Savleen Kaur
CWID: 10476867

Multi-Modal Hierarchical Attention Model for Developing Phishing Threat Intelligence

Abstract: One of the most persistent forms of cyber dangers, phishing website attacks, evolves and remains a major cyber hazard. To identify phishing websites, many detection methods have been presented earlier. The creation of deep representation-based methods capable of learning deep fraud cues for increased anti-phishing capacity was prompted by the limits of lookup systems (e.g., failing to address freshly generated assaults) and fraud cue-based methods (e.g., relying on feature engineering). These strategies, by focusing solely on URLs, overlook the other two crucial modalities of website content: textual information and visual design. Furthermore, the interpretability of these deep learning-based methods is constrained, lowering model trustworthiness and preventing actionable intelligence. This paper presents a multi-modal hierarchical attention model for phishing website detection that learns deep fraud cues from the three key modalities of website content.

Introduction: Phishing threat intelligence has long been an important part of cyber threat intelligence for security operations centers in businesses since phishing has remained a major and widespread cybersecurity risk for many electronic commerce (e-commerce) platforms where information like accounts, passwords, and credit card numbers from victims are compromised. As much as 18,000 active phishing websites have been recorded by 2019 per quarter.

Approach Earlier: Lookup systems, fraud cue-based approaches, and deep representation-based methods are the three categories of methodologies. Existing phishing website detection technologies have three major drawbacks. First, deep representation-based methods, is limited. Explainable phishing threat intelligence is critical for SOC analysts' decision-making when it comes to triaging, investigating, and remediating a potential attack. Second, only a few systems combine the deep representative fraud cues from many content modalities available on phishing websites to learn the deep representative fraud cues. At least three types of material are used to create the fraud cues: navigation content, informative content, and graphic content. These three modalities have distinct and complimentary purposes in deceiving victims, as well as fraud indications from each modality and their usefulness in phishing website detection. Third, deep learning strategies for harmonizing representations from several modalities are lacking. Because deep representations are formed by neural networks that are particular to the data modality, the deep pattern of each representation dimension is not always consistent across modalities. As a result, directly synchronizing deep representations from distinct modalities could be difficult.

MMHAM: The paper suggests using a multi-modal hierarchical attention model that employs a novel shared dictionary learning strategy for aligning representations from many modalities in the attention mechanism. In this paper, the suggested MMHAM not only learnt better deep cues for improved phishing detection, but it also produced a hierarchical interpretability system from which they could create phishing threat intelligence to inform phishing website detection at various levels. It can also be used to extract deep representations of fraud cues automatically from three modalities of content: the information content of webpage text, the navigation content of URL, and the visual content of images. Furthermore, the hierarchical attention mechanism is inspired by the hierarchical organization of different content modalities, where attention is assigned to each modality.

Navigation, information, and visual design modalities make up a website. The information modality is represented by the text on the webpage. As a result, team regards them as three modalities. In phishing detection, URLs, webpage content, and images each include independent but complementary fraud indications concerning navigation design traits, information design aspects, and aesthetic appearances. The URL is made up of several characters, while the webpage text is made up of words at the finer level. Multiple photos are also possible. The elements in a URL contribute to the fraud cues in different ways. Special characters, such as the @ symbol and unicode, may, for example, be more important than alphabetical characters. As a result, to assign value to each character, an attention mechanism is required. Words like "credit," "card," and "password" may be more beneficial in learning deep fraud signals because the webpage language delivers information with the goal of duping victims into providing sensitive information. Furthermore, the images may play a varied role inside the visual design modality. Images associated with fraudulent adverts, may contain more fraud indications than normal iconography. Attention should be paid to each aspect inside each modality collectively, as this will not only increase detection performance but also offer relevant CTI.

The paper suggests using a hierarchical attention method to analyze websites since the hierarchical structure of websites can aid in the detection of phishing sites. The multi-modal hierarchical attention mechanism is motivated by the variety of distinct modalities (MMHAM).
The encoder network from the multi-modal attention mechanism is used to detect phishing websites. The encoder network may get deep aligned representations for different modalities because of the multi-modal attention mechanism (i.e., URL, webpage text, and images).

There are two levels of attention processes in the encoder network. The classical attention mechanism computes the relevance score assigned to each modality's constituent at the first level. Attention model is introduced for each letter in the URL, each word in the webpage content, and each image among the photos. Character attention, word attention, and image attention are the terms used to describe them. These attention ratings are referred to as element attention because they are at the element level. The suggested multi-modal attention mechanism is the second level attention mechanism, which is designed to deal with representations from various modalities and assess the importance score for each. As a result, modality attention is the name given to this type of attention. A hierarchical interpretability system is thus built to provide interpretability at various levels, allowing for more informative interpretation. Furthermore, the model's hierarchical attention allows it to weigh distinct

modalities based on their relevance score, resulting in a better representation for improved anti-phishing capability.
Character Attention: On a character-by-character basis, the URL is examined. As a result, in both the modality representation and the encoder network, a character level LSTM model (character-LSTM) is used to encode the URL string. Because characters contribute differently to the URL's fraud cues, the attention method is used to focus on the most relevant characters in the construction of the URL representation.
Word Attention: To encode each word in the webpage content, a traditional deep learning model called word-level LSTM is used. To understand increased deep fraud cues, attention is introduced to pay attention to essential words in summarizing the website content.
Image Attention: To extract visual representations, a popular deep learning model called ResNet-50 is used. The pre-trained model is applied straight to the photos, yielding a vector with 2048 dimensions that summarizes the entire image, as is standard practice.

To validate the model, authors have taken more than 4300 websites from various domains like arts, business, computers etc and they've developed a testbed for it and ran multiple machine learning algorithms on it. Accuracy yielded was 97.26%, a high F1-score and a high precision. Furthermore, all the improvements showed statistically significant margins (p-value<0.0001) when compared to previous approaches for phishing website detection. This demonstrates the effectiveness of the planned MMHAM in detecting phishing websites. Also, deep learning-based methods tend to provide higher results, confirming the benefits of deep learning methods discovered in prior studies. The fraudsters jumble up "V", "U", "Y", and "O" to create combinations of characters that look like a legitimate URL. Furthermore, to dynamically offer the phishing website pages, fraudsters frequently use URL links with parameters.

Conclusion: The proposed methodology gives SOC analysts an automated and effective tool for detecting phishing websites that they can integrate into their existing cyber threat intelligence operations. Furthermore, the approach gives explainable phishing threat knowledge to help them make better decisions and perform further investigation. The patterns identified in the model (for example, each modality and key words) can help analysts not only identify phishing websites with confidence, but also prioritize resources to address different phishing threats based on explainable insights, and proactively take preventive measures for potentially impacted assets and personnel. With an innovative shared dictionary learning technique, the MMHAM incorporates multi-modal information such as navigation, textual information, and visual design. On the other hand, a hierarchical attention mechanism was implemented, allowing for a hierarchical interpretability system that allows for interpretable cybersecurity decision-making in the detection of phishing websites. Experiments demonstrate that MMHAM not only has the best detection results in terms of accuracy, precision, recall, and F1-score, but it also has a hierarchical interpretability system that can be used to determine the importance of each modality and the elements inside each modality. Furthermore, it was discovered that textual content plays a substantially more crucial role in phishing website detection, thanks to the hierarchical attention model. This demonstrated the need of using multi-modal information for detecting phishing websites.

In my opinion, the model does a remarkable job because if a website is discovered as a phishing website, the suggested approach can avoid assaults by providing users with clear notifications or disabling the page. Furthermore, the model's interpretability helps teach internet users how to be more resistant to phishing websites. Internet users will be less likely to be duped if they have a better understanding of phishing websites. Given the fact that most deep learning studies only focus on URL strings, this model's finding emphasizes the necessity of utilizing website description. Therefore, webpage description plays a significantly more important role in phishing website detection.

References:

https://doi.ieeecomputersociety.org/10.1109/TDSC.2021.3119323
https://www.computer.org/csdl/journal/tq/2022/02/09568704/1xDLLqovoxq