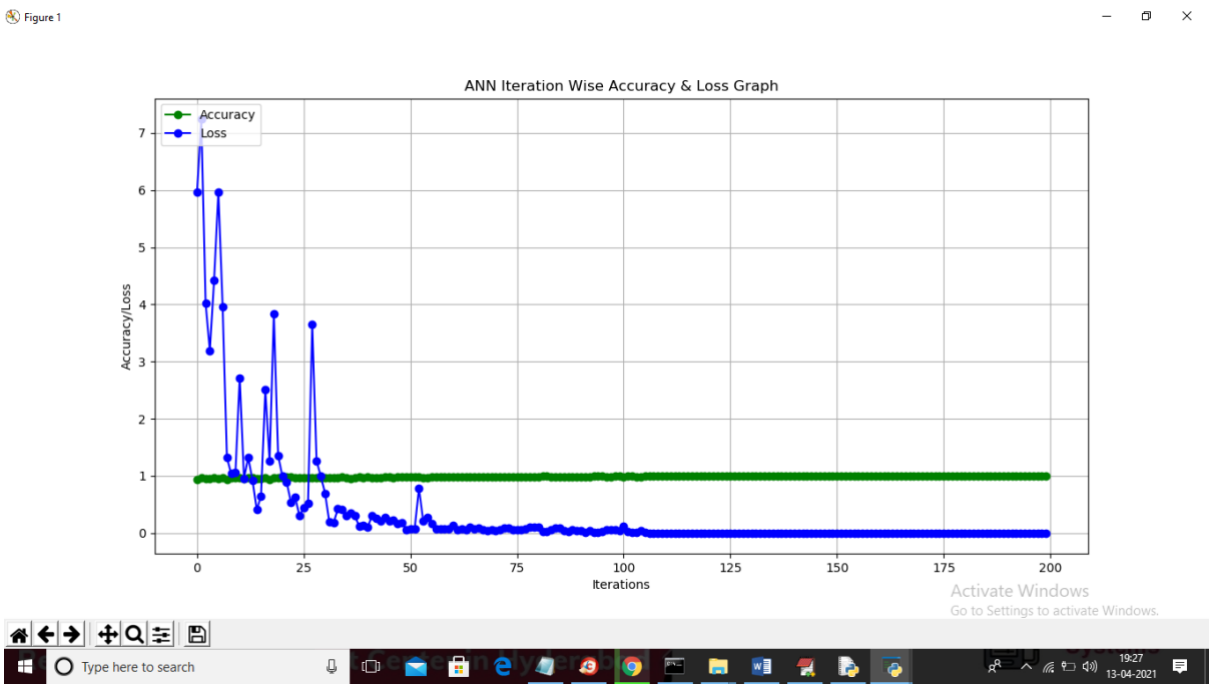
In above screen we can see ANN start iterating model generation and at each increasing epoch we can see accuracy is getting increase and loss getting decrease.



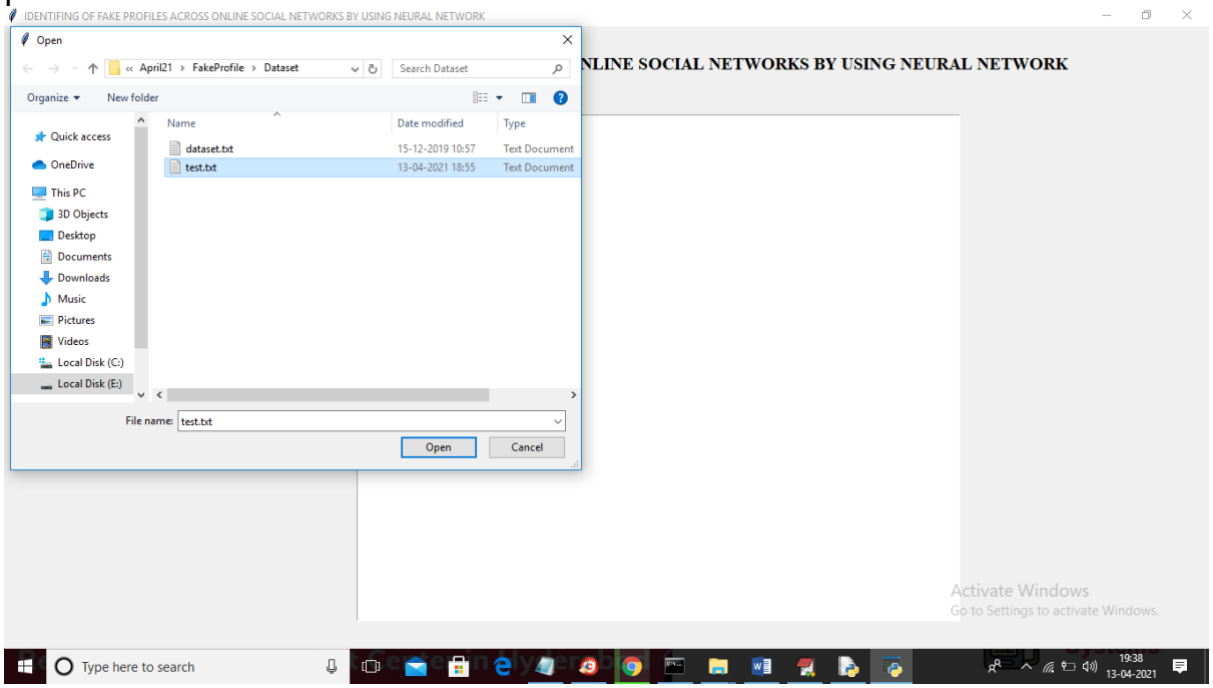
In above screen we can see after 200 epoch ANN got 100% accuracy and in below screen we can see final ANN accuracy



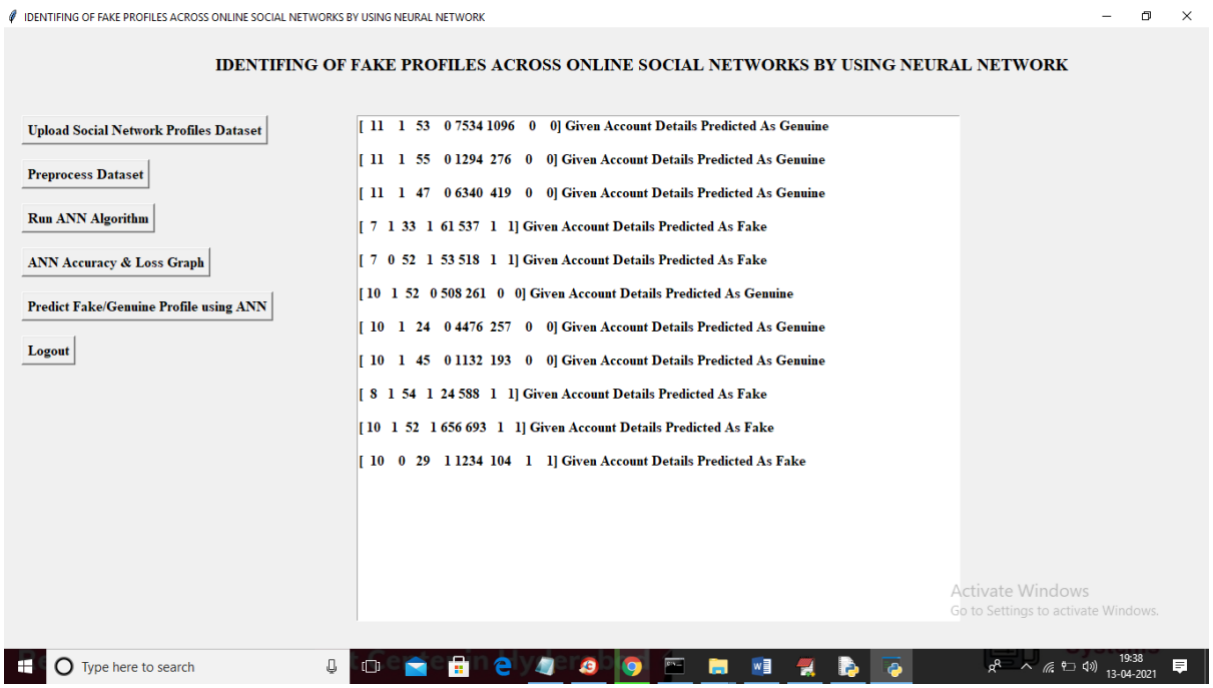
In above screen ANN model generated and now click on 'ANN Accuracy & Loss Graph' button to get below graph



In above graph x-axis represents epoch and y-axis represents accuracy/loss value and in above graph green line represents accuracy and blue line represents loss value and we can see accuracy was increase from 0.90 to 1 and loss value decrease from 7 to 0.1. Now model is ready and now click on 'Predict Fake/Genuine Profile using ANN' button to upload test data and then ANN will predict below result



In above screen we are selecting and uploading 'test.txt' file and then click on 'Open' button to load test data and to get below prediction result



In above screen in square bracket we can see uploaded test data and after square bracket we can see ANN prediction result as genuine or fake

Significance of Research

ANNs are basically known as lone performers, which are not intended in the production of the general network types. This software is used for practical application through its networks. The primary focus is on forecasting and data mining. The software tools are used as -

1. Darknet,
2. NeuroSolutions,
3. Neural Designer,
4. Keras,
5. Neuroph,
6. Tflearn,
7. Torch,
8. “Stuttgart Neural Network Simulator”,
9. ConvNetJS,
10. NVIDIA DIGITS,

The ANN process has the ability in the relearning process according to the newer data types. Due to the uncertainty and complexity, it is difficult in defining a particular analytical model. In the elaborate ritual, a powerful computer-based application can be used. Therefore, the optimization technique's principle lies in the optimization process through which both the constraints and object functions are evaluated into the simulation model (Wanda and Jie, 2020). For the combined simulation, the optimization techniques and ANN need to be provided with the practical means for higher complex optimization. In searching for the solution space, the 'multi-objective optimization algorithm" or NSGA-II is used with adaptive local search. In discrediting the event simulation model, both input-output data is used for the

generation of ANN in approximating the object function. Acting as an intelligent brain, this can train simulated data and accurate models.

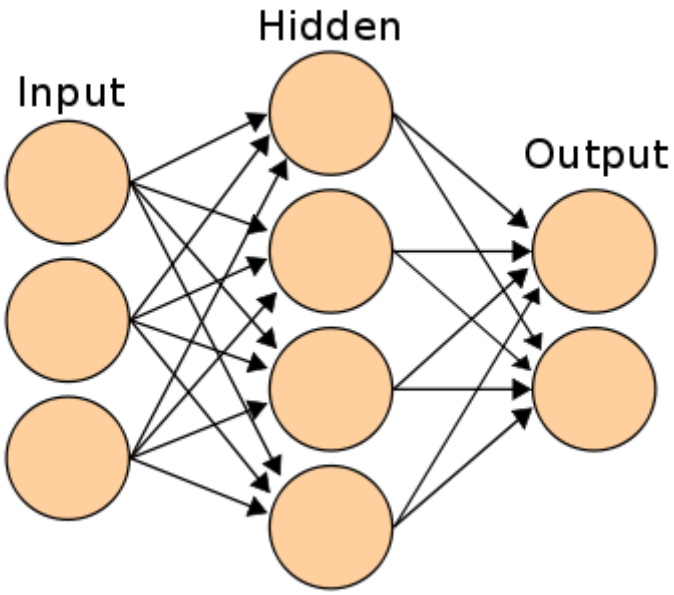


Figure3: ANN Framework

(Source:https://www.analyticsvidhya.com/wp-content/uploads/2014/10/ANN.png)

In between the nodes, the linkages are considered as the main factors (Zhang *et al.* 2020). By finding random weights of the linkages at the start of the algorithm, using the inputs for finding the linkages, searching the errors at the output nodes, weight calibration in between hidden and input nodes, defining the final linkage weights for the scoring of activation rate the framework is structured. Apart from this, by using hidden nodes and their linkages with the output, the output nodes' activation nodes can be found out.

4: Required Resources

In finding the resources, there are multiple modules that can be used. As the social network is a general site, by implementing artificial neural networks, different kinds of modules can be used in the detection of fake profiles. PyBrain is known as a modular within the machine learning library in using Python. Comparing the algorithm with predefined environments can offer better machine learning tasks. Scikit-learn are used for machine learning through Python (Meligy*et al.*, 2017). In predictive data analysis, it is considered as efficient tools. The sexmachine was created for publishing Python 3 compatible versions into PyPi. Without bugging, it can add definite improvements. In relation to this matplotlib is considered a comprehensive library used for animated, static, and interactive visualizations in python. This can make easy and more challenging things more efficiently to create. The ipython notebook is also known as the Jupiter notebook. In the computational environment, it can be combined with the execution of codes, plots and mathematics. Therefore ipython is also known as an interactive shell of python. A Jupiter kernel works with the code in the notebook.

Section 5: Required Skills

The activity of this related technique is from translating web pages into three virtual assistants to order groceries while conversing with chatbots in solving problems. Email servers are also using ANNs and deleting spam from the user inbox. Chatbots are also developed with ANNs as a "natural language processing". Pandas in the package of python are delivering their flexible, fast and flexible data structure in working with the level data types. For working with the array types, NumPy is used in the python library. The working function lies in the domain of "linear algebra". In relation to this pipe is known as the package manager of python. This is used as a distributed part of the standard library (Kaur and Sabharwal, 2018). Apart from this, the knowledge is required about Java, Python more clearly. In using the modules and packages, depth knowledge and preferred system configuration are preferable. Python 3.9 is used as the best version. Therefore, the ram, hard disk capacity, and the IDE packages are necessary for working with the programs.

6: Project Plan

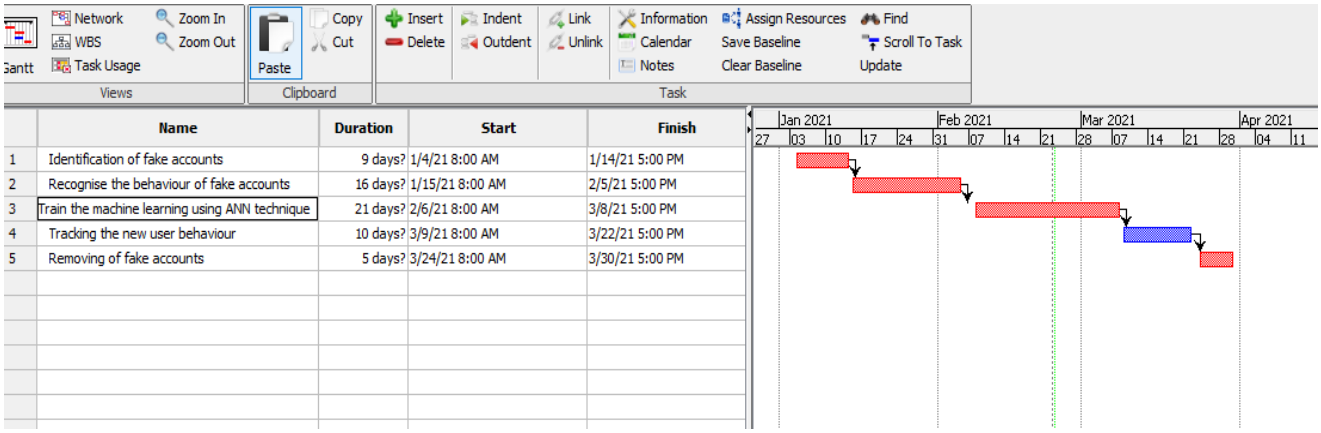


Figure4: Gantt chart

Conclusion

The Network traffic classification techniques are discussed in this paper to enhance some idea about Machine Learning algorithms for network traffic data. The analysis carried out definitely helps to a new analyst to make the decision about which Machine Learning algorithm is more appropriate for this application. Initially, the network traffic extraction is carried out to evaluate the different Machine Learning algorithm which is trained in later phase. The Machine Learning algorithms are used for managing the performance of network and classification of unknown applications.We then employ four basic Machine Learning algorithms to analyze the protocol. Further, the classifiers using different Machine Learning algorithms are developed to compare the accuracy for this network traffic data. We find that K-nearest neighbor (KNN) algorithm outperforms Naïve Bayes algorithm, Decision Tree and Support Vector Techniques in terms of accuracy which is due to the fact that KNN uses better classification criterion than Naïve Bayes and Decision Tree Algorithm. We find that KNN is most robust among the algorithms: NB, DT, and SVM for out training data set. It is also able

to maintain highest mean for accuracy. Feature extraction for classifying students based on their academic performance Dept. of CSE Malla Reddy Engineering College for Women(Autonomous Institution-UGC, Govt. of India) Page 45 6. REFERENCES [1] Chakraborty, A., J.S. Banerjee, and A. Chattopadhyay. Non-uniform quantized data fusion rule alleviating control channel overhead for cooperative spectrum sensing in cognitive radio networks. in 2017 IEEE 7th International Advance Computing Conference (IACC). 2017. IEEE. [2] Chakraborty, A., J.S. Banerjee, and A. Chattopadhyay, Non-uniform quantized data fusion rule for data rate saving and reducing control channel overhead for cooperative spectrum sensing in cognitive radio networks. Wireless Personal Communications, 2019. 104(2): p. 837-851. [3] Rueda, A. A survey of traffic characterization techniques in telecommunication networks. in Proceedings of 1996 Canadian Conference on Electrical and Computer Engineering. 1996. IEEE. [4] Shahbar, K. and A.N. Zincir-Heywood. How far can we push flow analysis to identify encrypted anonymity network traffic? in NOMS 2018-2018 IEEE/IFIP Network Operations and Management.

REFERENCE

1. Awasthi, S., Shanmugam, R., Jena, S.R. and Srivastava, A., 2020. Review of Techniques to Prevent Fake Accounts on Social Media.
2. Hajdu, G., Minoso, Y., Lopez, R., Acosta, M. and Elleithy, A., 2019, May. Use of Artificial Neural Networks to Identify Fake Profiles. In *2019 IEEE Long Island Systems, Applications and Technology Conference (LISAT)* (pp. 1-4). IEEE.
3. Kaur, J. and Sabharwal, M., 2018. Spam detection in online social networks using feed forward neural network. In *RSRI conference on recent trends in science and engineering* (Vol. 2, pp. 69-78).
4. Khaled, S., El-Tazi, N. and Mokhtar, H.M., 2018, December. Detecting fake accounts on social media. In *2018 IEEE International Conference on Big Data (Big Data)* (pp. 3672-3681). IEEE.
5. Meligy, A.M., Ibrahim, H.M. and Torky, M.F., 2017. Identity verification mechanism for detecting fake profiles in online social networks. *Int. J. Comput. Netw. Inf. Secur.(IJCNIS)*, *9*(1), pp.31-39.
6. Ramalingam, D. and Chinnaiah, V., 2018. Fake profile detection techniques in large-scale online social networks: A comprehensive review. *Computers & Electrical Engineering*, *65*, pp.165-177.
7. Wanda, P. and Jie, H.J., 2020. DeepProfile: Finding fake profile in online social network using dynamic CNN. *Journal of Information Security and Applications*, *52*, p.102465.
8. Zhang, J., Dong, B. and Philip, S.Y., 2020, April. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)* (pp. 1826-1829). IEEE