IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

DETECTION OF MALWARE USING BIDIRECTIONAL LSTM

K.AVINASH,

M.Tech, Dept of IT&CA,Andhra University College of Engineering,Visakhapatnam

ABSTRACT

With the rapid development of the Internet, the methods of cyber attack have become more complex and the damage to the world has become increasingly greater. Therefore, timely detection of malicious behavior on the Internet has become an important security issue today. This paper proposes an intrusion detection system based on deep learning, applies bidirectional long short term memory architecture to the system, and uses the MALWARE(numerical) data set for training and testing. Experimental tests show that the intrusion detection system can effectively detect the known or unknown malicious behavior of the network under the current network environment. Malware, short for “malicious software,” refers to any intrusive software developed by cybercriminals (often called “hackers”) to steal data and damage or destroy computers and computer systems. Examples of common malware include viruses, worms, Trojan viruses, spyware, adware, and ransom ware. Recent malware attacks have exfiltrated data in mass amounts. A Bidirectional LSTM, or bi-LSTM, is a sequence processing model that consists of two LSTMs one taking the input in a forward direction, and the other in a backwards directionIn bidirectional, our input flows in two directions, making a bi-LSTM different from the regular LSTM. With the regular LSTM, we can make input flow in one direction, either backwards or forward. However, in bi-directional, we can make the input flow in both directions.

Keywords:-Malware, LSTM, BI-LSTM, Cyber Criminal, Rransomware

1. INTRODUCTION

The term "malware" stands for "malicious software," which includes worms, Trojan horses, and other destructive programs. These programs have a wide range of capabilities, including the capacity to steal, encrypt, or destroy private information, change or hijack common computer activities, and keep an eye on online behavior. Show user approval. One category of malware includes computer viruses. It is frequently a program that is installed without the user's consent and has the potential to damage both the operating system and the physical (hardware) parts of a computer. The virus's effects include file deletion and size modifications. Erase the disc's whole, including all formatting. It is challenging to get data from the drive because the file allocation table has been destroyed. a variety of sound and graphic effects that are safe but frightening; a slowing down of the computer's operation until it crashes; and worms. Computer worms are malicious software applications that spread by using computer-to-computer communication. Worms and viruses have characteristics in that they both have the ability to reproduce, although worms do it on other computers rather than locally. I propagated to other systems using computer networks. Internet worms, instant messaging worms, and email worms are some examples of computer worms. Sharing files via a network.

Trojan horses

Trojan horses are "masked" programs that try to access the operating system using security flaws in the operating system. Unlike computer viruses, which can replicate themselves, Trojans cannot.

The following categories apply to different types of Trojan horses:

Backdoors: give the attacker remote Internet access to the victim's machine; Applications that read data from the keyboard and save it in files that can subsequently be viewed by the attacker or delivered directly to the email account are known as password stealers. logical bombs: these Trojans are capable of taking activities that endanger system security when certain conditions are met;

Denial of Service tools are programmes that transmit specific data sequences to the target audience, which is typically a website, with the aim of stopping that audience's Internet services.

•**The ransomware**

Ransomware is a type of malware that locks up a victim's computer and demands payment to unlock it. Based on the type of virus, the award is given and the victim's need to pay is formally justified. Some ransomware variations claim that the payment is necessary to unlock encrypted data, while others claim that doing so is the only way to avoid punishment from a government agency (typically the FBI or a local agency). Ransomware's capacity to encrypt private user data is one of its effects.

has the ability to delete specific files, multimedia, and other files that contain important data. They might also try to get rid of important system or other application components.

• Threats from ransomware can be used to steal identity data, irreplaceable personal documents, authentication names, passwords, and other confidential information. Additionally, they can abruptly halt the operation of anti-virus, anti-spyware, or other software by throttling its activities and deactivating essential system functions. Seventh Root Kit A root kit software program often exploits a vulnerability in the host system to get complete access to a system, alter it, and then secretly use its resources.can modify the "pHs" program on a Linux system such that it hides the root kit process while showing active processes.

2 . LITERATURE SURVEY

Author: Tort Trial & Insurance Practice Law Journal Roberta D.

Every company is exposed to cyber danger. The news reports confirm that cyberattacks are occurring more frequently, are more well developed, and are larger than before. They also cross boundaries of both industries and places. As serious cyberthreats grab daily headlines, regulations pertaining to data privacy and security are growing. Due to the rise in data security breaches, denial of service assaults, other attacks, and data loss, addressing and decreasing cyber risk is a top priority for businesses worldwide. There is no firewall or impenetrable security system, thus it is abundantly clear that network security cannot fully address the issue of cyber risk on its own. The whole approach taken by a business to manage, reduce, and maximize its cyber risk protection can . A company's overall strategy to handle, mitigate, and optimise protection against cyber risk can greatly benefit from the use of insurance. The Securities and Exchange Commission is aware of this fact. The SEC's Division of Corporate Finance has released recommendations on cyber security disclosures under the federal securities laws in response to "increasing frequent and serious cyber events." According to the guidance, businesses "should examine the appropriateness of their disclosure relating to cyber security risks and cyber events on a continuous basis" and that "acceptable disclosures may include" a "[d] description of applicable insurance coverage," among other things. An enhanced Android malware detection method based on permission-based features and an evolving hybrid neuro-fuzzy classifier .Several kinds of attackers have been drawn in by the rising popularity of Android devices and consumers. By using code obfuscation techniques, malware developers produce new versions of malware from older ones. The exponential rise in the production of malware varieties might have been made easier by malware that was obscured. Because signature-based malware detectors may easily dodge it and behavior-based malware detectors struggle to accurately identify it, obfuscated malware is difficult to detect. Therefore, a reliable technique for identifying disguised malware in Android-powered smart phones is necessary. In the literature on Android malware classification, there aren't many malware detection techniques that can locate disguised malware. These malware detection techniques, however, lacked the capacity to improve their performance through rule-based learning and evolution. Based on the concept of creating soft computing systems, this research proposes an evolving hybrid neuro-fuzzy classifier (EHNFC) for Android malware classification utilizing permission-based features. The proposed EHNFC can not only identify malware that has been hidden by fuzzy rules, but it also also of clustered permission-based features was added to an evolving clustering method for adapting and evolving malware detection fuzzy rules. By increasing cluster convergence and producing rules that are better suited to the input data, this change to the developing clustering algorithm raises the proposed EHNFC's classification accuracy. The suggested EHNFC outperforms a number of cutting-edge obfuscated malware classification algorithms in terms of false negative rate (0.05) and false positive rate, according to testing results (0.05). The findings also show that the proposal, in terms of accuracy, identifies Android malware more effectively than existing neuro-fuzzy systems (namely, the adaptive neuro-fuzzy inference system and the dynamic evolving neuro-fuzzy system).Detecting malware variants.Malware programmes, such as Trojan horses, worms, and viruses, are rampant everywhere. Research and data indicate that malware's effects are deteriorating. The main weapons in the fight against malware are malware detectors. For the purpose of identifying malicious software within a computer system, the majority of commercial anti-malware scanners keep a database of malware patterns and heuristic signatures. To create new stealth versions of their malicious programmes, malware authors use semantic-preserving code modification (obfuscation) techniques. Today's detection methods struggle to identify malware variants because they primarily focus on syntactic characteristics while ignoring the semantics of malicious executable programmes.To combat this new security danger, a strong malware detection technique is needed. In this thesis, we present a new methodology that, by examining the semantics of known harmful code, overcomes the limitation of current malware detection methods. The creation of a semantic signature, slicing analysis, and test data generation analysis are the three main analysis methodologies that make up the methodology. This method's main component is to approximate the semantics of malware code and to create signatures that may be used to recognise malware variants that may be disguised but are nonetheless semantically comparable. A programme test input plus semantic traces of known malware code make up a semantic signature. Finding a balance has been the main obstacle in creating our semantics-based method to malware variant detection.Finding a balance between increasing the detection rate (i.e. matching semantic traces) and performance, with or without accounting for the effects of obfuscation on malware variants, has proven to be the main issue in creating our semantics-based method to malware variant detection. In order to improve the creation of semantic signatures, we develop slicing analysis. We support our trace-slicing method with a theoretical finding that demonstrates the slicer's accuracy. Our malware detector's proof-of-concept implementation shows how the semantics-based analysis method could enhance existing detection technologies and make it harder for malware authors to create new viruses. Exploring programme semantics for the selection of an appropriate component of the semantic signature is another crucial aspect of this thesis, and for this, we present two new theoretical findings.Classifying unknown packing techniques for malware detection using entropy analysis Authors: Mahn Soo Choi, Hongzhe Li, Heejo Lee, and Munkhbayar Bat-Erdene.Around 80% of all currently active malware is packed, a number that has been steadily increasing. In this study, we suggest a system for categorising the regardless of whether they are malicious software or safe programmes, packing algorithms of provided unknown packaged executables. The entropy values of a certain executable are first scaled, and the entropy values of a specific region in memory are then represented symbolically. Symbolic aggregate approximation (SAX), which is known to be efficient for massive data conversions, is the foundation of our suggested approach. Second, we use supervised learning classification techniques to categorise the distribution of symbols,Our trials with a set of 324 packed benign programmes and 326 packed malicious programmes, each with 19 packing algorithms, show that our method can identify the packing algorithms of provided executables with a high accuracy of 95.35%, a recall of 95.83%, and a precision of 94.13%. On the basis of incremental aggregate analysis and SAX representations of the entropy values, we suggest four similarity metrics for identifying packing techniques. The fidelity similarity measurement, which is from 2 to 13 greater than the other three metrics, shows the best matching result among these four metrics, with an accuracy rate ranging from 95.0 to 99.9%.A framework for intrusion detection for mobile phones based on specifications Authors: Sencun Zhu, Zhi Xu, and Ashwin Chaugul.Malware is increasingly becoming more prevalent on mobile devices as a result of the mobile market's rapid rise. One characteristic of a lot of malware that is frequently discovered on mobile devices is that it always tries to access private system functions on the device in a covert and sneaky method. For instance, the malware might secretly communicate with the device's audio peripherals or send messages automatically without the user's knowledge or consent. We introduce SBIDF, a Specification Based Intrusion Detection Framework, which uses keypad or touch screen interrupts to distinguish between malware and human action, to detect the illegal malicious activities.In the proposed framework, we use an application independent specification to describe the typical behaviour pattern. This specification is written in Temporal Logic of Causal Knowledge (TLCK), and it is enforced to all third-party applications on the mobile phone during runtime by watching the inter-component communication pattern among crucial components. Our analysis of the behaviour of simulated real-world malware demonstrates our capability to identify all types of malware.

This paper presents a malware detection method based on network behavior evidence chains. The proposed new method will detect the specific network behavior characteristics on three different stages as connection establishment, operating control, and connection maintenance.

3.IMPLEMENTATION STUDY

Methods for Malware Analysis:

The creation of efficient methods for identifying infected files requires malware analysis. This analysis entails looking at the goals and operations of a malware programme. In order to describe how malware functions and what impact it has on the system, three separate analysis methodologies are used, although their time and skill requirements are significantly different.

Static evaluation

Another name for it is code analysis. In other words, malicious software code is examined to learn more about how it operates. Tools for disassembly, decompilation, debugging, and source code analysis are used in this reverse engineering technique. As this technique has no execution time overhead, we shall use it exactly as is.

Dynamic evaluation

Also known as behavioural analysis. During execution in a secluded setting like a virtual machine, simulator, or emulator, infected files are examined. Following file execution, the system's behaviour and impacts are observed.

Hybrid research

This method is suggested as a way to get beyond static and dynamic analysis' constraints. In order to improve the comprehensive analysis of malware, it first analyses the specification of the signature for every malicious code and then combines it with the other behavioural parameters. Owing to this method, hybrid scanning is more advanced than static and dynamic scanning.

Signaturedetection

Using signature-based detection, a threat's individual identifier is established and made public so that the threat can be discovered. to be determined later. In the context of a virus scan, this may be a special code template that is attached to a file or something as straightforward as the hash of a malicious file. Known. The file may be flagged as malicious if that particular pattern or signature is found again. As malware has evolved, its creators have started to use novel methods like polymorphism to alter the pattern each time an object spreads from one system to another.

Detection of behaviour

In contrast to signature-based scanning, which reveals The heuristic scan [5] employs rules and/or algorithms to seek for commands that may suggest intent or evil if the signatures detected in the file match those of a known malware database. Certain heuristic scanning techniques can identify malware using this technique without the need for a signature. Because of this, the majority of antivirus products use signature and heuristic methods to detect any malware that may attempt to avoid detection.

Determine features

A variation on behavior-based detection known as feature detection aims to reduce the frequency of false alarms that are typically **associ**ated with it. Program characteristics that characterise the security behaviour of crucial programmes serve as the foundation for characteristic detection. Instead of ostensibly recognising specific attack patterns, this entails watching programme executions and spotting deviations from the specification and its behaviour. This method is similar to that for detecting anomalies, but it differs in that it is based on characteristics created by hand to record the system's behaviour rather than using machine learning methods**.**

Machine learning

Machine learning is a class of methods that enables software applications to predict outcomes considerably more accurately without being specifically designed. Building algorithms that take input data and apply statistical analysis to forecast output data while output data is updated as much input data become valid is the fundamental tenet of machine learning. Machine learning procedures are comparable to data mining and predictive modelling procedures. For both, it is necessary to look for specific patterns by date and modify software operations accordingly. A lot of people are also aware with machine learning because of online purchasing and the customised marketing they see.

The two types of machine learning algorithms are supervised and unsupervised, respectively.

Supervised algorithms

They need a data researcher or data analyst with machine learning expertise to provide the required input and output data as well as provide feedback on the precision of the predictions, which is especially important during algorithm training. Data scientists choose which traits or factors the model should take into account while making predictions. The algorithm will apply what it has learned to new data after the training is over. Regression and classification issues are within the category of supervised learning challenges. When the output variable is a category, such as "red" or "blue" or "illness" and "no disease," a classification difficulty exists.Regression: When the output variable has a real value, such "dollars" or "weight," a regression problem exists. Recommendation and time series prediction are two typical sorts of challenges built on top of classification and regression, respectively. Popular supervised machine learning algorithms include the following: Regression issues using linear regression. Random forest for regression and classification.

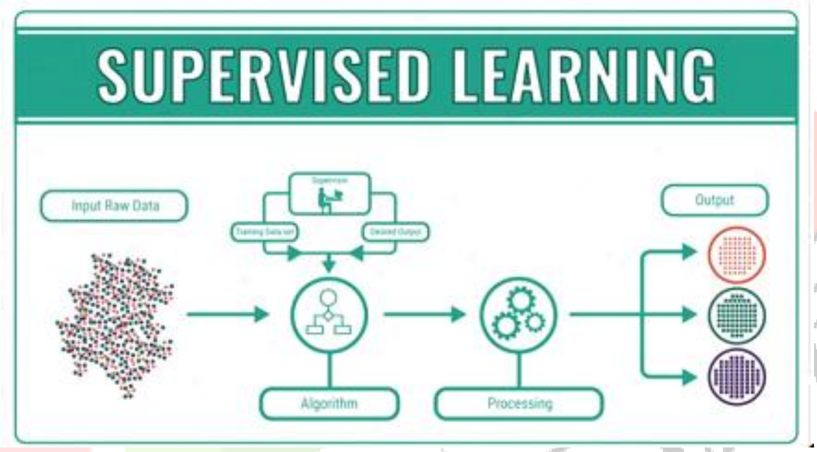

Fig 1:_ algorithms without supervision

4.**PROPOSEDWORK AND ALGORITHM**

We propose a flexible framework in which one may utilize different machine learning techniques to effectively discriminate between malicious files and clean files while striving to limit the number of false positives.

After being successfully tested on medium-sized datasets of malware and clean files, the concepts underpinning this framework were subjected to a scaling-up technique that allows us to work with very large datasets of malware and clean files.

The capacity to recognize hazardous files before they are executed, simplicity of use, speedy identification, and detection of polymorphic malware are all advantages.

- Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step.
- LSTM was designed by Hochreiter & Schmidhuber.
- It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information.
- As the gap length increases RNN does not give an efficient performance.
- LSTM can by default retain the information for a long period of time.
- It is used for processing, predicting, and classifying on the basis of time-series data.

LSTM has a chain structure that contains four neural networks and different memory blocks called cells.

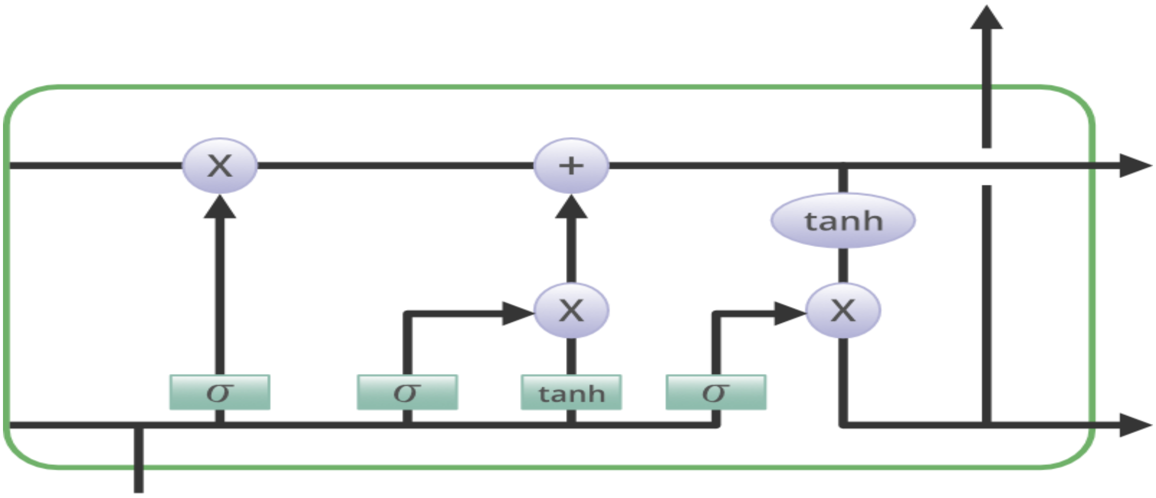

Fig 2:- proposed model of LSTM Structure

5. METHODOLOGIES

5.1 DATA PREPROCESSING

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So for this, we use data preprocessing task.

5.2 DATA EXPLORATION

Data exploration refers to the initial step in data analysis in which data analysts use data visualization and statistical techniques to describe dataset characterizations, such as size, quantity, and accuracy, in order to better understand the nature of the data.

5.3 MODEL CREATION

A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data**.**

5.4 TRAINING AND TESTING

Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy**. You'll need a new dataset to validate the model because it already “knows” the training data.**

5.5 PREDICTION

Training data is the initial dataset you use to teach a machine learning application to recognize patterns or perform to your criteria, while testing or validation data is used to evaluate your model's accuracy. You'll need a new dataset to validate the model because it already “knows” the training data

5.6 BILSTM Model

The architecture of bidirectional LSTM comprises of two unidirectional LSTMs which process the sequence in both forward and backward directions. This architecture can be interpreted as having two separate LSTM networks, one gets the sequence of tokens as it is while the other gets in the reverse order. Both of these LSTM network returns a probability vector as output and the final output is the combination of both of these probabilities.

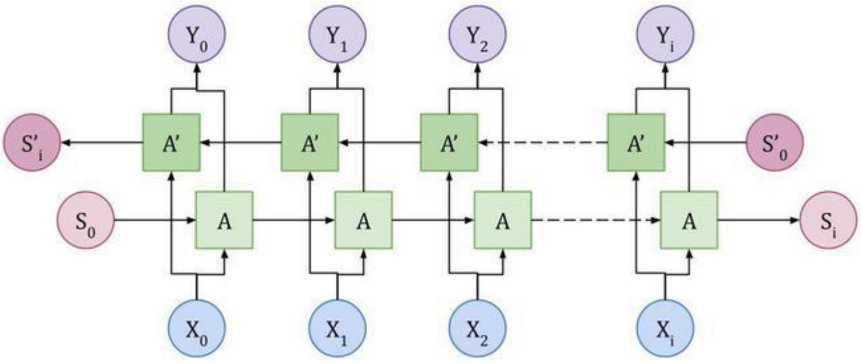

Fig 3 :- proposed Bidirectional LSTM layer Architecture

7.CONCLUSIONANDFUTUREWORK

The aim of this paper is to present a deep learning approach to the malware problem. Due to the sudden growth of malware, we need automatic methods to detect infested files. In the first phase of the work,data set is created using infested and clean executables, in order to extract the data necessary for the creation of the data set, we used a script created in Python. After creating the data set, it must be ready to train machine learning algorithms. This project proposes a bidirectional LSTM algorithm based on convolutional neural network and independent recurrent neural network. Experimental results show that compared with other methods, the proposed BI-LSTM algorithm has improved the accuracy of malicious webpages detection. The dataset I trained and tested get 99% accuracy. This project can reach the application level with the help of a library called pickle, to save what the algorithm has learned and then we can test a new file to see if it is clean or infected. Static analysis has also proven to be safer and free fr the overhead of executontime.

8. REFERENCES

[1] Malware Types and Classifications, Bert Rankin, 28.03.2018, published in LastLine, last accessed 12.09.2018.

[2] A Brief History of Malware - Its Evolution and Impact, Bert Rankin, 05.04.2018, published in LastLine, last accessed 12.09.2018.

[3] Detecting malware through static and dynamic techniques, Jeremy Scott, 14.09.2017, published in NTT Security, last accessed 12.09.2018.

[4] Hybrid Analysis and Control of Malware, Kevin A. Roundy and Barton P. Miller, International Workshop on Recent Advances in Intrusion Detection, pp. 317-338, 2010, Springer.

[5] Advanced Malware Detection - Signatures Vs. Behavior Analysis John Cloonan Director of Products, Lastline, 11.04.2017, published in Infosecurity Magazine, last accessed 12.09.2018.

[6] What is Machine Learning? Daniel Faggella, 12.08.2017, published in techemergence, last accessed 12.09.2018.

[7] Data mining, Margaret Rouse, Search SQL Server last accessed 12.09.2018, the article can be found here.
• https://searchsqlserver.techtarget.com/definition/data-• • mining• •

[8] Supervised and Unsupervised Machine Learning Algorithms, Jason Brownlee, 16.03.2016, published in Machine Learning Algorithms, last accessed 12.09.2018.