UDC 004.773.3

DETECTION OF PHISHING AND TRACKING SYSTEMS IN EMAILS USING NEURAL NETWORKS

ALIAKSANDR LABANAU, KATSIARYNA HATSIKHA
Polosk State University, Belarus

This paper describes techniques and mechanisms built on the basis of neural networks that will help in the identification of phishing and tracked emails.

Currently, almost every Internet user uses an e-mail to register on various services, exchange corporate information and for other needs. An e-mail is a cheap and convenient way of communication between people. However, it can be used to access personal user data, as well as fraudulent activities such as phishing [1].

The principle of work of most tracking services is based on the injection of some picture in the email body that is located on an external server. When the recipient opens the email, a request is sent to the tracking service. In the response the text that is loaded by the browser or the client application of the user comes, at this moment images, styles and other elements of the HTML page from external servers are downloaded, one of which can be a tracking service [2].

Phishing is one of the problems of a modern network, it is a fake Web-site by which cybercriminals steal confidential information about a user without their knowledge (for example, the name, the password, the bank or card information, etc.). The purpose of attackers when creating a phishing URL is to deceive users, resulting in a fake resource that will receive the personal and financial data of the victim [3].

Most of the existing systems for detecting phishing resources use the following approaches to be safe from this type of threat:

– "Black list" approach. Determine if the requested URL is in the list of malicious resources. The drawback of this approach is that a black list cannot normally cover all phishing and tracking sites. Since the creation of a fraudulent website, it may take a long time before it is added to the list. And this time gap between starting and adding a suspicious website to the list can be enough for tracking systems.

– Heuristic approach meets several related website metrics to classify it as a phishing or this website. Unlike the "black list" method, using a real-time heuristic approach, you can recognize newly created fake Web sites.

Based on bellow, an heuristic approach was used to develop software to identify phishing sites or e-mail tracking systems. All references to objects that contains URLs to third-party resources (images, buttons, styles and other objects) are transferred to a system that as a result of its work determines whether the site is real and does not contain any information that will help track the message, or the investigated resource is phishing. To implement this algorithm, neural networks were used in addition to such systems as Google PageRank and Alexa rating.

In this paper, we used 14 different metrics, which will be used to create lexical and statistical analysis of URLs (Fig. 1):

– Presence of a domain in Alexa Rank, Google Page Rank.

– Domain length, subdomain length, path length. Phishing resources try to use a domain similar to the current resource.

– Entropy of URL address. The higher this value, the more complex the URL. Because the URLs of phishing resources have random text, you can try to find them by their entropy.

– The ratio of the length between parts of the address.

– Some service defined symbols (about 5 metrics) - like '@' or '-'.

– Number of suspicious words, Euclidean distance, Kolmogorov-Smirnov statistics, Kulbak-Leibler divergence.

When choosing neural networks, we select recurrent neural networks (RNN) – this is a type of network capable of simulating consecutive patterns [4]. A distinctive feature of RNN is the concept of time for the model, which, in turn, allows you to process sequential data on one element at a time and study their successive dependencies [5]. The limitation of RNN is the inability to recognize the correlation between elements of more than 5 or 10 steps from each other. The model that overcomes this problem is LSTM [6]. Therefore, LSTM blocks were used to build a model that receives the URL as an input parameter and predicts whether the URL corresponds to a phishing event or a monitored service (Fig. 2).

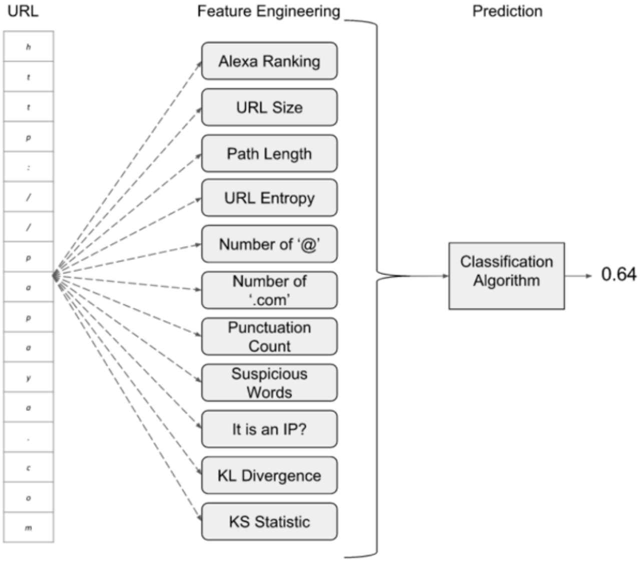
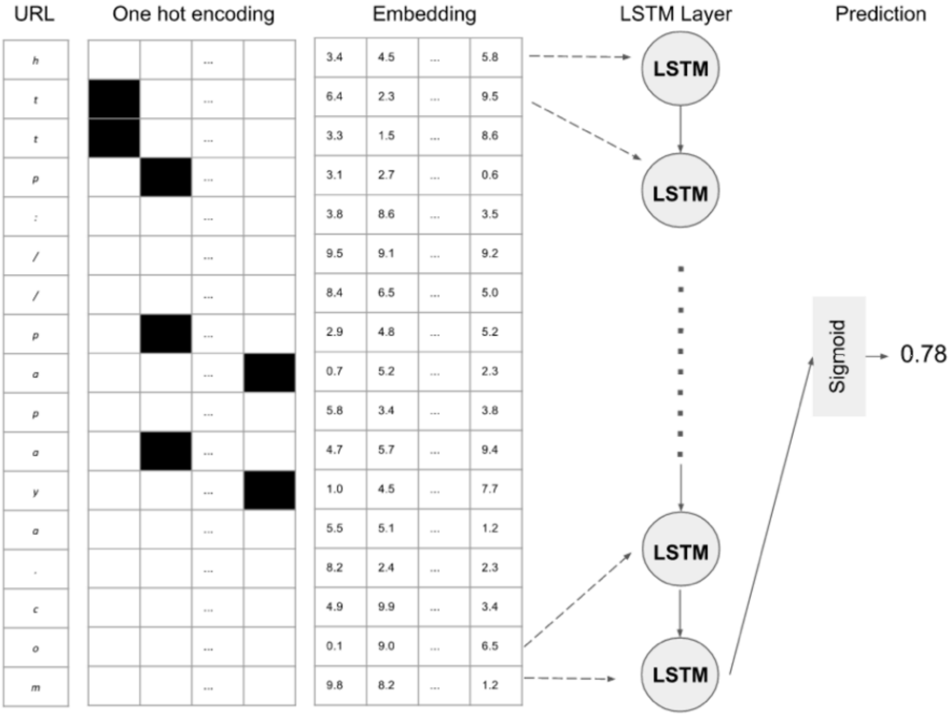


Fig. 1. The approach used to classify phishing URLs



h	3.4	4.5	...	5.8
t	6.4	2.3	...	9.5
t	3.3	1.5	...	8.6
p	3.1	2.7	...	0.6
:	3.8	8.6	...	3.5
/	9.5	9.1	...	9.2
/	8.4	6.5	...	5.0
p	2.9	4.8	...	5.2
a	0.7	5.2	...	2.3
p	5.8	3.4	...	3.8
a	4.7	5.7	...	9.4
y	1.0	4.5	...	7.7
a	5.5	5.1	...	1.2
.	8.2	2.4	...	2.3
c	4.9	9.9	...	3.4
o	0.1	9.0	...	6.5
m	9.8	8.2	...	1.2

Fig. 2. Neural network for classification of URLs of phishing based on LSTM blocks

After training the neural network using two million different URLs that contained 5% of phishing sites, the implemented software passed the test, which resulted in the indicators shown in Table 1.

Table 1. – Results of the operation of the system based on LSTM

Fold	AUC	Accuracy	Recall	Precision	F1-score
0	0.999044	0.9871	0.991114	0.983203	0.987143
1	0.999106	0.987921	0.989549	0.986359	0.987952
2	0.999141	0.987844	0.98716	0.988506	0.987833
Average	0.999097	0.987622	0.987622	0.986023	0.987642
Std dev	4e-05	0.00037	0.00037	0.002178	0.000357

After analysing the results, we can conclude that parsing the URL by their samples is a good way to detect phishing sites and email tracking systems. Creating a system to secure against phishing and tracking systems based on URLs recognition is much faster and more efficient than full-text analysis. LSTM can evaluate URLs with the speed of 942 addresses per second. The LSTM model uses only 581 KB of memory to process the test data set. It is a crucial moment while installing the software into the systems with a limited access and the amount of resources.

REFERENCES

1. Phishing Activity Trends Report, 3rd Quarter 2016, Tech. Rep. : APWG, December, 2016.
2. Email Tracking [Electonic resource]. – https://en.wikipedia.org/wiki/Email_tracking.
3. Why Phishing Works : SIGCHI Conference on Human Factors in Computing Systems, 2006 / R. Dhamija, J. D. Tygar, and M. Hearst. – P. 581–590.
4. Dietterich, T. Machine learning for sequential data: A review : Structural, syntactic, and statistical pattern recognition / T. Dietterich. – 2002. – P. 1–15.
5. Lipton, Z.C. A Critical Review of Recurrent Neural Networks for Sequence Learning / Z.C. Lipton. – 2015. – P. 1–38.
6. Gers, F.A. Learning to forget: continual prediction with LSTM. Neural computation / F.A. Gers, J. Schmidhuber, and F. Cummins. – 2000. – Vol. 12, № 10. – P. 2451–2471.