Turkish Journal of Computer and Mathematics Education *Vol.11 No.01 (2020), 991 - 997*
Research Article

Methods for Detecting Fake Profiles in Online Social Networks Using Artificial Neural Networks

Dr. Koparthi Suresh[1] , Dr. RVVSV Prasad[2]
[1]Professor, Bhimavaram Institute of Engineering & Technology, Pennada, Bhimavaram, India
[2]Professor, Swarnandhra College of Engineering & Technology, Narsapur, - 534275, India
sureshkgrl@gmail.com[1], ramayanam.prasad@gmail.com[2]

ABSTRACT

In light of the current situation, most individuals are participating in online social networks. Everyone, from kids to adults, spends a lot of time on these sites, either learning new things or getting in touch with friends and family in more effective ways. However, today's social media platforms are plagued by a large number of bogus accounts that exploit security flaws in order to steal from the sites or commit cybercrimes themselves.

Keywords: Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Naive Bayes, and Decision Tree (SVM)

INTRODUCTION

In the context of computer security, the term "malware" is used to describe any malicious programme. One of the biggest security concerns with the rapid development of technology is malicious software [1]. Malware, or malicious software, refers to any software created with the intention of causing damage to a computer system by means such as stealing data or spying on its users. Depending on their intended damage or behaviour, malicious programmes like Trojans, viruses, rootkits, worms, and spyware fall under one of many categories [13]. With a growing family of malicious software, anti-virus scanners are unable to keep up, since millions of pieces of software are still vulnerable. Kaspersky Lab estimates that there were 5,638,828 unique hosts compromised in 2018 1. More than 33 billion records will be taken by thieves in 2023, according to another study by Juniper Research 2. These days, it's not easy to spot malware since hackers may use any number of techniques they find online.

Moreover, the widespread availability of anti- detection techniques allows anybody to launch an attack or create malicious software without needing any special training or expertise. Furthermore, attackers are using techniques that need them to swiftly update to a newer version in order to evade detection. In light of the fact that even a single malware attack may cause significant harm to data and catastrophic losses, protecting computer systems is one of the most essential jobs for consumers and enterprises. Due to massive amounts of money being lost and the prevalence of assaults, effective and trustworthy methods of detection are essential. There are two main types of malware detection: static analysis, which involves analysing a compiled file or programme, and dynamic analysis, which involves analysing the behaviour of the programme while it is running, including things like how much power it uses, how much memory it accesses, and how much of the device's network it uses [2].

Static analysis is concerned with analysing malicious code without running the programme; it employs techniques including file format inspection, text extraction, anti-virus scanning, fingerprinting, and

disassembly to uncover the file's behavioural features. In order to do a dynamic analysis, the file being analysed must be monitored in real time while it is being performed in a virtual machine. Malware detection methods may be broken down further into signature-based and heuristics-based approaches [6]. However, this approach's precision isn't always sufficient for detection, leading to a great deal of erroneous positive and negative results. There is an immediate need for a different kind of detecting system. Therefore, machine learning-based methods may greatly improve system security. Malware execution may be classified as either infected or not using a machine learning model. In order to determine whether or not a suspicious file is harmful, malware analysis needs highly developed detection capabilities. It's also put to use while trying to determine what kind of virus an object is. It is possible to apply machine learning techniques to figure out what constitutes a typical behaviour and then look for anything that seems out of the ordinary. Because of this, it can safeguard users' computers and halt threats considerably more quickly than in the past.

Random Forest (RF), Support Vector Machine (SVM), Na vc Bayes (NB), Logistic Regression (LR), and AdaBoost [3] are just few of the many machine learning approaches shown to be effective in malware identification and classification. Our research adheres to the procedures outlined in the cited literature [4]. This paper's experiments are grounded on 306 attributes gleaned through a sandbox's observations of the files' real behaviour (Heuristic). On a set of 984 malicious and 172 benign files, they used machine learning methods to conduct binary and multi-class categorization. Random Forest obtained the maximum accuracy, 95.69%, for multi-class classification and 96.8%, for binary classification. Through the use of a Random Forest classifier, we are able to focus on the most relevant traits while excluding those that are less crucial to our task. As a result, our results are more precise than those found in [4], the study we based our work on. Based on our tests, Decision Tree has the greatest accuracy (98.2%) for binary classification, whereas Random Forest has the highest accuracy (95.8%) for multi-classification. The remaining sections of this work will be structured as follows. In Section 2, we discuss our further findings. Our approach and data sets are described in Section 3, and the findings are presented in Section 4. Our findings and suggestions for further research are presented in Section 5.

Issues in Research

The prevention of phishing assaults, in which personal information is stolen over the internet, is a major problem in the context of phoney profiles. It's a common tool for data theft by cybercriminals. These profiles are engaged in everything illegal, including the detection of passwords, the dissemination of irrelevant material, and the promotion of awareness. This may lead to obscurity in the long run if you are able to master the crisis and use it to your benefit. In order to reduce instances of cybercrime like trolling, hacking, and cyberbullying, this must be determined.

Explanation for Conducting Research

All social media accounts need to be protected against cybercriminals, thus it's important to be able to detect them, and a model based on artificial neural networks may help with this.

Strategy for Research

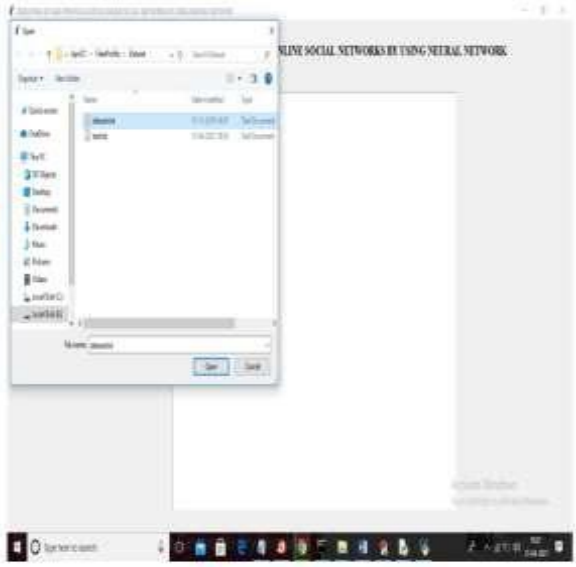
For this purpose, a new "artificial neural network" component has been added to the computer infrastructure. It's designed to mimic the way the human brain stores and processes data. For studies of this kind, the inductive method might be explored. Taking a look at the current processes and circumstances, we can see this via the regularities and patterns in the system. Using an ANN model properly is essential for gaining a technological edge. It's like the theoretical underpinnings of AI that finally allow us to prove the impossible by human standards. For this reason, "artificial neural networks" (ANNs) are presented as a method of modelling, enabling the human nervous system by

means of a learning approach. This kind of detection reveals information about the target through "actions taken by the user. An important part of the reporting of anomalies is the input of users. Users' social impact may be measured using these two categories. There are two goals here: the first is to identify the user's influence on others, and the second is to elevate the user's status. The assessment also takes into account the "feature that can be seen with the naked eye

SCREENSHOTS

Double-click the 'run.bat' file to launch the project.

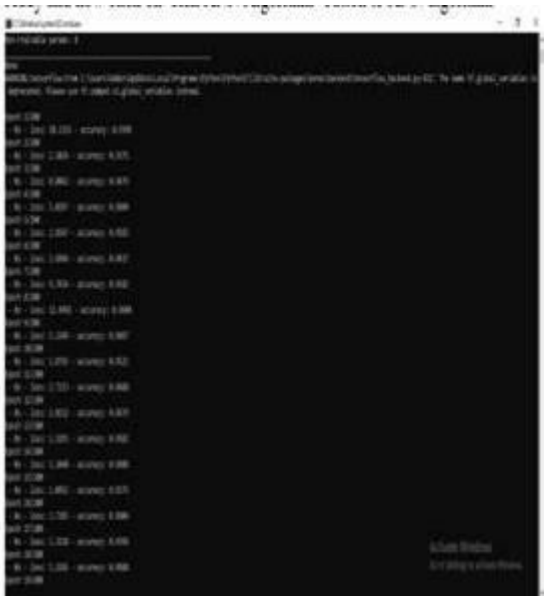
For the aforementioned screen, choose the "Upload Social Network Profiles Dataset" option.



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 „Pre-process Dataset" button to remove missing values and to split dataset into train and test part

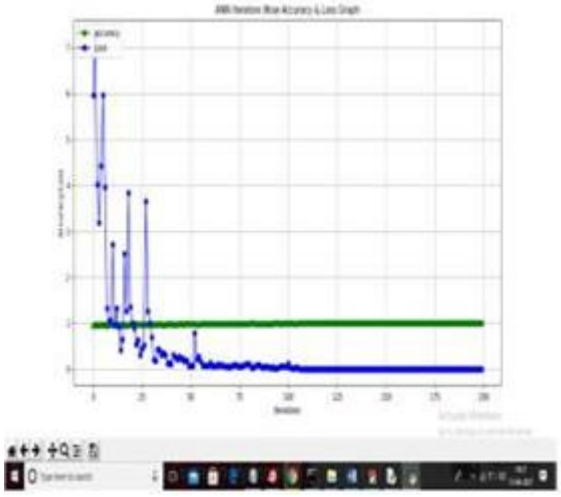
The following page shows that there are a total of 600 records in the dataset, 480 of which will be used for training the ANN and 120 for testing it. Once the dataset has been prepared, the user may click the "Run ANN Algorithm" button to begin the training and testing processes.



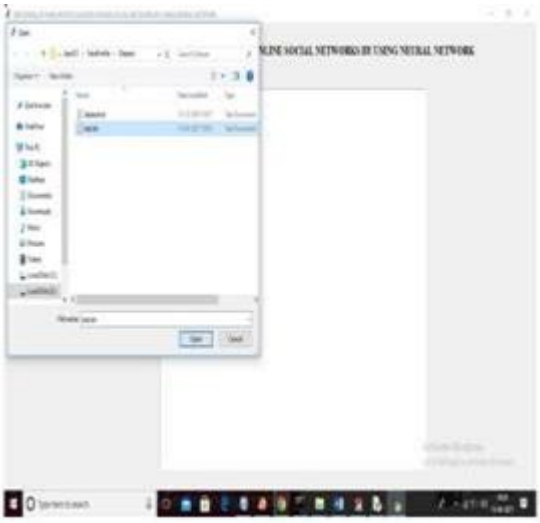
On the above screen, we can see that the ANN has begun iteratively generating a model, and that with each new epoch, the model's accuracy and loss are improving.

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



We can observe that the accuracy went from 0.90 to 1 and the loss value went from 7 to 0.1 by looking at the preceding graph, where the x-axis denotes epoch and the y-axis represents accuracy/loss value. Now that the model is complete, you may submit your test data by clicking the "Predict Fake/Genuine Profile using ANN" button, and the ANN will make a prediction, as shown below.



In the preceding window, we load test data by choosing the test.txt file, clicking the Open button, and then examining the prediction results.

Above, the uploaded test data can be seen in square brackets, and the ANN prediction result, authentic or fraudulent, can be seen immediately after.

The Value of Research

General-purpose networks are not designed to have ANNs integrated into their architecture, hence ANNs are often seen as outliers. Through its many interconnected systems, this software is put to actual use. Predictive modelling and data mining are the main areas of interest.

The applications serve as –

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

The ANN method may adapt its learning to different data types via a relearning procedure. The unpredictability and complexity make it hard to define a specific analytical model. A potent computer-based programme may be employed in the intricate ritual. Thus, the essence of the optimization method is in the optimization procedure, which incorporates the evaluation of constraints and object functions into the simulation model (Wanda and Jie, 2020). The optimization methods and ANN used in the integrated simulation need to be given realistic tools for achieving more complicated optimization. Adaptive local search is employed with the "multi- objective optimization technique,"

also known as NSGA-II, to discover the solution space. Both input and output data are employed in the construction of ANN to approximate the object function, hence undermining the event simulation model. This serves as a computational hub, allowing for the training of synthetic data and reliable models.

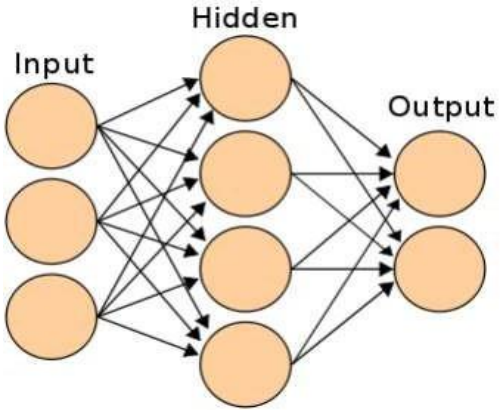


Figure3: ANN Framework

The connections among the nodes are the most important aspects to take into account (Zhang et al. 2020). The framework is organised by first generating random linkage weights, utilising the inputs to generate the linkages, searching the errors at the output nodes, calibrating the weights between the hidden and input nodes, and finally defining the linkage weights for the scoring of activation rate. In addition, the activation nodes of the output nodes may be identified by using hidden nodes and the connections between them and the output.

The Necessary Means

There are a number of modules available for use in the resource discovery process. The social network being generic in nature allows for the use of several modules for the identification of fraudulent profiles, all of which may be implemented using artificial neural networks. Among Python's machine learning libraries, PyBrain is well recognised as a versatile, modular option. Improved machine learning tasks may be obtained by comparing the algorithm to specified contexts. Python's scikit-learn packages may be used for machine learning (Meligyet al., 2017). As such, they are regarded as effective instruments in the field of predictive data analysis. The sexmachine was developed to facilitate the distribution of PyPi packages that are compatible with Python 3. Fixing bugs isn't necessary for it to make noticeable improvements. Matplotlib is a Python package used for creating dynamic, static, and interactive visualisations. This may streamline the production of both simple and complex tasks. Another name for the ipython notebook is the Jupiter notebook. In a computer lab setting, it may be used in tandem with the running of algorithms, graphs, and calculations. Thus, ipython is often referred to as a Python interactive shell. The notebook's code is compatible with a Jupiter kernel.

Appropriate Abilities

This linked method encompasses a wide range of actions, such as web page translation into three different types of virtual helpers, grocery store ordering with the help of chatbots, and the resolution of difficulties. ANNs are also being used by email servers to filter out unwanted messages before they reach the inbox of a user. Artificial neural networks (ANNs) are used in the creation of "natural language processing" chatbots. Python's pandas library provides a quick and versatile data structure for dealing with tabular data. NumPy is a module of the Python library used for manipulating arrays of various kinds. The

operative function is properly classified as "linear algebra" related. Python's package management is often mentioned in connection with this conduit. This is a component of the standard library and is supplied with it (Kaur and Sabharwal, 2018). Aside from this, a solid grounding in Java and Python is also necessary. Using the modules and packages effectively requires in-depth familiarity of the system and the user's chosen configuration. As of late, the best version of Python has been 3.9. Accordingly, the availability of RAM, sufficient hard drive space, and IDE packages are essential for using the software.

Project Plan

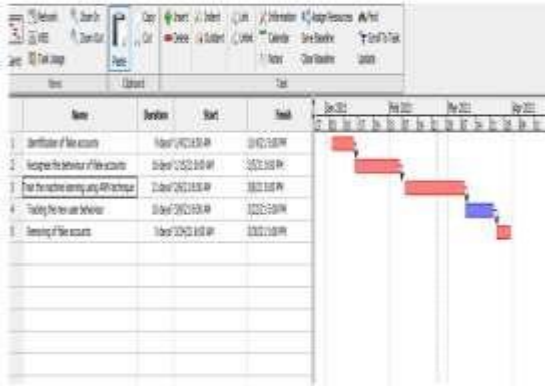


Figure4: Gantt chart

Conclusion

In this article, we describe the methods for classifying network traffic in order to have a better understanding of the Machine Learning algorithms that may be used to such data. A novice analyst would benefit greatly from the analysis performed here when deciding which Machine Learning algorithm to use. At first, we take data from the network in order to test the learned Machine Learning algorithms. Algorithms based on Machine Learning are used to manage network performance and categorise unidentified apps. Finally, we conduct our analysis of the protocol using four foundational Machine Learning methods. Also, classifiers are constructed for this network traffic data using several Machine Learning techniques, and their accuracy is compared. Due to its superior classification criteria, the K-nearest neighbour (KNN) approach surpasses the Naive Bayes algorithm, the Decision Tree technique, and the Support Vector Machine technique. Among the three algorithms we use for training (KNN, DT, and SVM), we find that KNN is the most stable. It may also keep the greatest mean for accuracy. Student classification using extracted features for academic achievement Computer Science and Engineering Department Malla Reddy Engineering College for Women (Autonomous Institution- University Grants Commission, Government of India) Page 45

References

[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).