Phishing detection by classification of URLs using Machine learning Algorithms

1Keshav Adkar, 2Shaurya Agarwal

Students
Department of Software Engineering
SRM Institute of Science Technology
Chennai 603203, India

Abstract- Since phishing does not initially appear to be malicious, it is difficult to track down or defend against. Additionally, the cost and difficulty of carrying out such attacks are significantly declining. As it is challenging to know if an URLs safe or not, we created a model which serves in classifying URLs into safe or legitimate class. URLs contain components of its page and hence are used to identify the purpose of the web page without checking its actual content. We are proposing a method based on a deep learning algorithm (Bi-Directional Long short-term memory), which predicts the status of URL without the use of domain expertise or manual feature extraction with more significant accuracy points than already existing systems.

Keywords: Phishing, deep learning; URL classification; cybercrime; recurrent neural networks; bidirectional long short-term memory

1. INTRODUCTION
Phishing is an act of deceiving an online user to gather personal data by acting as a trustworthy institution or entity. Phishing is a significant threat to all users on the internet. Attackers attempt phishing to obtain sensitive information such as usernames, passwords, and credit card details through electronic communication. It is done through calls, emails, websites and electronic mediums. For example, one can receive a mail which looks like a regular request, but it may download malicious software and can gather saved passwords. These attacks are fabricated in a way that leads consumers to expose their financial data such as usernames and passwords in fraudulent websites pretending as legitimate ones.

The "phishing" name appeared in the mid-1990s among hackers trying to dupe AOL users into revealing their login information. The "ph" is an element of a tradition of whimsical hacker spelling and was influenced by the term "phreaking," short for "phone phreaking," an original form of hacking that involved using sound tones to make free phone calls through telephone handsets.

 Nowadays, phising attacks can be originated from anywhere in the world at small costs by people with little technical skills. Institutions trying to protect their users from these attacks are having a difficult time dealing with a large number of rising sites, which must be classified and labelled as malicious or safe before users can access them. Also, the number of attacks and phishing webpages are on rising. According to recent data, phishing has risen by 1300% in 4th quarter of 2019.

The method of analysing URLs instead of the entire content gives a decisive advantage in identifying malicious websites. By this method, we reduce computational cost and effort significantly, which is required by sophisticated methods like the content analysis. For this work, we are focusing on all classification modeltechnique, specifically naive bayes, for analysing the URLs. Catboost and XGboost models perform exceptionally at detecting long patterns in sequences. They have solved different text analysis problems recently. It learns directly from URL’s sequence of characters to find a pattern. When each model trains on a portion of data and evaluates a different subset of data, this is known as ordered boosting. Ordered boosting has advantages, such as improving robustness to unknown data.

2. RELATED WORK
Earlier, there were systems which used to analyse the content of webpages to predict their status. In 2007, Yue Zhang, Jason Hong, and Lorrie Cranor introduced a method that identifies phishing websites based on the retrieval of information about webpage content, providing about 95% accuracy. In 2011, Mona Ghotaish Alkhazae and Omar Abdullah Batarfi presented a phishing classification method that validates the reliability of a web page by looking into the source code. In 2013, Jian Mao, Pei Li, Kun Li, Tao Wei, and Zhenkai Liang presented a new approach to identify phishing efforts that use an algorithm to assess the web page's legitimacy is determined by how similar it seems to other pages. These systems demanded a lot of operational costs and computational time. It is far more efficient to analyse an URL and its nature than to analyse the entire content.

Recently, machine-learning algorithms and other statistical methods predicted the nature of URLs. In 2010, Maher Aburrous, Md Alamgir Hossain, Keshav Hahal, and Fadi That developed an intelligent fuzzy data mining algorithm to classify phishing sites that disguised as online banking sites. They were able to conclude from their research that the URL and the domain identity are essential criteria for evaluating
phishing sites. Over time, with research, it was inferred that the random forest algorithm is best suited for such problems.

In another paper, Bahnsen showed that LSTM could give at least five higher accuracy points than the random forest. As LSTM learns the sequential pattern, we do not need to extract features manually. Additionally, there is little to no need for domain-specific knowledge when using this algorithm. However, even then, it holds room for improvisation through some modification in the algorithm.

3. PROPOSED METHODOLOGY
We propose a system which uses various machine learning algorithms for predicting the malicious status of URLs.

We used XGBoost and Catboost that can enhance model performance on sequence classification problem statements such as this. It is useful in problems where all time steps of the input sequence are available, both on an inverted replica of the input sequence and the first on the input sequence as-is. This method can provide added context to the network and result in faster and even fuller learning on the problem. Regression and Classification algorithms were also used to get the best possible output with highest precision and accuracy.

3.1 Dataset
In total, numerous URLs were scrapped to form dataset—half of them legitimate and half of the phishing. The legitimate URLs came from just Common Crawl, an organisation which collects web crawl data. The phishing URLs came from Phishtank, an anti-phishing site which has a dataset of millions of phishing URLs. One can easily see the similarity between legitimate URLs and phishing URLs. The URL’s are classified by dividing it into various categories such as LongURL, ShortURL and so on into numerous categories.

Two-third of the dataset will be used for training purpose, while the remaining one-third will be used for validation of model.

RESULTS
We implemented various ML models in colab with theanoas backend. As already mentioned, we have divided our dataset into three folds. Two folds will be used for training, the model, while the remaining fold will be used for validation of the model. Furthermore, in each fold, 90% of dataset was used for training and the remaining 10% for internal validation. In the figure, the learning curve for the model is shown, and it’s clearly visible that after 15th epoch accuracy was over 99%.

In table 1, results for the model are shown. The Naïve Bayes Classifier model has an average accuracy of 99.26%, with a standard deviation of just 0.03%, suggesting stable performance across the different folds. Moreover, as shown in Fig. 7, the AUC of the model is very high, having an average of 99.91%.

In table 1, results for the model are shown. The Naïve Bayes Classifier model has an average accuracy of 99.26%, with a standard deviation of just 0.03%, suggesting stable performance across the different folds. Moreover, as shown in Fig. 7, the AUC of the model is very high, having an average of 99.91%.
The training time for system is almost 256 minutes. During the process, 0.623 MB of memory was consumed.

CONCLUSION/FUTURE WORK
The model provides proof that Naïve Bayes Classifier provides higher accuracy points as compared to other ML models. We get this result as our model propagates in both directions, hence finding better-fitted patterns with almost the same computational power. With these results, we can easily conclude that a pattern of URLs can act as a deciding factor in classifying a web page into phishing or legitimate site. The accuracy can be increased further by changing the order of neural network layers. For that, further research is necessary with different orders of layers. Use of Ind RNN will also be compared with the existing system to check the best-suited model for this problem statement.

ACKNOWLEDGEMENT
Our deepest gratitude to Dr. K. Kottilingam, an Associate Professor of S.R.M Institute Of Science and Technology in Chennai, for providing us with the opportunity to do research and for their important guidance during the project completion.

REFERENCES:
1. Alejandro Correa Bahnsen, Eduardo Contreras Bohorquez, Sergio Villegasy, Javier Vargasy and Fabio A. 2016. Gonz'alez, "Classifying Phishing URLs Using Recurrent Neural Networks".
2. Yuchen Liang1, Jiandong Deng, and Baojiang Cui, “Bidirectional LSTM: An Innovative Approach for Phishing URL Identification”, 2019
10. Jason Brownlee, “How to Develop a Bidirectional LSTM for Sequence Classification in Python with Keras”, machinelearningmastery.com

<table>
<thead>
<tr>
<th>Fold</th>
<th>AUC</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.999</td>
<td>0.992</td>
<td>0.991</td>
<td>0.9832</td>
<td>0.9871</td>
</tr>
<tr>
<td>1</td>
<td>0.999</td>
<td>0.992</td>
<td>0.989</td>
<td>0.9863</td>
<td>0.9879</td>
</tr>
<tr>
<td>2</td>
<td>0.999</td>
<td>0.992</td>
<td>0.987</td>
<td>0.9885</td>
<td>0.9878</td>
</tr>
<tr>
<td>Average Std. Dev.</td>
<td>0.999</td>
<td>0.992</td>
<td>0.989</td>
<td>0.9860</td>
<td>0.9876</td>
</tr>
</tbody>
</table>