Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 47, S. 13-17, Ocak 2023
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 47, pp. 13-17, January 2023
Copyright © 2023 EJOSAT
Research Article

Developing and Evaluating an Artificial Intelligence Model for Malicious URL Detection

Fatih Tiryaki[1*], Ümit Şentürk[2], İbrahim Yücedağ[3]

[1*] Düzce University, Departmant of Cyber Security, Düzce, Turkey, (ORCID: 0000-0002-4855-1069), fatih.tiryaki@mgu.edu.tr
[2] Abant Izzet Baysal University, Faculty of Engineering, Departmant of Computer Engineering, Bolu, Turkey, (ORCID: 0000-0001-9610-9550), umit.senturk@ibu.edu.tr
[3] Düzce University, Departmant of Computer Engineering, Düzce, Turkey, (ORCID: 0000-0003-2975-7392), ibrahimyucedag@duzce.edu.tr

(3rd International Conference on Engineering and Applied Natural Sciences ICEANS 2023, January 14 - 17, 2023)

(**DOI:** 10.31590/ejosat.1234556)

ATIF/REFERENCE: Tiryaki, F., Şentürk, Ü. & Yücedağ, İ. & (2023). Developing and Evaluating an Artificial Intelligence Model for Malicious URL Detection. *European Journal of Science and Technology*, (47), 13-17.

Abstract

Today, the increased use of the internet has become important in our lives and new communication technologies, social networks, e-commerce, online banking, and among other applications have a significant impact on the promotion and growth of business. In our study, we aimed to work with a large dataset and to achieve the best results in detecting malicious URL addresses using an artificial intelligence model. A 7-layer RNN model was used in the study, and two similar national and international datasets were combined and merged to create a big new dataset consisting of 579,112 URL addresses. Then, this new data set is divided into training and test sets. first data set was trained at the model and then the second data set was processed test. When this data set was processed in our model, we achieved a success rate of over 91%. This rate is a very good result of detecting malicious url addresses. Your contribution with this work is significant in developing more effective methods for detecting harmful sites as internet usage increases, parallel use of artificial intelligence models makes detection of such sites more effective and potentially assist users in protecting from various types of cyber-attacks is targeted.

Keywords: Malicious URL, Cyber Security, Artificial Intelligence, RNN Model, Accuracy

Kötücül URL Tespitinde Yapay Zekâ Modeli Geliştirme ve Değerlendirilmesi

Öz

Günümüzde internetin her geçen yıl kullanımın artmasıyla hayatımızda çok önemli bir hale gelmiş ve yeni iletişim teknolojileri, sosyal ağlar, e-ticaret, çevrimiçi bankacılık dâhil olmak üzere birçok uygulamada işlerin teşvik edilmesinde ve büyütülmesinde önemli bir etkiye sahiptir. Yaptığımız çalışmada, kullandığımız yapay zekâ modeli ile zararlı URL adreslerini tespitinde büyük bir veri seti ile çalışılması ve en iyi sonucu elde etmek hedeflenmiştir. Çalışmada 7 katmanlı RNN modeli kullanılmış, modelde çalıştırmak üzere ulusal ve uluslararası birbirine benzer iki adet veri seti birleştirilmiş, 579.112 adet URL adresinden oluşan devasa bir yeni veri seti oluşturulmuştur. Daha sonra bu yeni veri seti eğitim ve test setlerine ayrılmıştır. İlk olarak veri setimiz modelde eğitilmiş ve ardından ikinci veri seti test edilmiştir. Bu veri seti modelimizde işlendiğinde %91'in üzerinde bir başarı oranı elde edilmiştir. Bu oran zararlı url adreslerini tespit etmesinde çok iyi bir sonuçtur. Bu çalışmamızla, internet kullanımı arttıkça zararlı sitelerin tespiti için daha etkin yöntemlerin geliştirilmesine önemli katkı sağlamakta, yapay zeka modellerinin paralel kullanımı bu tür sitelerin tespitini daha kolay hale getirmekte olup ve potansiyel olarak kullanıcıların çeşitli siber saldırı türlerinden korunmalarına yardımcı olması hedeflenmektedir.

Anahtar Kelimeler: Kötücül URL, Siber Güvenlik, Yapay Zekâ, RNN, Doğruluk

[*] Corresponding Author: fatih.tiryaki@mgu.edu.tr

1. Introduction

The use of the internet has become increasingly widespread and important in our lives. New communication tools and social networks have increased the use of the internet. [1]. In particular, the Covid-19 pandemic that swept the world in December 2020 further accelerated the process of internet use and widespread use, and the number of users roaming the internet has increased exponentially [2]. It is estimated that by 2022, the global number of internet users will approach 5 billion, with 2 billion websites. Banks, educational institutions, universities, businesses, and academic and scientific studies are all open to internet use and data sharing of all kinds. As the internet has become so widely used and indispensable in our lives, secure internet use has come to the forefront. As the number of users and websites on the internet has increased, cybercriminals have started to use website addresses, or URLs, for attacks. This malicious use of URLs is a major threat to cybersecurity and is referred to as "malicious URL." The rate of malicious URLs has increased significantly in recent years, the use of artificial intelligence-machine learning techniques has become increasingly important in detecting and preventing these attacks [3].

Chen et al. processed a dataset consisting of 200,000 URLs, including 100,000 normal URLs and 100,000 Malicious URLs, using the YOLO algorithm on an artificial neural network [4]. The features extracted were used to evaluate malicious URLs by a bidirectional LSTM recurrent neural network algorithm, and it was claimed that a success rate of 90% was achieved [5]. Ahammad et al. investigated lexical and domain-based features using random forests with a dataset of 35,300 URLs, comparing Decision Tree, Light GBM, Logistic Regression, and Support Vector Machine. They found that the Light GBM algorithm had the best result, with a success rate of 86% [6].

Kumar et al. designed a static feature-based Kaya using a dataset of 350,000 URLs, claiming a success rate of 90% [7]. Bharadwaj et al. claimed to have reached a success rate of 89% using a GloVe-based YSA model on 227,909 URLs [8]. Vecile et al. claimed a 79% accuracy rate using machine learning (LSTM) models on a dataset of 68,908 URLs [9]. Zahao et al. processed 300,000 URLs using their own detection algorithms, achieving a success rate of 90% [10]. Paydey et al. claimed a success rate of 85% using a machine learning-based deep learning model on a dataset of 20,000 URLs [11].

In this study, the following contributions have been made to the solution of the problem of detecting malicious URL addresses:

- A 7-layered deep neural network RNN model has been designed using features obtained from URL address information and with the necessary optimization processes.

- According to the studies in the literature review, a wider and universal data set is used and the availability and performance of the model is verified through the trained data set.

- It has been shown that the model used has a higher Accuracy rate than other models.

In the second section of our study, the working environment and the method used are discussed. In the third section, the URL structure, the data set used in the study, the Model Architecture, and the training of the model. In the fourth section, testing of the model, evaluation metrics, and comparison of the method are carried out. In the fifth section, the results obtained are evaluated.

2. Experimental Environment

2.1. Working Environment

In this study; an i5-7300U 2.60 GHz processor and 8 GB memory computer was use for the implementation. The Windows 10 Pro operating system was installed on the computer used and a wired internet connection was provided. Python was used as the programming language and colab.research.google.com was used as an internet-based platform.

2.2. Method Used

The training and testing of the data set used was carried out. It is showing the procedures of the study carried out figure 1. The data set section, where the data sets are combined, the processing section where the data is mixed, the model section where the RNN model is created, tokenizer, and epoch are made and artificial neural network section where the training is carried out and the results are obtained, are formed.

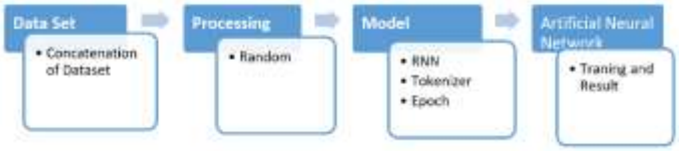


Figure - 1 URL Processing Diagram

PROPOSED SYSTEM

2.3. URL Structure

URL (Uniform Resource Locator) is the 'address' of any data such as a file or web page on the internet. It addresses specifies the source of the data. URLs consist of 4 parts [12]. These parts is showing at the figure 2.

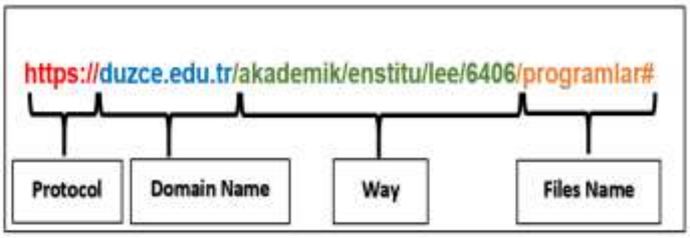


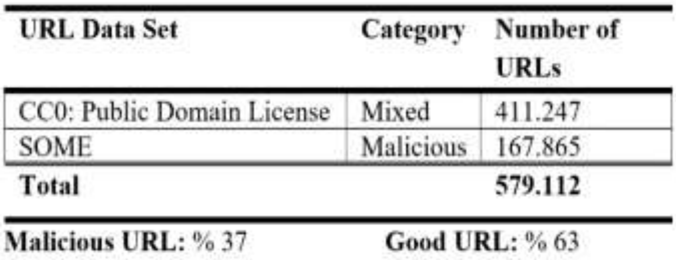
Figure - 2 URL Structure

URL addresses consist of four main parts; the first part is the protocol section and the connection type is determined. The “S” is for the security protocol. The second part is the domain name of the URL address. In the third part, the link path of the file within the URL address is shown. In the example, the link path in the file path within the institute subfolder in the academic main folder is specified. The fourth part is the name of the file.

2.4. Data Set Structure

The performance of the data set is largely dependent on the size of data set and the fact that it’s a more comprehensive international data. Great care was taken in preparing our real data set.

Table - 1 URL Data Set Number



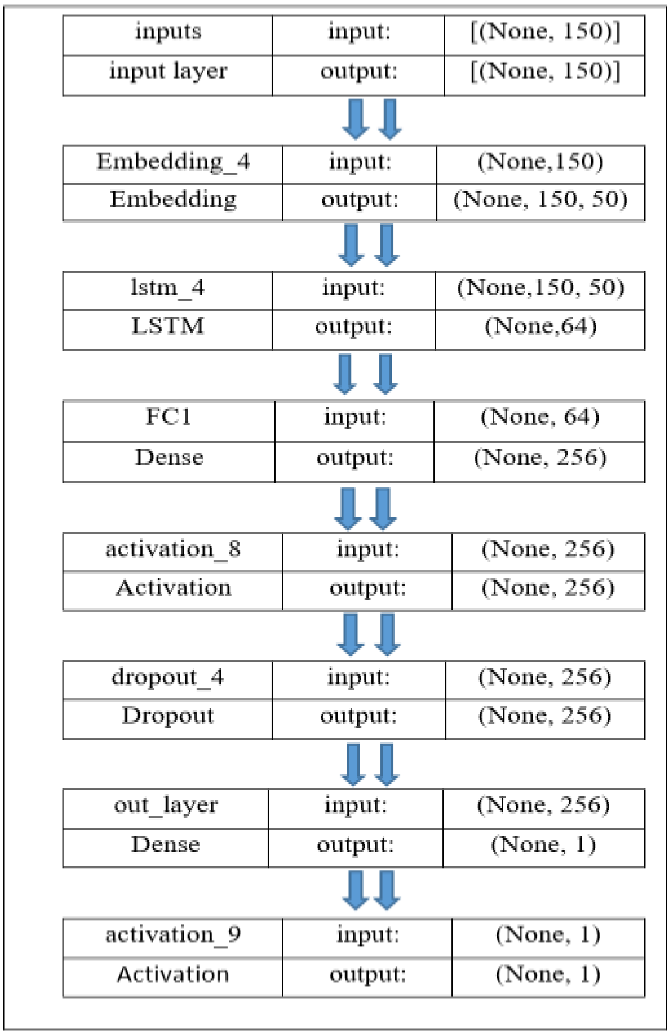
URL Data Set	Category	Number of URLs
CC0: Public Domain License	Mixed	411.247
SOME	Malicious	167.865
Total		**579.112**

Malicious URL: % 37 **Good URL:** % 63

Looking at Table 1, you can understand number the URLs and rates of malicious URLs. When preparing our final data set, a total of 579,112 URLs were used, including 411,247 URLs distributed by Kaggle under the CC0 Public Domain License [13] and 167,865 malicious URLs distributed by the National Cyber Intervention Center (SOME) [14] in the malicious links section on September 12, 2022. Thus, a large national and international data set was reached. Since the information in both data sets is the same and has the same purpose, they were combined as a single data set. The with new data set, is intended to create a wider and transparent data set with 37% malicious URL addresses and 63% non-malicious URL addresses.

2.5. Model Architecture

We completed the training of our data set by inserting it into a 7-layered RNN model and tried to reach the real accuracy value by processing our test data through the trained RNN model.



inputs	input:	[(None, 150)]
input layer	output:	[(None, 150)]
Embedding_4	input:	(None,150)
Embedding	output:	(None, 150, 50)
lstm_4	input:	(None,150, 50)
LSTM	output:	(None,64)
FC1	input:	(None, 64)
Dense	output:	(None, 256)
activation_8	input:	(None, 256)
Activation	output:	(None, 256)
dropout_4	input:	(None, 256)
Dropout	output:	(None, 256)
out_layer	input:	(None, 256)
Dense	output:	(None, 1)
activation_9	input:	(None, 1)
Activation	output:	(None, 1)

Figure - 3 Model Flow

Let's briefly describe the flow of our 7-layered RNN model. In the model creation, an embedding layer that can generate a maximum of 150 characters and a maximum of 50 words, a LSTM layer that generates a 1x64-dimensional feature vector, which enables deeper learning and reduces errors, and a 256-node output layer were used [15]. The relu function was used as the activation function in the output layer because I used multi-class data. Then, the binary_crossentropy was used as the loss function, the Adam algorithm as the optimization algorithm, and the accuracy as the metric. The test success was measured by repeating 10 epochs in the model.

2.6.Model Training

The data sets from two different sources were combined. In this study, the RNN model of artificial intelligence method was applied in Python. The newly prepared data set was defined in the system, the URL addresses were mixed, and then they were divided into training and test sets. These URL addresses allocated 20% of them (115,822) to the training set and 80% to the test set. In this way, the data sets were more accurately trained and the most accurate result was tried to be obtained.

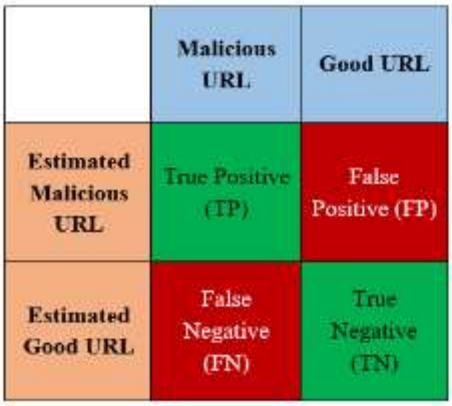
3. Results and Discussion

3.1. Model Testing

The test model was created with consisting of 463,289 URLs and predictions were made. Accuracy was calculated and seen that a result of over 91%. This rate is proof that the used model has high sensitivity and accuracy.

3.2. Evaluation Metrics

Precision, sensitivity, accuracy, and f score metrics were used for evaluate performance of the model.



	Malicious URL	Good URL
Estimated Malicious URL	True Positive (TP)	False Positive (FP)
Estimated Good URL	False Negative (FN)	True Negative (TN)

Figure - 4 Accuracy Table

True Positive (TP) and True Negative (TN) are area where model is correctly predicted while False Positive (FP) and False Negative (FN) are area where model is incorrectly predicted. These parts show at the figure 4.

- True Positive (TP): Malicious URL correctly predicted,
- True Negative (TN): Normal URL correctly predicted,
- False Positive (FP): Normal URL incorrectly predicted,
- False Negative (FN): Malicious URL incorrectly predicted.

- **Accuracy:** It is the ratio of the number of malicious URLs correctly predicted to total number of URLs. The calculation formula is as follows [16].

$$\textbf{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \qquad (1)$$

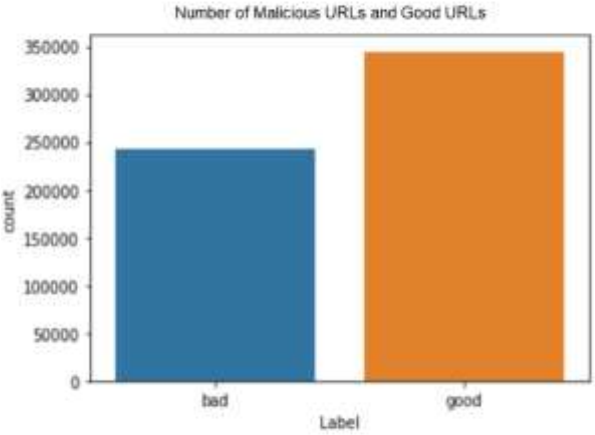


Image -1 Success rate of accuracy

Image 1 show the number of malicious-normal URLs and rate of accuracy.

- **Precision:** It is a metric that expresses what percentage of positive predictions are made for the situations that need to be predicted positively. The calculation formula is as follows [17].

$$\textbf{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (2)$$

- **Recall:** It is the ratio of the number of malicious URLs correctly predicted to number of normal URLs incorrectly predicted. The calculation formula is as follows [16].

$$\textbf{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \qquad (3)$$

- **F Score** It is a value used to evaluate precision and sensitivity values together. It is the harmonic mean of the two values. The calculation formula is as follows [17].

$$\textbf{F Score} = \frac{2 \text{ x } (\text{Precisionx Recall})}{\text{Precision} + \text{Recall}} \qquad (4)$$

A high accuracy value shows that the model has a successful classification ability for all types of URLs in general. A high sensitivity value indicates that the identification is as successful for the relevant type. A high precision value guarantees that the high accuracy value is obtained for the relevant type. A high f score value shows that the system's stability in classification results for all types is as high [16].

3.3. Comparison with Other Articles

In other articles that have been reviewed in the literature review, data sets ranging from 35,000 to 300,000 have been used, and machine learning models (decision tree, Gbm, logistic regression, etc.) and artificial neural network models (Yolo, Lstm, Glove-based, etc.) have been used to achieve success rates between 19% and 90%.

In this study, a more accurate and performance-oriented data set was targeted by using more URL addresses (579,112) and a 7-layer RNN model was used to achieve success rates above 91%.

4. Conclusion

The use of the internet has become increasingly widespread and important every year. The widespread is use of the internet and cyber criminals is used of internet website addresses, or URLs, by as targets for attacks. Therefore have caused the need for early detection of bad URLs to be very important.

In this study, two separate data sets with the same features, obtained from [13] and [14] were combined to create a large new data set of 579,112 URLs. The RNN model was processed in 7 layers, with 20% of the created data set used for training and 80% used for testing. The highest accuracy rate of 91% was obtained as a result of the performance analysis performed.

References

[1] Nora A. A. and Narmatha C (2022), A Systematic Approach for Malware URL Recognition, 2022 2nd International Conference on Computing and Information Technology (ICCIT) Jan. 25 - 27, 2022/ FCIT/UT/KSA.

[2] M. Alsaedi, F. A. Ghaleb, F. Saeed, J. Ahmad and M. Alasli. (2022), Cyber Threat Intelligence-Based Malicious URL Detection Model Using Ensemble Learning, Sensors 2022, 22, 3373. https://doi.org/10.3390/ s22093373.

[3] Bu, S.-J.; Kim, H.-J.(2022) , Optimized URL Feature Selection Based on Genetic-Algorithm- Embedded Deep Learning for Phishing Website Detection. Electronics, 2022, 1, 1090. https://doi.org/10.3390/ electronics11071090.

[4] Z. Chen Y. Liu, C. Chen, M. Lu and X. Zhang (2021), Malicious URL Detection Based on Improved Multilayer Recurrent Convolutional Neural Network Model, Hindawi Security and Communication Networks Volume 2021, Article ID 9994127, 13 pages https://doi.org/10.1155/2021/9994127.

[5] R. H. GBURI (2021), Detection of Malicious URLs Using Machine Learning, Yök Tez:704886.

[6] SK H. Ahammad, S. D. Kale, G.D. Upadhye et al (2022), Phishing URL detection using machine learning methods, Advances in Engineering Software 173 (2022) 103288.

[7] G. M. Kumar, Sri. S. K. Alisha and Sri. V. B. Murthy (2022), Detecting Mobile Malicious Webpages In Real Time, Journal of Engineering Sciences Vol 13 Issue 07,2022, ISSN:0377-9254.

[8] R.Bharadwaj, A. Bhatia, L. D. Chhibbar, K. Tiwari and A. Agrawal (2022), Is this URL Safe: Detection of Malicious URLs. Using Global Vector for Word Representation | 978-1-6654-1332-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICOIN53446.2022.9687204.

[9] S. Vecile, K. Lacroix, K. Grolinger and J. Samarabandu (2022), Malicious and Benign URL Dataset Generation Using Character-Level LSTM Models, 2022 IEEE Conference on Dependable and Secure Computing (DSC) | 978-1-6654-2141-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/DSC54232.2022.9888835.

[10] H. Zhao and Z. Chen (2022), Malicious Domain Names Detection Algorithm Based on Statistical Features of URLs, 2022 IEEE 25th International Conference on Computer Supported Cooperative Work in Design (CSCWD) | 978-1-6654-0527-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSCWD54268.2022.9776264.

[11] A. Pandey and J. Chadawar (2022), Phishing URL Detection using Hybrid Ensemble Model International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Vol. 11 Issue 04, April-2022.

[12] D. J. Bell, B. D. Loader, N. Pleace and D. Schuler (2004), Cyberculture: The Key Concepts, Routledge LONDON AND NEW YORK

[13]https://www.kaggle.com/datasets/teseract/urldataset/14.09.2022

[14] https://www.usom.gov.tr/adres./12.09.2022

[15] Ü. Şentürk, İ. Yücedağ and K. Polat (2018), Repetitive neural network (RNN) based blood pressure estimation using PPG and ECG signals, 2018 2Nd international symposium on multidisciplinary studies and innovative technologies (ISMSIT).

[16] R. S. Arslan, A Deep Learning Model for Malicious Url Filtering, European Journal of Science and Technology Special Issue 29, pp. 122-128, December 2021.

[17] H. Karamollaoğlu, İ. Yücedağ and İ. A. Doğru (2021), Customer Churn Prediction Using Machine Larning Methods: A Comparative Analysis, UBMK'2021 6th International Conferance on Computer and Engineering – 139.