Detecting Phishing URLs using the BERT Transformer Model

Denish Omondi Otieno*, Faranak Abri†, Akbar Siami Namin* and Keith S. Jones*
*Texas Tech University, †San Jose State University
Email: *deotieno@ttu.edu, †faranak.abri@sjsu.edu, *akbar.namin@ttu.edu, *keith.s.jones@ttu.edu

***Abstract*—Phishing websites many a times look-alike to benign websites with the objective being to lure unsuspecting users to visit them. The visits at times may be driven through links in phishing emails, links from web pages as well as web search results. Although the precise motivations behind phishing websites may differ the common denominator lies in the fact that unsuspecting users are mostly required to take some action e.g., clicking on a desired Uniform Resource Locator (URL). To accurately identify phishing websites, the cybersecurity community has relied on a variety of approaches including blacklisting, heuristic techniques as well as content-based approaches among others. The identification techniques are every so often enhanced using an array of methods i.e., honeypots, features recognitions, manual reporting, web-crawlers among others. Nevertheless, a number of phishing websites still escape detection either because they are not blacklisted, are too recent or were incorrectly evaluated. It is therefore imperative to enhance solutions that could mitigate phishing websites threats. In this study, the effectiveness of the Bidirectional Encoder Representations from Transformers (BERT) is investigated as a possible tool for detecting phishing URLs. The experimental results detail that the BERT transformer model achieves acceptable prediction results without requiring advanced URLs feature selection techniques or the involvement of a domain specialist.**

***Index Terms*—Social Engineering, Phishing URLs, BERT Transformer Language Model.**

I. Introduction

Driven by events no one could have foreseen the post-pandemic rapid digitization makes cybersecurity everyone's business. The future promises more connected systems, out of the office to remote workplaces, into sophisticated digital clouds, distributing more data and consequently more organized adversaries. Phishing websites pose a challenge to governments, organizations as well as to unsuspecting individuals because of their close resemblance to benign web pages [1]. In addition, judging the safety of web pages is something that even cybersecurity experts struggle to do accurately without additional information [2]. With the bulk of the pandemic season disruptions behind us, the society and by extension the cybersecurity community is feeling a sense of calm, composure and confidence in returning to normalcy. However, the feeling is evidently quickly vanishing particularly in the information technology security posture and being replaced by elevated levels of new and devastating cyber-risks concerns [3].

Positively, the global widespread post-pandemic adoption of the World Wide Web, as a leading information retrieval service of the internet, supporting knowledge dissemination [4] has brought about monumental improvements and developments in communications, e-commerce as well as along chains of supply, that are almost completely digital to list a few. The World Wide Web's global popularity has not only enhanced services, as well as increased economic activities but has also pushed the cybersecurity community beyond its comfort zone by presenting cyber-criminals with new avenues to dupe unsuspecting users [4]. Correspondingly, each new digital development comes with new cyber-risks [4], leaving the cybersecurity community with more work to do and in tough economic environments.

A look into the post-pandemic and beyond, the cybersecurity community can expect a return to a harsher reality with ever-increasing cyber-risks [3]. The 2023 Verizon Data Breach Investigation Report (DBIR) [5], acknowledges social engineering attacks to be often very effective and extremely lucrative for attackers, with stolen credentials, phishing, together with the exploitation of vulnerabilities to be three ways in which cyber-criminals access privileged information. The report [5], further determines an increase in business email compromises i.e., pretexting attacks which it notes at 50% of the incidents within the social engineering space. The human element is documented to be the weakest security link in computer security systems [6], [7], [8] and the 2023 Verizon Data Breach Investigation Report [5], similarly observes that 74% of all breaches include aspects of the human element, with people being involved either via error, use of stolen credentials, misuse of privileges as well as social engineering [5].

The 2023 Voice of the CISO ProofPoint report [3], considers that cyber-attackers look set to wreak more havoc as it envisions email fraud at 33%, insider threats at 30%, cloud account compromise at 29%, DDos attacks at 29%, supply chain disruptions at 27%, ransomware attacks at 27%, smishing texts/vishing calls at 27% as well as malware dissemination at 26% to be cyber-risks of major concerns across the security posture [3]. The ProofPoint report [3], additionally notes that hybrid work setups are evolving into a mainstay for most, with data extortion becoming the rule rather than the exception [3]. Moreover, the insights into CISO challenges, expectations, as well as priorities report [3] establish that the increasing commercialization of the dark-web exploit tools, initial-access brokers and "as-a-service" attack infrastructures, to be increasingly threatening if not yet, to make cyber-attacks more open to anyone with ill intents [3].

Armed with enough time, motivation and unchained ethical constraints, attackers are leveraging the capabilities of the World Wide Web to infiltrate networks, comprise digital clouds, as well as getaway with privileged data from institutions of every size, across every industry without sparing any jurisdiction [3]. A well-known and effective scheme attackers use is baiting unsuspecting users into clicking on compromised URLs [4]. Such threats to digital peace among others [3], [5], [8], to list a few, highlight the urgent need for improved cyber-defenses.

Necessarily, cybersecurity concerns are at the forefront of research organizations, companies and governments alike [4]. Consequently, significant progress has been made to help mitigate phishing risks [9] including plug-ins and extensions, comprehensive blacklists, whitelists, machine learning algorithms as well as anti-phishing games [9], [10]. However, after all these efforts, the cybersecurity community observes that there is no anti-phishing solution considered as optimum [9]. Further, it is also rather easy for attackers to deceive the URLs blacklisting services by slightly modifying a few components of the URL string [4], as well as some researchers hold that there is no static cybersecurity technical defense approach that can solely mitigate threats introduced by user behavior [9].

Amid the growing concerns around cybersecurity, understanding and mitigating attacks, as well as navigating the threat landscape still remains a matter of protecting people, systems and data. This study, therefore critically examines phishing URLs along with benign ones and seeks to contribute to the field of knowledge by:

1) Focusing on analyzing phishing websites not by their contents but rather by their URLs, to extract, visualize as well as correlate URL features in a bid to uncover elements, attributes, characteristics and trends along with distinct distinguishable patterns between phishing and benign URLs;
2) Additionally, the study, seeks to apply the self-attention-based Bidirectional Encoder Representations from Transformers, (BERT) to investigate its capabilities in detecting phishing websites by use of their URLs.

The rest of the paper is structured as follows. Section II, reviews the literature detailing the different techniques explored in phishing mitigation strategies. Section III, presents a brief background of the technical concepts utilized in this study as Section IV, establishes the methodology of the study while the experimental setup is presented in Section V. Further, Section VI, reports and elaborates on the findings as Section VII, concludes and sketches the future works.

II. Related Work

The research community is and has conducted large-scale studies on URLs detection as well as proposed different mitigation strategies. From blacklisting services to lists-based, anti-phishing games, visual similarity techniques, content based models as well heuristics learning procedures among others [1], [4], [9], [10] symbolise the combined intense efforts to address phishing.

Le et al. [1] suggest, it is imperative to detect malicious URLs in a precise and timely manner. Le et al. [1] observe that traditional malicious URLs detection methods such as blacklisting may not be exhaustive and might not detect newly generated malicious URLs. Le et al. [1] further, witness several machine learning malicious URLs detection strategies, some using lexical properties of the URL string by extracting Bag-of-words like features, followed by applying machine learning models such as support vector machine (SVMs). However, Le et al. [1] note that, a number of malicious URLs detection techniques require substantial manual feature engineering. As well as others suffer from the inability to effectively capture semantic meaning and sequential patterns in URL strings [1]. Le et al. [1] moreover, contribute to the science of addressing malicious URLs by putting forward an end-to-end deep learning framework to learn a nonlinear URL embedding for malicious URL detection directly from the URL.

Vanhoenshoven et al. [4] address the detection of malicious URLs using several machine learning techniques i.e. Naıve Bayes, Support Vector Machines, Multi-Layer Perceptron, Decision Trees, Random Forest along with K-Nearest Neighbors and their numerical simulations results suggest that the classification methods achieve acceptable prediction rates and in particular, Random Forest and Multi-Layer Perceptron attain higher accuracies [4].

Inspired by the evolving nature of the phishing websites, Chatterjee and Namin [11], introduce a deep reinforcement learning approach, capable of adapting to dynamic behaviors of phishing websites to learn features associated with the phishing web pages for enhanced detection [11]. Chatterjee and Namin [11] model the pinpointing of phishing web pages through Reinforcement Learning (RL), where Chatterjee and Namin [11], detail that an agent learns the value function of a given URL to perform classification tasks. Chatterjee and Namin [11] additionally, map the sequential decision making process using a deep neural network based implementation of RL. Chatterjee and Namin [11] further, observe the average of the relevance measures from different runs of their proposed model to be Precision of 0.867, Recall of 0.88, Accuracy of 0.901 and an F-Measure of 0.873.

Yamoun et al. [12], present RoBERTa transformer a variant of BERT, in its pre-trained version without fine-tuning vs Term Frequency Inverse Document Frequency, (TF-IDF) for websites content-based classification. Yamoun et al. [12] study two approaches, a mono-multi classification into 16 classes, as well as different binary classification tasks for each of the 16 classes following the one vs. all strategy. Yamoun et al. [12] suggest better results with RoBERTa embeddings compared to TF-IDF features observing an Accuracy of 68% for the mono-multi-classification, along with an Average of 90.69% for binary classifications. Yamoun et al. [12] further, view that for the mono-multi classification results, RoBERTa and TF-IDF embeddings are practically the same at 68% vs 68.2%. However, Yamoun et al. [12] maintain that RoBERTa embeddings distinctly outperform TF-IDF features in binary classifications with a difference in Accuracy of +5.41%.

Praksh et al. [13] suggests a predictive blacklisting technique for identifying phishing attacks. Praksh et al. [13] as well notice that while URL blacklisting has been effective to some degree its reliance on the exact match features with a blacklisted entry might make it easy for cyber attackers to evade it by employing simple modifications to the URL. Praksh et al. [13] moreover, propose five heuristics, to enumerate simple combinations of known phishing pages, to discover new phishing web pages URLs along with an approximation matching algorithm that dissects a URL into multiple base components that are further matched individually and specifically against entries in a blacklist. Praksh et al. [13] assess that their approximation matching algorithm leads to few false positives as well as negatives at 3% and 5% respectively. Besides Praksh et al. [13] in their evaluations with real-time blacklist feeds, discover around 18,000 new phishing URLs from a set of 6,000 new blacklist entries.

Ma et al. [14] explore the use of statistical methods from machine learning in classifying site reputation based on the relationship between URLs along with the lexical and host-based features that characterize them. Ma et al. [14] with experimental results of Naive Bayes, Support Vector Machine (SVM), and Logistic Regression report (95–99)% Accuracy from the classifiers, detecting malicious web pages from their URLs, with modest false positives.

Apart from relying on different phishing websites identification techniques i.e., honeypots, features recognitions, manual reporting and web crawlers among others. Mohammad et al. [15] for instance, critically explore legal proposals as well as educational processes as countermeasures to curb phishing. Although, Mohammad et al. [15] find useful proposals they however, argue that law enforcement has the downside of time, as phishing websites have a short average time to live while phishing attacks are performed swiftly with the immediate possibility of the attacker/s disappearing into cyberspace as quickly as possible after the attack. Mohammad et al. [15] further, note that education is an effective countermeasure technique however, they [15] present that eliminating phishing via education is a long-winded process where users need to dedicate a great amount of time to study and understand phishing, as well as attacks are becoming so sophisticated to the level that cybersecurity experts are also being deceived.

The phishing attacks countermeasures expand far and wide and Sheng et al. [10] argue that games can be an effective way of educating people about phishing along with their related cyber-risks. Sheng et al. [10] describe a science learning, principle-based, iteratively refined online game that teaches users good cyber hygiene practices to help them avoid phishing attacks. Sheng et al [10] evaluate the online game by testing the ability of the game participants in identifying fraudulent web pages before as well as after spending 15 minutes interacting with one of the three Sheng et al. [10] suggested anti-phishing training activities, i.e., engaged playing the game, studying an anti-phishing tutorial developed based on the game as well as reading existing online training messages/materials. Sheng et al. [10] seek to contribute to the phishing mitigation knowledge by affirming that participants who played the content interactive game proved to be better at identifying fraudulent websites in relation to others. On the other hand, Luga et al. [9] in looking at client-side defenses write that user passwords can be enhanced by post-password actions i.e., technical methods applied after a user enters a correct password to mitigate situations where passwords mistakenly fall into attackers hands.

The review of literature speaks towards a constant effort by the cybersecurity research community to address the issue of phishing attacks from all possible angles. Maneriker et al. [16] additionally, remark that earlier phishing detection relied on the use of standard machine learning classifiers. However, recent research is instead proposing the use of deep learning models for phishing URLs detection tasks [16]. Concurrently, Maneriker et al. [16] point out that text embedding investigations using transformers are leading to state-of-the-art results in many Natural Language Processing (NLP) tasks and this study, seeks to advance this line of science by investigating the effectiveness of the Bidirectional Encoder Representations from Transformers, (BERT) in detecting phishing websites by use of their URLs.

III. Technical Background

This section briefly elaborates on some of the technical concepts as well as the key techniques employed in this research.

A. Transformer

Vaswani et al. [17] introduce a novel architecture designated the Transformer, based on attention mechanisms, dispensing with recurrence and convolutions. The Transformer, a model architecture, relies on self-attention mechanism to draw global dependencies between input and output [17], while allowing for significant parallelization [17]. Self-attention enables the Transformer to detect connections between different elements and assess the importance of the connections [17]. The novel network architecture (Transformer) [17], allows for simultaneous sequence processing where model training can be sped up through parallelization due to positional embeddings and Multi-head attention.

Vaswani et al. further [17], document that Transformers are more parallelizable and require significantly less time to train. A closer look into the less time to train aspect shows that word embeddings transform inputs into vector representations as positional embeddings encode the position of each token in a sequence [17]. The sequence-to-sequence element (Seq2Seq) of Transformers transforms input sequence of vectors into output sequence through a series of encoder and decoder layers [17], as Attention and Self-attention aspects of the Transformer work with the sequence that is being encoded (Self-attention) while Attention assists in understanding and analyzing different parts of a sequence being generated by the decoder [17]. Research [16] and [17] acknowledge, Transformers to be efficient in NLP tasks, and the layers of the Transformer encoder and decoder contain fully connected feed-forward

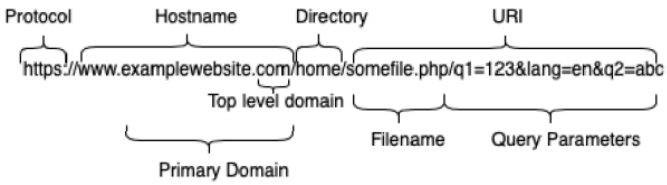


Fig. 1: The Structure of a URL as depicted by [11].

network [17], as Multi-head attention allows for the learning of multiple ways of weighing input sequences [17].

B. Bidirectional Encoder Representations from Transformers

Scientific studies [16], [17] affirm that Transformer learning is and has revolutionized the field of NLP. Devlin et al. [18] introduce a language representation model, The Bidirectional Encoder Representations from Transformers (BERT) which uses encoders in a transformer as a sub-structure to pre-training models for NLP tasks. BERT's model architecture is a multi-layer bidirectional Transformer encoder a type of neural network architecture introduced by Vaswani et al. [17]. BERT [18], applies the two-way Transformer Encoder to learn the information before and after each word to obtain better word vector representation. BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [18]. Each encoder layer in BERT further consists of two sub-layers, a feed-forward neural network and self-attention mechanism. The feed-forward neural network allows BERT to capture complex patterns and associations between words by applying the non-linear transformations to the output of the self-attention while self-attention permits BERT to weigh the importance of words based on their relationship to other words. Devlin et al. [18] suggest that fine-tuning introduces minimal task-specific parameters and some of the advantages of BERT is that the model does not need to be retrained, rather it can be fine-tuned with one additional output layer to create state-of-the-art models for a wide range of tasks without substantial tasks-specific architecture modifications [18]. The bi-directionality capability of BERT moreover, allows it to read input from either right to left or left to right, simultaneously. Devlin et al. further [18], affirm that BERT is conceptually simple and empirically powerful thus in this study the research investigates BERTs capabilities in detecting phishing web pages by use of their URLs.

C. The Structure of Uniform Resource Locators (URLs)

Chatterjee and Namin [11] explain that a typical URL has two fundamental parts 1) The Protocol, and 2) The Resource Identifier, where Chatterjee and Namin [11] additionally, present that the Protocol specifies the Uniform Resource Locator to be used in communication between a user and the web server. While the Resource identifier indicates the internet protocol (IP) address or the domain space where a resource is located. Chatterjee and Namin [11] further, stipulate that a colon as well as two forward slashes separate the protocol from the resource identifier as depicted in Figure 1.

Mamun et al. [19] point out that URL lexical features are the textual properties of a URL such as URL length, length of the hostname as well as tokens found in the URL, etc. A further look at the structure of Figure 1, reveals several possible segments of a URL such as Hostname, Primary domain as well as Top level domain, Directory, Filename, Query parameters and the Uniform Resource Identifier (URI). Encouragingly, Zeng et al. [20] describe that with the assistance of the Domain Name System (DNS) normal users only need to query the relevant domain names and no longer have to explicitly deal with IP addresses while surfing the web. However, for the unsuspecting user not aware of the intricacies of web technologies, attackers can obfuscate a URL such that the actual domain name might not be easily identifiable as it's nested deep inside a URL. Zeng et al. [20] further, discuss domain-squatting, a premeditated attempt by attackers to register perceptively confusing domain names thereby tricking unsuspecting users into querying them, for example, a name of a renowned brand might be "greatproducts.com", the attacker consequently might as well register "greatproducts.org", "greatproducts.biz" or "greatproducts.net", etc. in order to confuse the unsuspecting users. Additionally, Zeng et al. [20] moreover, detail five squatting types namely typo-squatting, bit-squatting, homograph-squatting, sound-squatting, and combo-squatting.

IV. Methodology

This section describes the methodology applied in this study. Figure 2, conveys the steps employed in the study and they are: Data Pre-Processing I and II, Feature Engineering, Primary Domain Text Analysis as well as Classification using the Bidirectional Encoder Representations from Transformers.

1) *Data Pre-Processing I and II.* Models output analyses are only as good as the data they're based on. Data pre-processing is an important preliminary step that might refer to the process of cleaning, transforming, as well as the integration of data to make it ready for analysis. To enhance the quality of data so as to promote the extraction of meaningful insights, this study, categories Data pre-processing into Data Pre-Processing I and Data Pre-Processing II, with the aim of identifying and fixing possible errors. In Data Pre-Processing I, the study, performs data cleaning to remove the (NaNs) i.e., null values, as well as it drops all the non URLs based columns. However, it does not remove the punctuations or delete the special characters as they form an integral part of feature difference analysis between phishing and benign URLs. The study, moreover, further performs data balancing as part of Data Pre-Processing I to try and prevent the experiments model from becoming biased towards one class. In Data Pre-Processing II, the study, applies data labeling to clearly class the URLs into phishing and benign, while data shuffling allows the study to shuffle the data before making a split/division so that each split/division has an accurate representation of the dataset.
2) *Feature Engineering.* Feature Engineering is a flexible way of extracting features from documents. Features extracted

from URLs might represent or potentially manifest the concern under investigation, i.e., variations in URLs features might represent the occurrence of a phishing or benign URL. This study employs URLs feature extraction to extract lexical and host-based features of the URLs. The motivation for analyzing the features is based on the fact that phishing URLs pose serious cybersecurity concerns in a space where data privacy remains mission-critical as privacy continues to increase in importance, in institutions, for organizations, as well as to individuals and therefore the study extracts features of the URLs to try and quantify the differences between phishing and benign URLs. The host-based features obtained from the host-name properties of URLs might as well allow for analysis of 'who', 'where', 'when', and 'how a website is hosted. This study's incentive behind the extraction of host-based features is that there might be a significant difference between phishing and benign website deployment tactics, as well as reputations.

3) *Primary Domain Text Analysis.* Data visualizations (i.e., graphs, charts, infographics etc.) are valuable ways of communicating information at a glance, but what if the given data is text-based? As such using a word cloud might highlight important textual data points. Word-Clouds have been widely used to present themes and contents in texts either for summary as well as for visualizations. Word-Clouds are visual representations of text data in the form of tags, which are typically single words whose importance is visualized by way of their size as well as color. Word-Clouds or tag clouds give greater prominence to words that appear more frequently in a source text. The bigger, as well as the bolder, the word appears, the more often it's used within a given textual data. This study visualizes the primary domains of the phishing as well as the benign URLs to examine if there exists distinguishable important textual points from the primary domains of the two types of URLs.
4) *The Transformer Model and Evaluation.* A model is mainly at the center of machine learning or artificial intelligence applications. Models can identify patterns as well as relationships in the given data with a level of accuracy, speed and sophistication that humans cannot match. Models instructions are induced from a set of data and based on probabilistic assessments, models can be used to prescribe an action, make predictions as well as recommendations. This study, accordingly, leverages on the ability of the Bidirectional Encoder Representations from Transformers to binary classify URLs as either phishing or benign. Evaluating a model plays a vital role in giving insights into the performance of a model. Model evaluation uses metrics to help analyze the outputs/patterns or predictions of a model. Metrics such as Confusion Matrix, Accuracy, Precision, Recall, F1 score, as well as the Area under Curve (AUC) are some of the elements of model evaluation looked at in this study.

V. Experimental Setup

To carry out the research objectively and efficiently the study employs:

1) The URL dataset (ISCX-URL2016) from The Canadian Institute for Cybersecurity [21]; and
2) Phishing URLs from PhishTank [22], a collaborative clearing house for data and information about phishing on the Internet [22].

Google Colab, a Google research product is adopted as the preferred experiments environment. Python libraries such as Pandas, a fast powerful, data analysis and manipulation tool. NumPy i.e., (Numerical Python) together with their correlate components, along with a plotting library as well as the urllib.parse a module that defines standard interface to break URLs strings into components are some of the packages coupled with their related modules leveraged in this study.

A. Data Sets

The URL dataset (ISCX-URL2016) by The Canadian Institute for Cybersecurity [21] and as further elaborated by [19] consists of over 35,300 benign URLs, around 12,000 spam URLs, about 10,000 phishing URLs and more than 11,500 URLs related to malicious websites as well as over 45,450 URLs belonging to Defacement URL category. This study randomly selects 35,000 benign URLs and a further total of 9,964 phishing URLs from The Canadian Institute for Cybersecurity (ISCX-URL2016) dataset [21]. In its bid to balance the phishing URLs the study randomly extracts 25,036 phishing URLs from the PhishTank dataset retrieved on the 11th day of November 2022. A preliminary data preparation drops all the columns of the PhishTank dataset [22] except the URL column. The study subsequently merges the 9,964 phishing URLs from The Canadian Institute for Cybersecurity (ISCX-URL2016) dataset [21] with the 25,036 phishing URLs extracts from the PhishTank dataset [22] to achieve 35,000 phishing URLs.

B. Data Preparation

Data labeling denotes the process of identifying items of raw data and subsequently adding one or more labels to them to specify their contexts to the machine learning/deep learning model/s. The randomly extracted 35,000 benign URLs from The Canadian Institute for Cybersecurity (ISCX-URL2016) dataset [21] are thus labeled as (0) while the output of merging the 9,964 phishing URLs from The Canadian Institute for Cybersecurity (ISCX-URL2016) dataset [21] with the randomly extracted 25,036 phishing URLs from the PhishTank dataset are grouped with the label (1). In a bid to avoid over-fitting the study maximizes to achieve a balanced dataset. The randomly extracted 35,000 benign URLs as well as the earlier merged 35,000 phishing URLs are likewise further consolidated to create a new data frame. Data shuffling in this study moreover, serves to randomize the order in which the data is presented to the Bidirectional Encoder Representations from Transformers. To further explore the selected datasets, the study develops Python scripts in which it utilizes different select libraries

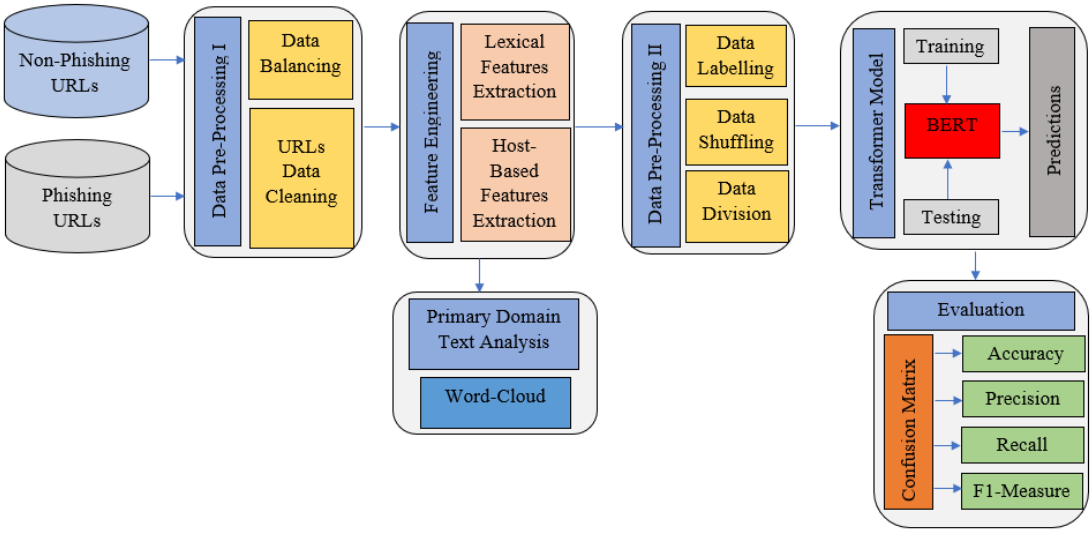


Fig. 2: The flowchart of methodology.

and their related modules for its analysis, plotting, as well as manipulations. An analysis of the merged new data frame reveals 70,000 URLs for this study.

C. Classification

URLs classification might be done by providing the classification of the URL string regardless of the page's content as well as classifying the page and providing the classification specific for that page based on its contents. To leverage on the ability of the Bidirectional Encoder Representations from Transformers in its experiments. This study, splits the balanced dataset into 75-25 ratio with (75%) being training and (25%) testing, for evenness of the data the study applies stratification "stratify" with the aim being to ensure the same ratio of both categories are loaded for each case in an attempt to prevent over-fitting.

VI. Results and Analysis

This section reports the results of this study.

1) URLs Features Comparison: Chatterjee and Namin [11] argue that there are certain characteristics of a URL that might help in distinguishing phishing URLs from benign URLs. For example, Chatterjee and Namin [11] list: long URLs as well as the presence of an IP address in a URL to be probable indicating features of a suspect URL. This study likewise proposed to investigate a set of select features between the phishing and benign URLs namely:

1) The Length of the URLs.
2) Length of the Hostnames.
3) Length of the Paths.
4) Primary Domains.
5) Count of Alphabets.
6) Count of Digits.
7) Count of Punctuations.
8) The use of IP.
9) Use of URL shortening services and
10) A look at HTTP as well as HTTPs Protocols.

TABLE I: Mean-Values of selected URLs features.

#	Feature	Benign	Phishing
1	URL-Length	115.009314	61.796600
2	Hostname-Length	11.808629	31.501943
3	Path-Length	74.950600	18.066629
4	Alphabets	76.970943	47.212629
5	Digits	15.085229	3.837486
6	Punctuations	22.953143	10.746486

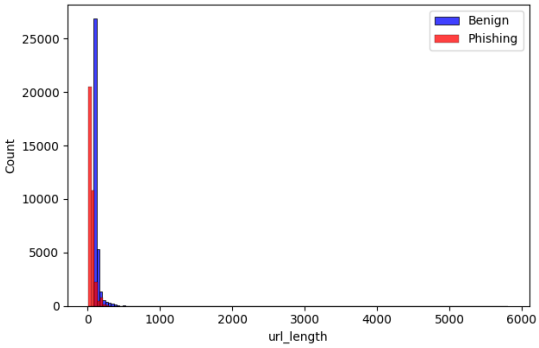
The experimental results suggest that when it comes to the length features as demonstrated in Figure 3. On average benign URLs are longer than phishing URLs, Figure 3a. While phishing URLs at most depict longer hostname-lengths in comparison to benign URLs, Figure 3b as well as the path-lengths of the benign and phishing URLs record significant difference with benign URLs having longer path-length compared to phishing URLs, Figure 3c. Table I, # 1, # 2, and # 3, report similar observations on the mean-values of the URLs lengths, hostname-lengths as well as the path-lengths.

The study further sets to investigate the characters in a URL to find out their effectiveness in differentiating phishing URLs from benign URLs. The study looks at the distribution of alphabets, Figure 4a along with Table I, # 4, and observes that the experimental results suggests that as per the mean-values Table I, # 4, benign URLs have more alphabetical characters than phishing URLs, an observation captured by Figure 4a as well. The trend is closely observed in Figure 4b along with Table I, # 5 as well as in Figure 4c along with Table I, # 6. Where the results suggest that benign URLs on average, Table I, # 5 have more numerical characters as well as punctuations, Table I, # 6, compared to phishing URLs.

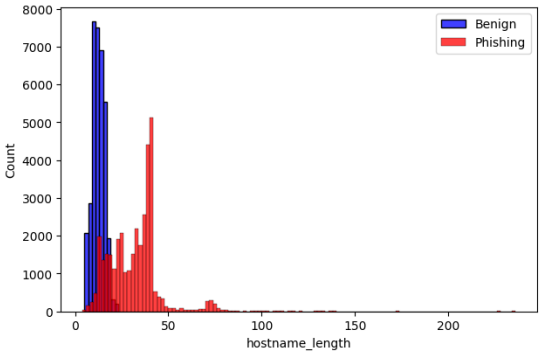
The study accepts that a comparison of length features along with alphabets, digits as well as a look at the punctuations might not offer conclusive distinctive difference between phishing and benign URLs. Consequently, the study suggests taking a further examination at the use of URLs shortening services, as well as investigating the presence of an IP address in a URL among the URLs. As depicted in Table II, the study,

TABLE II: Use of Shortening Service as well as IP.

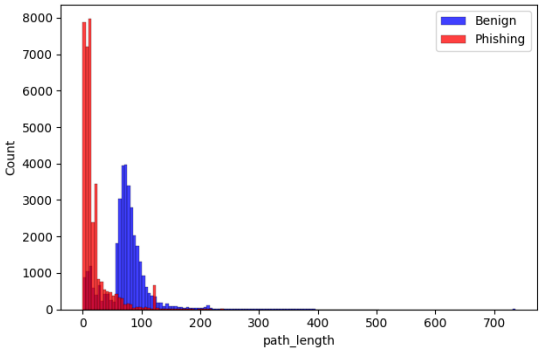
#	Layout	Benign	Phishing
1	Shortened	0.069657	0.171657
2	Use-of-IP	0.001714	0.006029



(a) URL-Length.



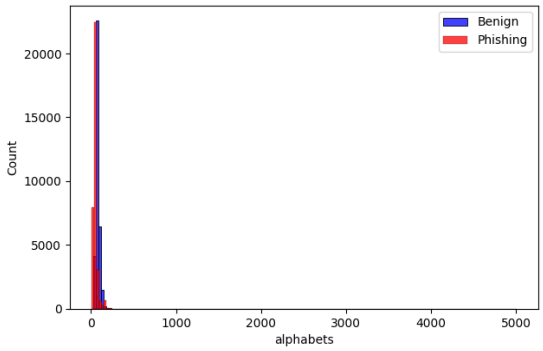
(b) Host-Name Length.



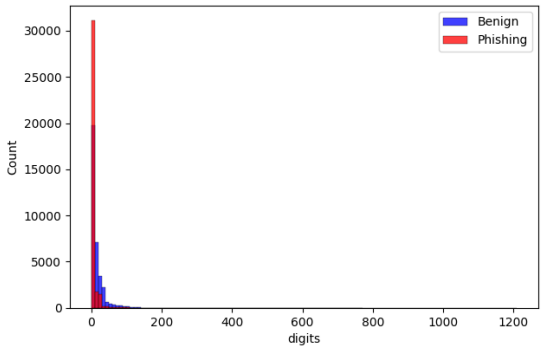
(c) Path-Length.

Fig. 3: Length Features.

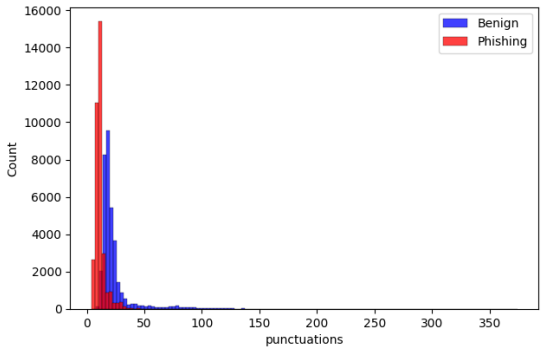
observes that the experimental results imply that a number of phishing URLs utilize the URLs shortening services as compared to benign URLs, as well as the use of IP address in the URL is prevalent in phishing URLs than in benign URLs. However, the study notes that the difference between the use of shortening services along with the presence of an IP address in a URL might not be significant enough to warrant a distinctive classification between phishing and benign URLs but rather offers valuable insights in relation to the use of shortening services, as well as the presence of an IP address in a URL as pointed out in Table II.



(a) Alphabets.



(b) Digits.



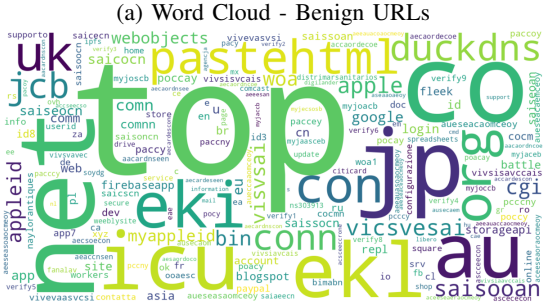
(c) Punctuations.

Fig. 4: Alphabets, Digits and Punctuations.

2) Word-Clouds: Word-Clouds can be applied to give an intuitive and visually appealing overview of a text by distilling text down to those words that appear with the highest frequency [23]. Text summarization is one popular application area for Word-Clouds [23] as such a summarization might be helpful in learning about the number and kind of topics present in a body of text [23].

This study in its quest to detect phishing URLs takes a look into the primary domains of the phishing and benign URLs. Figure 5, helps this study judge whether a given text or group of texts is distinctively relevant to differentiate between phishing and benign URLs. Despite the study observing a difference between phishing and benign URLs primary domains texts summarizations (Word-Clouds) it maintains that the difference might not be conclusive enough to warrant a binary classification between phishing and benign URLs.

(a) Word Cloud - Benign URLs



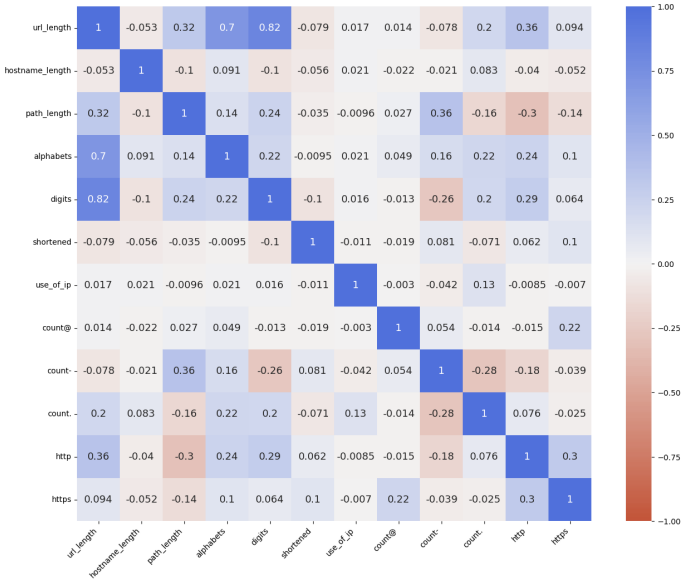
(b) Word Cloud - Phishing URLs

Fig. 5: Benign and Phishing URLs Word-Clouds.

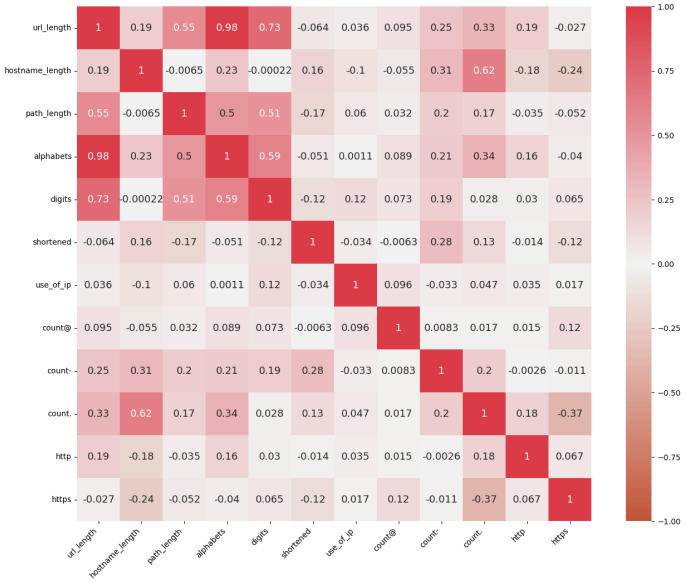
However, as Heimerl et al [23] points out that Word-Clouds generated for a body of text can serve as starting points for a deeper analysis, this study subsequently suggests using the differences of text summarizations in Figure 5a and Figure 5b to serve as continuing points for further analysis into the detection of phishing URLs.

3) Spearman Features Correlation Matrixes: Correlation depicts the strength of a relationship between two variables. It can be explained as a statistic that measures the degree to which two variables move in relation to each other. Typically a correlation matrix is ”square” with the same variables visible in the rows and columns. A correlation matrix is useful for displaying the pairwise correlation coefficient values. This study, seeks to find if a correlation exists between URLs features and moreover, aims to examine if a correlation pattern difference is observable between phishing and benign URLs. The features of interest at this point of the experiment include: URL-length, hostname-length, path-length, alphabets, digits, shortened (use of shortening services), use-of-IP, Http along with Https. While the punctuations on focus at this point include the count of @, count of - and the count of ., as shown in Figure 6.

Figure 6a and Figure 6b, both observe a systematic pattern of 1.00s from their top left to their bottom right depicting that each variable perfectly correlates with itself. Figure 6a, points out a significant positive correlation between URL-lengths and digits as well as between URL-lengths and alphabets. While, Figure 6b, similarly depicts a significant positive correlation between URL-lengths and alphabets, URL-lengths and digits as well as URL-lengths and path-length. Figure 6b, moreover, notes a positive correlation between count of . and the hostname-length, an observation not replicated in Figure 6a. Further analysis of Figure 6a and Figure 6b suggest that



(a) Features Correlation - Benign-URLs.



(b) Features Correlation - Phishing-URLs.

Fig. 6: Spearman Features Correlation Matrixes (Heat-maps) of Benign vs Phishing URLs.

methods such as heat maps and correlation plots might not be conclusively sufficient in differentiating between phishing and benign URLs.

4) BERT Model and Evaluation: The study, acknowledges that building a machine learning/artificial intelligence or deep learning model works on a constructive feedback principle i.e., build a model, get its feedback from metrics, make possible improvements, and persist until desirable classification accuracy is reported. The evaluation metrics accordingly, assist this study, explain the performance of the Bidirectional Encoder Representations from Transformers as applied in this work.

The study, examines the evaluation metrics of Precision, Recall, Accuracy and the F1-score, as well as it observes a

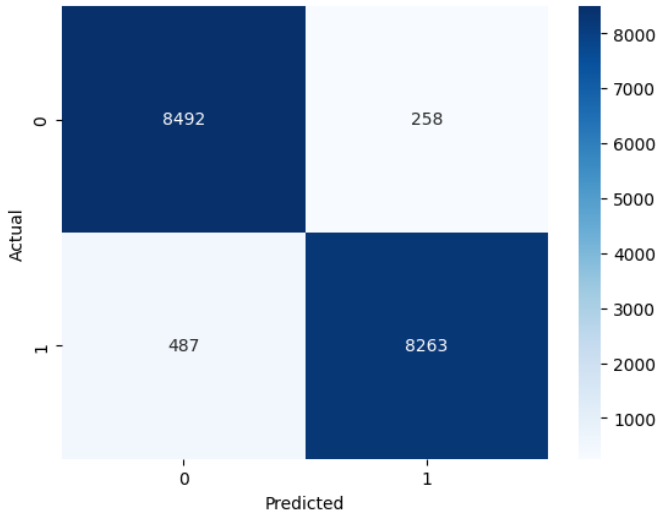
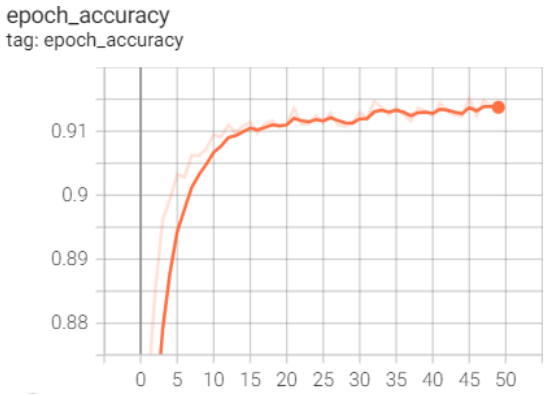


Fig. 7: Confusion Matrix.

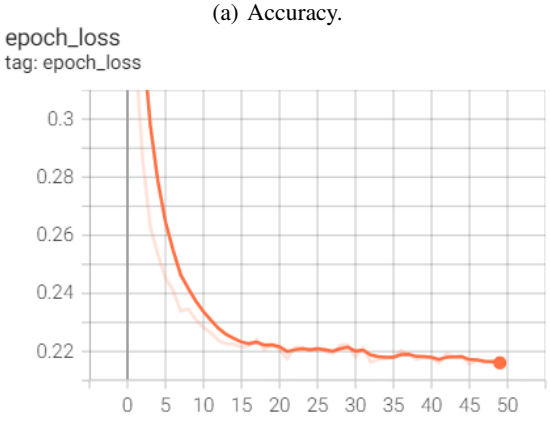
Confusion Matrix. A look at the study's, Confusion Matrix, Figure 7, points out that the experiment yields modest False Positives as well as False Negatives, as classification algorithms or models, predict/classify data into finite sets of classes but since models are not perfect, some data points might be classified incorrectly.

In its training phase, the experiment, trains its model for 50 epochs. Figure 8b, the loss curve, with its X-axis, denoted by values 0 to 50 implying the 50 epochs and the Y-axis recording values 0.22 to 0.3, details the training process and the direction in which the experiments model learned. An analysis of Figure 8b, suggests that the loss of the study's experiment trained model is almost always lower on the training than the validation and possibly continued training could have led to over-fitting. Precision, the ratio between the True Positives and all the Positives, observes a Precision of 0.95 for benign URLs as compared to 0.97 for phishing URLs, as observed in Table III. Recall, the measure of the experiment's model correctly identifying True Positives, reports a Recall of 0.94 for phishing URLs compared to 0.97 for benign URLs as well as the harmonic mean of Precision and Recall i.e., The F1-score, details an F1-Score of 0.96 for both the two types of URLs in question, Table III.

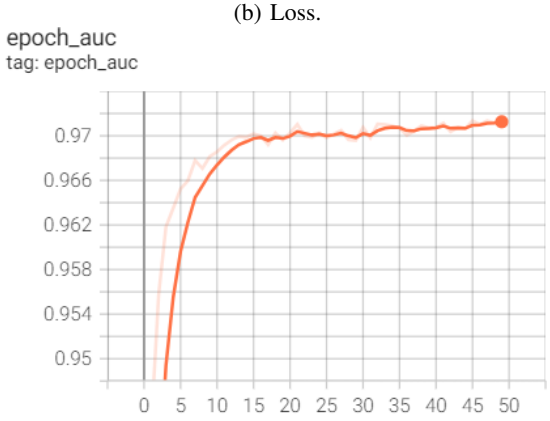
Figure 8c, the area under the curve, created by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings, records values 0.95 to 0.97 in its Y-axis in relation to the 50 epochs accounted for on its X-axis. An examination of the area under the curve for this experiment's model depicts a classifier that can to some degree correctly distinguish between phishing and benign URLs, with Table III, reporting the ratio of the total number of correct predictions and the total number of predictions (Accuracy) to be 0.96. Figure 8a, (Accuracy), with its X-axis recording the 50 epochs and the Y-axis accounting for values from 0.88 to 0.91 tries to elaborate how the experiment's model performance is growing over time, suggesting the model is improving with experience (it's learning) especially at the



(a) Accuracy.



(b) Loss.



(c) Area Under the Curve.

Fig. 8: Model Learning Curves.

TABLE III: BERT Transformer Model Classification.

	Precision	**Recall**	**F1-Score**
Benign URLs	0.95	0.97	0.96
Phishing URLs	0.97	0.94	0.96
Accuracy			0.96

beginning but overtime it reaches a plateau, indicating the model might not be able to learn further.

In trying to further understand the effectiveness of the the Bidirectional Encoder Representations from Transformers as a classification model to help mitigate the phishing menace. The experiment compares its findings with those of the study "The Application of the BERT Transformer Model for Phishing

Email Classification" an investigation by [8], performed on the Body contents of phishing emails rather than the URLs and observes that encouragingly, similarly BERT reports encouraging levels of accuracy at 0.93 as established by [8].

VII. Conclusion and Future Work

This study examines the detection of Phishing URLs with the objective being to realise a classification model that can correctly classify URLs as either phishing or benign. The study's experimental results detail that the Bidirectional Encoder Representations from Transformers achieves acceptable prediction results without requiring advanced URLs feature selections techniques or the involvement of a domain specialist. Additionally, the study, observes in Section II, a good number of machine learning-based classification models as well as several proposed phishing attacks countermeasures that can be adapted to help in mitigating the phishing menace. The study, moreover, seeks to add to this field of knowledge by presenting the possibility of developing and incorporating effective and robust classifiers using machine learning-based Transformers.

Acknowledgment

This research was supported by the U.S. National Science Foundation (Awards#: 2319802 and 2319803) as well as by the U.S. Office of Naval Research (Award#: N00014-21-1-2007). Opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the NSF or the ONR.

References

[1] H. Le, Q. Pham, D. Sahoo, and S. Hoi, "Urlnet: Learning a url representation with deep learning for malicious url detection. arxiv 2018," *arXiv preprint arXiv:1802.03162*, 2018.

[2] K. Althobaiti, N. Meng, and K. Vaniea, "I don't need an expert! making url phishing features human comprehensible," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[3] "2023 voice of the ciso global insights into ciso challenges, expectations and priorities," ProofPoint - Annual Report, 2023.

[4] F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof, and M. Köppen, "Detecting malicious urls using machine learning techniques," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2016, pp. 1–8.

[5] "Dbir 2023 data breach investigations report," Verizon Data Breach Investigations Report, 2023.

[6] M. A. Sasse, S. Brostoff, and D. Weirich, "Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security," *BT technology journal*, vol. 19, no. 3, pp. 122–131, 2001.

[7] Z. Yan, T. Robertson, R. Yan, S. Y. Park, S. Bordoff, Q. Chen, and E. Sprissler, "Finding the weakest links in the weakest link: How well do undergraduate students make cybersecurity judgment?" *Computers in Human Behavior*, vol. 84, pp. 375–382, 2018.

[8] D. O. Otieno, A. S. Namin, and K. S. Jones, "The application of the bert transformer model for phishing email classification," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2023, pp. 1303–1310.

[9] C. Iuga, J. R. Nurse, and A. Erola, "Baiting the hook: factors impacting susceptibility to phishing attacks," *Human-centric Computing and Information Sciences*, vol. 6, pp. 1–20, 2016.

[10] S. Sheng, B. Magnien, P. Kumaraguru, A. Acquisti, L. F. Cranor, J. Hong, and E. Nunge, "Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish," in *Proceedings of the 3rd symposium on Usable privacy and security*, 2007, pp. 88–99.

[11] M. Chatterjee and A.-S. Namin, "Detecting phishing websites through deep reinforcement learning," in *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2019, pp. 227–232.

[12] L. Yamoun, Z. Guessoum, and C. Girard, "Transformer roberta vs. tf-idf for websites content-based classification," in *Deep Learning meets Ontologies and Natural Language Processing, 3rd International Workshop, in conjunction with ESWC 2022*, 2022.

[13] P. Prakash, M. Kumar, R. R. Kompella, and M. Gupta, "Phishnet: predictive blacklisting to detect phishing attacks," in *2010 Proceedings IEEE INFOCOM*. IEEE, 2010, pp. 1–5.

[14] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond blacklists: learning to detect malicious web sites from suspicious urls," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 1245–1254.

[15] R. M. Mohammad, F. Thabtah, and L. McCluskey, "Tutorial and critical analysis of phishing websites methods," *Computer Science Review*, vol. 17, pp. 1–24, 2015.

[16] P. Maneriker, J. W. Stokes, E. G. Lazo, D. Carutasu, F. Tajaddodianfar, and A. Gururajan, "Urltran: Improving phishing url detection using transformers," in *MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM)*. IEEE, 2021, pp. 197–204.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[19] M. S. I. Mamun, M. A. Rathore, A. H. Lashkari, N. Stakhanova, and A. A. Ghorbani, "Detecting malicious urls using lexical analysis," in *Network and System Security: 10th International Conference, NSS 2016, Taipei, Taiwan, September 28-30, 2016, Proceedings 10*. Springer, 2016, pp. 467–482.

[20] Y. Zeng, T. Zang, Y. Zhang, X. Chen, and Y. Wang, "A comprehensive measurement study of domain-squatting abuse," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[21] U. o. N. B. Canadian Institute for Cybersecurity, "Url dataset(iscx-url2016)." [Online]. Available: https://www.unb.ca/cic/datasets/url-2016.html

[22] PhishTank. [Online]. Available: https://phishtank.org/index.php

[23] F. Heimerl, S. Lohmann, S. Lange, and T. Ertl, "Word cloud explorer: Text analytics based on word clouds," in *2014 47th Hawaii international conference on system sciences*. IEEE, 2014, pp. 1833–1842.