RESEARCH ARTICLE OPEN ACCESS

Taking a Deep Bi-Directional LSTM Approach for Tackling the Malicious URL Detection Problem

Amalesh A, Gowthamy J

[1],[2] Computer Science, SRM Institute of Science & Technology, Chennai, Tamil Nadu - India

ABSTRACT

Network security can be compromised by several malicious domains. There have been considerable attempts to identify unknown Domain Generated Algorithms or DGA-generated with unique techniques. In this effort, a calculation dependent on DBLSTM model is imagined. The expected outcome demonstrates that the said classifier can achieve better precision, performing considerably better than conventional approaches.

Keywords: - Deep Bi-directional LSTM, DGA-generated domains, Neural Networks, Logistic Regression, Support Vector Machine.

I. INTRODUCTION

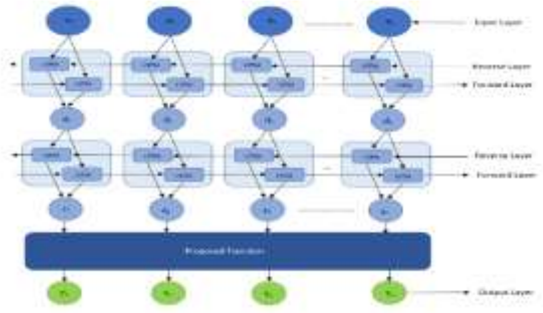
Ever since technology bloomed, the Internet was advancing at an exceptional speed. As of late, progress was not just made in the advancement of media but in energy networks as well. Ever since technology bloomed, the Internet was advancing at an exceptional speed. As of late, progress was not just made in the advancement of media but in energy networks as well. The goal is to create a intelligent and well-connected system that allows assimilation of data and commercial outflow. Zhong's research in 2016 suggested SDEI architecture that separates control, and data into distinct structures. The design allows for future energy flow programmability and peer-to-peer energy distribution. It will also be fascinating to observe the role of DL algorithms in the evolution of the Internet. Later, Chen embraced a certain deep learning method to achieve the intended strategy.

While an internet system includes many users and devices, it is essential to ensure the system's security. It is difficult for the field to blossom without network security. There have been many instances where a simple interruption in the passage of exchange causes huge losses in data. Since 2016, it is reported that approximately 1.4 million unaware individuals have fallen prey to cyber predators and there are many more stuck in this unfortunate situation. Furthermore, such malware causes significant network damage that renders the central command centers futile for over a period of months. Since major economies depend on the internet, and if cyber predators are aware how to target the said centers, the world's digital infrastructure is at risk. As a result, there is a requirement for an effective strategy for dealing with the possible threat.

II. CORE METHOD

Traditional machine learning approaches have been developed by researchers to identify harmful domains. However, with such approaches, the classifier would rely on human-influenced characteristics. This study used a deep learning approach to keep the characteristics objective and let the machine develop them on its own. Because a domain may be thought of as a series of events, this study used Long Short-Term Memory (LSTM). Although LSTM preserves and uses knowledge from the past, it does not conserve or benefit from eventual data. The researchers, eventually, chose Deep Bidirectional LSTM classifier.

Essentially, every letter is an info feature. A forward and a reverse LSTM sequence are then used to process these characteristics independently. The result of the forward series is determined using positive info, whereas the output of the reverse sequence is determined using negative info. The results are integrated and sent into the proposed module, which normalizes the data to PD and generates the concluding result. **Fig (1)** depicts a DBLSTM Classifier structure.



III. PROPOSED METHOD

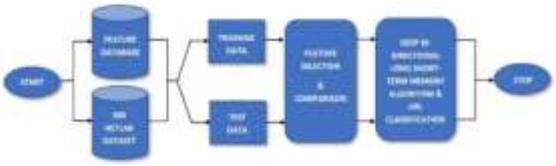

START
FEATURE DATABASE
360 NETLAB DATASET
TRAINING DATA
TEST DATA
FEATURE SELECTION & COMPARISON
DEEP BI DIRECTIONAL LONG SHORT-TERM MEMORY ALGORITHM & URL CLASSIFICATION
STOP

Fig (2). *Illustration of the system architecture.*

Fig (2) depicts the proposed system's architectural diagram which begins with the procurement of datasets, followed by feature engineering and comparison. Testing set size is then compared accompanied by the comparison of classifiers. Finally, LSTM network adjustment is done before running the Unknown DGA simulation.

A. *Procurement Phase*

In the procurement division, accessible datasets from Kaggle, Alexa, and 360 NetLab are acquired.

- **Safe URLs:** Over thousands of datasets are acquired from the Alexa and Kaggle dataset and serve as test data
- **DGA URLs:** Thousands of dangerous datasets are acquired from 360 NetLab as test data.

B. *Feature Engineering*

Lexical characteristics of URLs are used as features because they are simple to analyze and can be used on a broad variety of DGA families. The following are the lexical characteristics that are considered in this approach:

- **Length of the URL(l):** A regular URL is generally shorter than a malicious URL since it maintains the norm.
- **Vowel(V%) & Digits percentage(d%):** Vowels should be considered as they make the intelligible while numbers in an abnormal sequence make them complex.
- **Dots(.):** They resemble distinct sections of a URL. If present in excess, the domain's trustworthiness suffers.
- **Secure test(ST) and malignant test(MT):** A typical website contains .com, .biz, etc as their top-level domain, or in simpler words, their suffix. Hence, suffixes are collected and used as features.

C. *Feature Comparison*

The accuracy of an individual feature is assessed during the experiment with LR and SVM and the contrast is illustrated in **Fig.(3).**

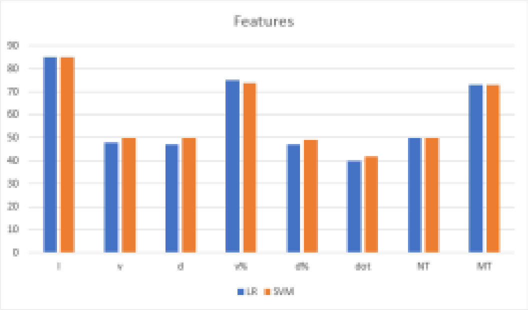


Fig (3). *Comparison of features between LR and SVM.*

Fig. (3) illustrates that any feature, regardless of model, may provide an accuracy rate of more than 40%. The chosen characteristics stand out from the others since they achieve an accuracy rate of more than 60%. These findings show that different characteristics have varying effects on overall accuracy.

D. *Contrasts Of Test Sets*

Ensuing feature comparison, comes an experiment that compares the efficiency of several models. By employing features and the training dataset domains, the testing set size is changed while the rest remain unchanged.

As the testing set size grows larger, both models show a general tendency of deterioration. Furthermore, while SVM outperforms LR when the testing set size is small, and vice versa when the size is increased.

E. *Contrast Of Classifiers*

The efficiency of DBLSTM model is compared to existing models. With the size constant, a dataset is chosen at random from the sources is utilized to improve the test set's impartiality and the contrast in efficiencies is depicted in **Fig (4).**

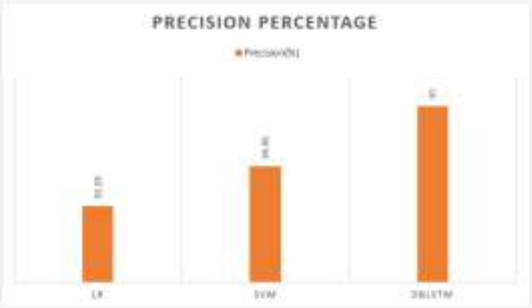


Fig. (4) *Precision % vs Classification Models*

When the LSTM model is used, the accuracy rate clearly improves (**Fig. (4)**). When a dataset is chosen at random, the accuracy rate improves to 98%.

F. LSTM Network Adjustment

This section analyses the effects of the model adjustment performed by modifying the length of one and all char in the domain in addition to number of LSTM layers on DGA-generated domains identification as to investigate the strength of different criterions on detection of DGA-generated domains. Exact impacts are depicted in **Fig. (5)** and **(6)**.

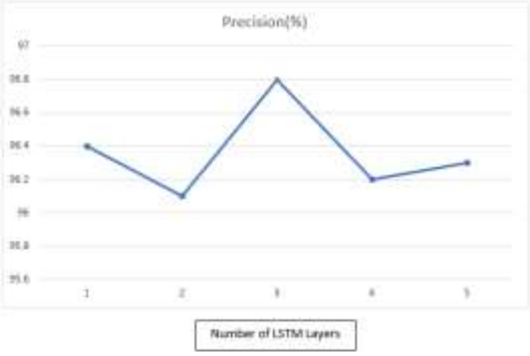


Fig (5). *Accuracy (%) vs. LSTM layers*

Fig. (5) demonstrates that for the identical function, the model performs best when the concealed size is set to three while the rest are held constant before altering the size.

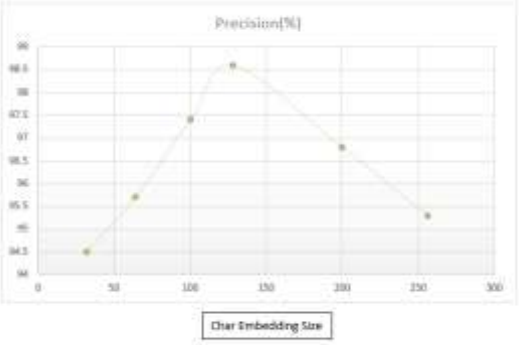


Fig. (6) *Accuracy % vs Char Embedding Size*

Fig.(6) indicates that, when provided an identical job, the model performs best when the embedding size of char is equal to 128, while all other criterions remains the same.

G. Unidentified DGA Simulation

The mechanisms in the previous sections were tested using the same origin for both train and test sets. In this part, the suggested technique is evaluated against datasets of several thousand domains created by the following mechanism to mimic unidentified DGAs. The technique generates a URL of length 20 in the prescribed format with each component created at random. The accuracy of the method is observed in **Fig (7)**.

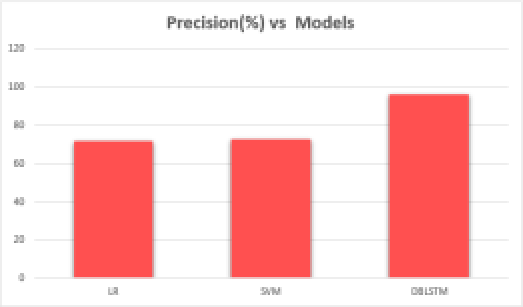


Fig (7). *Accuracy rates after DGA simulation*

The accuracy of our proposed method is at 96 percent, whereas the conventional methods has dropped substantially to less than 75 percent **(Fig. (7))**.

IV. CONCLUSIONS

This paper offers various techniques for detecting DGA produced domains based on URL characteristics. The outcome demonstrates that the DBLSTM method outperforms other approaches. The deep learning technique provided in the study has the potential to be pertinent in the field of cybersecurity, particularly in network security, to distinguish assaults launched by various domain creation methods.

To be sure, adjustments may be done to get better results. First, by picking additional lexical characteristics, the significant drop in accuracy rate of traditional methods may be mitigated. Other linguistic characteristics can be studied further. However, it also demonstrates the benefit of DL techniques because the DBLSTM method avoids feature engineering. Further analysis will be conducted to mimic cyberattacks and evaluate the DL method's efficiency in actual time setting.

ALGORITHMS	**ACCURACY RATE (%)**	**ACCURACY RATE AFTER DGA-SIMULATION (%)**
SVM	**94.95**	**72.24**
LR	**93.59**	**71.7**
DBLSTM	**97**	**97.02**

Table (1)*. depicts the contrast in the accuracy rates of algorithms before and after DGA-simulation.*

REFERENCES

[1] Vinayakumar R, Sriram S, Soman KP, Mamoun Alazab, "Malicious URL Detection Using Deep Learning"

[2] Yuchen Liang, Xiaodan Yan. "Using Deep Learning to Detect Malicious URLs", 2019 IEEE International Conference on Energy Internet (ICEI), 2019.

[3] Zhiqiang Wang, Xiaorui Ren, Shuhao Li, Bingyan Wang, Jianyi Zhang, Tao Yang, "A Malicious URL Detection Model Based on Convolutional Neural Network" International Symposium on Security and Privacy in Social Networks and Big Data, SocialSec 2020: Security and Privacy in Social Networks and Big Data pp 34-40.

[4] Gianluca Stringhini, "Detecting Spammers on Social Networks" Neurocomputing, Volume 159, Pages 27-34.

[5] M. Antonakakis, R. Perdisci, Y. Nadji, N. Vasiloglou, S. Abu-Nimeh, W. Lee, and D. Dagon, "From throw-away traffic to bots: detecting the rise of DGA-based malware," in Presented as part of the 21st {USENIX} Security Symposium ({USENIX} Security 12), pp. 491–506, 2012.

[6] Sangho Lee, Jong Kim, "WarningBird: Anear Real-Time Detection System for Suspicious URLs in Twitter Stream" IEEE Transactions on Dependable and Secure Computing (Volume: 10, Issue: 3, May-June 2013).

[7] Neda Abdelhamid, Aladdin Ayesh, Fadi Thabtah, "Phishing Detection based Associative Classification data mining" October 2014 Expert Systems with Applications.

[8] Elaheh Biglar Beigi, Hossein Hadian Jazi, Natalia Stakhanova, Ali A. Ghorbani, "Towards Effective Feature Selection in Machine Learning-Based Botnet Detection Approaches" 2014 IEEE Conference on Communications and Network Security.

[9] O. E. David and N. S. Netanyahu, "Deepsign: Deep learning for automatic malware signature generation and classification," in 2015 International Joint Conference on Neural Networks (IJCNN), pp. 1–8, IEEE, 2015.

[10] Justin Ma, Lawrence K Saul, Stefan Savage, Geoffrey M. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs"

[11] Ben Athiwaratkun, Jack W. Stokes, "Malware Classification with LSTM and GRU Language Models and a Character-Level CNN" 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[12] Toshiki Shibahara, Kohei Yamanishi, Yuta Takata, Daiki Chiba, Mitsuaki Akiyama, Takeshi Yagi, Yuichi Ohsita, Masayuki Murata, "Malicious URL Sequence Detection Using Event De-Noising Convolutional Neural Network" 2017 IEEE International Conference on Communications (ICC).

[13] Doyen Sahoo, Chenghao Liu, Steven C.H. Hoi, "Malicious URL Detection using Machine Learning: A Survey" ARXIV 2017

[14] Ignacio Arnaldo, Ankit Arun, Sumeet Kyathanahalli, Kalyan Veeramachaneni, "Acquire, Adapt, and Anticipate: Continuous Learning to Block Malicious Domains" in 2018 IEEE International Conference on Big Data.

[15] Ting-Fang Yen, Alina Oprea, Kaan ONarlioglu, Todd Leetham, William Robertson, Ari Juels, Engin Kirda. "Beehive: Large-Scale Log Analysis for Detecting Suspicious Activity in Enterprise Networks".