Düzce Üniversitesi
Siber Güvenlik ve Dijital Ekonomi

Düzce University
Journal of Cybersecurity and Digital Economics
https://dergipark.org.tr/tr/pub/cdej

Research Article

Blocking Fraudulent Websites Using Artificial Intelligence

Semih Güner[a], *Büşra TAKGİL*[a]*

[a]Duzce University, Faculty of Engineering, Computer Engineering, Duzce, Türkiye.
[*]**Corresponding author:** busratakgil@duzce.edu.tr

Article Information:
Received: 09/05/2025 Revision: 26/05/2025, Accepted: 10/06/2025

ABSTRACT

As technology and digitalization play an increasingly important role in our lives, individuals, businesses, and governments that adapt to this trend are also becoming more vulnerable to cyberattacks. Among these attacks, phishing attacks are the most common. In these attacks, scammers use fake websites or emails to obtain your login credentials and other sensitive information. With the growing importance of cybersecurity, cybersecurity companies, academics, and governments have also begun to develop anti-phishing systems to counter such attacks. This study investigates how effective the architectures developed in the field of artificial intelligence in recent years can be in detecting phishing attacks through URLs. The performance of different artificial intelligence architectures was compared in the study. According to the results, the BERT architecture was the best performing network with an accuracy rate of 98%. While the DistilBERT architecture also had high test results, it gave incorrect results for some URLs. The CNN architecture, on the other hand, achieved a success rate of 91%, although it is older than the Transformer architecture.

Keywords: *Deep learning, Artificial intelligence, Transformer, Phishing attacks.*

I. INTRODUCTION

Developments in the internet and information technologies have also greatly increased access to these technologies. The increase in the number of devices that can connect to the internet and the decrease in their prices have played an important role in this increase. According to the International Telecommunication Union (ITU), 67% of the world's population will have access to the internet in 2023 (International Telecommunication Union, 2023). The conveniences that the internet and technological tools have brought to our lives have also brought with them some important security problems. Among these problems, the one with the highest potential to cause financial losses is "Phishing" attacks. In phishing attacks, attackers trap users by preparing e-mails, SMSs or social media messages that appear to come from a real institution or person. These messages usually mention an emergency and the user is made to enter the requested information or click on the link. According to FBI data for 2021, a total of $52 million was lost in the United States in 2021 because of phishing attacks (Federal Bureau of Investigation, 2021). Defense methods against these attacks include whitelisting and blacklisting, heuristic detection, visual similarity detection, machine learning, and its subfield deep learning. The aim of this study is to contribute to defense methods developed against phishing attacks with deep learning techniques. The study compares the success of different types of deep learning models in the URL classification task. Though the effects of deep learning models are studied in the literature, the results of studies found in other techniques and literature results were compared in this study.

Cite as: Güner, S. & Takgil, B. (2025). Blocking Fraudulent Websites Using Artificial Intelligence. *Düzce Üniversitesi, Siber Güvenlik ve Dijital Ekonomi*, 1(1): 22-28.

Jain and his colleagues tried to detect whether a website was used for phishing purposes using machine learning. Feature extraction was performed on the URL and Naive Bayes and SVM algorithms were trained with these features. This data was classified into 14 different categories, such as whether the URL was an IP address or not, and whether it contained more than two subdomains. Experts who performed training with two different data sets, including 10 thousand and 25 thousand URLs, observed that the best result was obtained with SVM in the training with 25 thousand URLs (Jain & Gupta, 2018). Mittal and his colleagues performed a classification study on 1000 URLs with BERT. 1000 links in the data set were divided into 4 categories as "well-intentioned, falsification, phishing, malware". Data augmentation was performed on the data. In this way, the scarce and unbalanced data was made available for efficient use in the study. A 96% success rate was recorded after training and testing (Mittal et al., 2023). Jawade and his colleagues performed phishing site classification using the Fast.AI convolutional neural network library. Using the ISCX-URL2016 dataset containing 36,707 data, the researchers measured 98.92% accuracy rate in training and testing (Jawade & Ghosh, 2021).

Studies in the literature focused on specific models and did not provide a specific comparison among other models. Originality of this study is comparing the findings with other methods in the literature, while also creating and comparing different techniques under deep learning.

II. METHOD

The dataset used in the study was created by combining data from three different sources. The first of these datasets is the "Malicious and Benign URLs" dataset, which is publicly available on the website Kaggle (Siddharth Kumar, 2019). Since this dataset consists mostly of international URLs, it was deemed essential to include Turkish-origin URLs in the dataset. The second dataset was downloaded from the malicious URLs publicly available to the National Cyber Incidents Combating Center (USOM) (USOM, 2024). To expand the dataset, the third dataset was used, which collects URLs of international phishing websites from the PhishTank website (PhishTank, 2024). A total of 755,922 URLs were obtained. While 345,738 of these URLs were benign URLs, 410,184 were phishing URLs.

Although the data was collected, not all the data could be used for training and testing because sufficient processing resources were not available. From the total dataset, 100,000 normal and 100,000 phishing labeled random data were drawn; these data were separated in an 80/20 ratio for training and validation. In this study, BERT, DistilBERT, and CNN architectures were trained for the sequence classification task. The datasets were separated as the 200,000 URL dataset announced above for transformer architectures and the 20,000 URL dataset for the network in the CNN architecture. BERT and DistilBERT networks were used by extracting from the APIs provided by the Python library Keras (Chollet et al., 2024). The CNN model was created using high-level functions provided by Keras. The design of the CNN model is given in Table 1.

Table 1. Design of the Generated CNN Model.

Layer Type	Output Size	Number of Trainable Parameters
Text Vectorization Layer	(None, 100)	0
Embedding Layer	(None, 100, 100)	1,000,200
Dropout	(None, 100, 100)	0
Conv1D	(None, 32, 128)	89,728
Conv1D	(None, 9, 128)	114,816
Conv1D	(None, 2, 64)	41,024
Global Max Pooling	(None, 64)	0
Frequent Neural Network	(None, 128)	8,320
Dropout	(None, 128)	0
Frequent Neural Network	(None, 1)	129

III.FINDINGS AND DISCUSSION

As a result of the training, it was observed that the fastest predictive network was CNN, then DistilBERT, then BERT; the BERT network was the model that made the most correct predictions among the three models. The DistilBERT model gave successful results in post-training tests; however, it almost always predicted URLs it

had never seen incorrectly. Dataset change and hyperparameter change were tried to solve the problem; however, the problem persisted. Therefore, DistilBERT was not considered successful in solving this problem. Validation and manual testing results of DistilBERT are given in Table 2 and Figure 1.

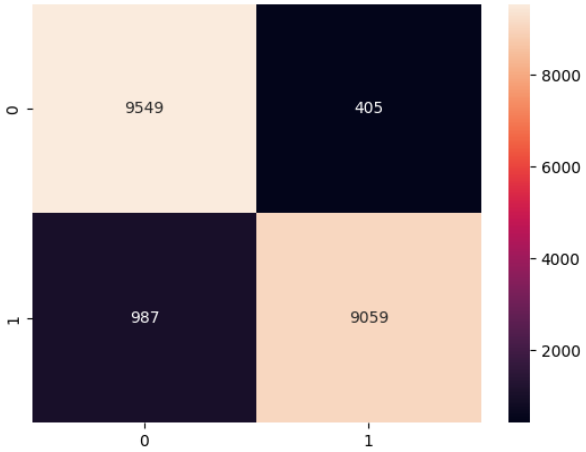

Figure 1. Confusion matrix of DistilBERT.

Table 2. Outputs corresponding to random inputs given to DistilBERT.

URL Input	Input After Preprocessing	Output
https://www.google.com.tr	google com tr	Phishing
https://www.google.com	google com	Phishing
https://usom.gov.tr	usom gov tr	Phishing

Although BERT was trained on the same dataset as DistilBERT, it did not experience the same problem. BERT was trained with a batch size of 64 and 4 epochs. As can be seen from the results, the BERT network achieved a 98% success rate in most measurements. The desired distribution was also displayed in the confusion matrix. Table 3 shows the results of the BERT Artificial Neural Network.

Table 3. BERT Artificial Neural Network Test Results.

	Precision	Recall	F1-score
Benign	0.98	0.97	0.98
Malignant	0.97	0.98	0.98
Accuracy			0.98
Macro avg.	0.98	0.98	0.98
Weighted avg.	0.98	0.98	0.98

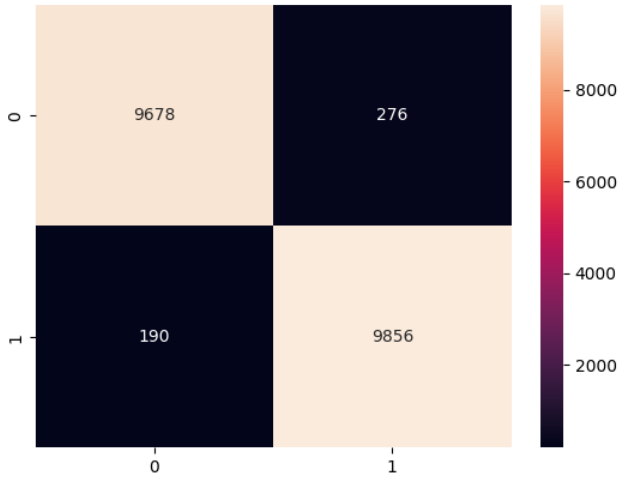

Figure 2. Confusion Matrix Obtained from Test Data İn The BERT Model.

The hyperparameters selected for CNN are ADAM optimization algorithm, Binary crossentropy loss function, epoch number 3, batch size 16. The results obtained from CNN training are given in Figure 3 and Table 4.

Table 4. CNN Test Results.

	Precision	Recall	F1-score
Benign	0.94	0.87	0.90
Malignant	0.88	0.94	0.91
Macro avg.	0.91	0.91	0.91
Weighted avg.	0.91	0.91	0.91

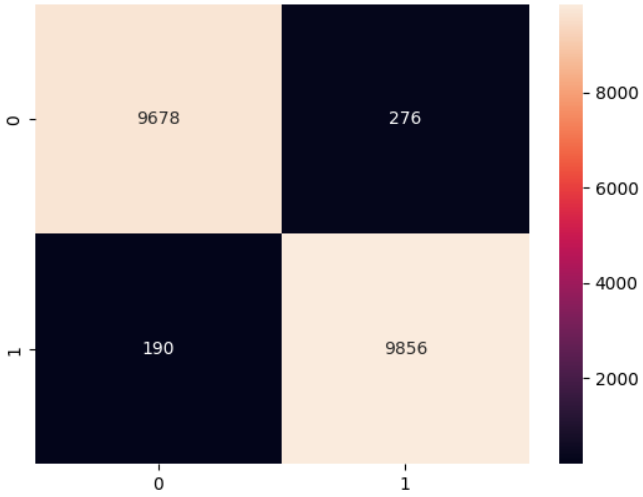

Figure 3. Confusion Matrix Obtained From Test Data İn CNN Network.

IV. DISCUSSION

To evaluate our work, Table 5 provides comparisons of the studies in the literature with the model we developed. It has been observed that the most successful systems in the studies are deep learning systems. However, it is emphasized that continuous research should be done to develop faster and more effective systems. Today, many

different methods and techniques are being researched to increase the performance of deep learning systems and make them more efficient. Since none of these studies are used by the end user and the results are not shared by companies if they are adapted to the systems, the research remains theoretical. These efforts should not remain only theoretical but should be supported by practical applications. The lack of real-world applications prevents the potential of the developed systems from being fully evaluated. For this reason, it is of great importance that the research does not remain only at an academic level but is applied to real-world problems in cooperation with the industry. Companies sharing the results they obtain by implementing such systems will contribute to the further development of research and the production of more effective solutions.

Table 5. Comparison of the obtained results with the studies in the literature.

	Dataset Size	**Type of Training**	**Model**	**Accuracy**
Jain et al., 2018	10000	Machine Learning	Naive Bayes (NB), SVM	NB %64.74, SVM %76.04
	25000			NB %76.87, SVM %91.28
Khushi et al.,2023	1779	Deep Learning	BERT	%96,00
Jawade et al.,2021	36707	Deep Learning	FASTAI CNN	%98,92
Parekh et al., 2018	(Unspecified)	Machine Learning	Random Forests	%95,00
Semih & Busra, 2024	200000	Deep Learning	BERT, DistilBERT (DB)	BERT %98.00, DB %93.00
	20000		CNN	%91,00

V. RESULTS

As a result of the study, three neural networks were trained for URL classification. The results obtained during this training process were compared with the existing literature and evaluated. The findings clearly show that the transformer architecture has a great promise, especially in sequence classification tasks. The high accuracy rates and speed provided by the transformer architecture reveal the potential of this technology. Artificial neural networks, whose success rate has been increased to over 99% and accelerated, can be embedded into real systems in the future. Artificial neural networks with accelerated and high success rates to be integrated into real systems will make great contributions to the prevention of phishing attacks.

DECLARATIONS

Acknowledgements: Any person and/or institution can be acknowledged in this section.

Author Contributions: All work was done by B.T and S. G.

Conflict of Interest Statement: Author declares no conflict of interest.

Supporting/Supporting Organizations: This research has not received any external funding.

Ethical Approval and Participant Approval: This article does not contain any studies on human or animal subjects. Scientific and ethical principles were followed during the preparation of this study and all studies used are given in the references.

Plagiarism Statement: This article was scanned with a plagiarism program. No plagiarism was detected.

Availability of Data and Materials: Data sharing is not valid.

Use of AI Tools: The Author declare that they did not use Artificial Intelligence (AI) tools in the creation of this article.

REFERENCES

Chollet, F., & others. (n.d.). Keras. Retrieved May 6, 2024, from https://keras.io

Federal Bureau of Investigation, 2021. Internet Crime Report 2021. Retrieved May 10, 2024, from www.ic3.gov.tr

International Telecommunication Union., 2023. Statistics. Retrieved May 11, 2024, from https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx

Jain, A. K., & Gupta, B. B. ,2018. PHISH-SAFE: URL features-based phishing detection system using machine learning. In Advances in Intelligent Systems and Computing (Vol. 729, pp. 467–474). https://doi.org/10.1007/978-981-10-8536-9_44

Jawade, J. V., & Ghosh, S. N., 2021. Phishing Website Detection Using Fast.ai library. Proceedings - International Conference on Communication, Information and Computing Technology, ICCICT 2021. https://doi.org/10.1109/ICCICT50803.2021.9510059

Mittal, K., Gill, K. S., Chauhan, R., Singh, M., & Banerjee, D., 2023. Detection of Phishing Domain Using Logistic Regression Technique and Feature Extraction Using BERT Classification Model. 2023 3rd International Conference on Smart Generation Computing, Communication and Networking, SMART GENCON 2023. https://doi.org/10.1109/SMARTGENCON60755.2023.10442975

PhishTank. (n.d.). What's PhishTank. Retrieved May 4, 2024, from https://phishtank.org/faq.php#whatisphishtank

Siddharth Kumar. (2019). Malicious And Benign URLs. Retrieved May 4, 2024, from https://www.kaggle.com/datasets/siddharthkumar25/malicious-and-benign-urls