Research Paper

Nizar Azar

**Paper Title:** Phishing Website Detection through Multi-Model Analysis of HTML Content  
**Authors:** Furkan Çolhak, Mert İlhan Ecevit, Bilal Emir Uçar, Reiner Creutzburg, Hasan Dağ  
**Publication Date:** January 2024  
**Link:** https://arxiv.org/pdf/2401.04820

**Problem Statement**

The academic paper that I will be analyzing tackles the consistent and evolving threat of phishing by focusing on the analysis of HTML content from websites. As phishing attacks become more complex and sophisticated, conventional methods (like URL checking or simple blacklists) struggle to outsmart these smart cyber criminals. The authors of this paper try to improve detection accuracy by designing a model that extracts and fuses both structured (numeric/tabular) and unstructured (textual) features directly from the HTML source of webpages.

**Dataset Used**

MTLP Dataset: - C¸ olhak, Furkan: MTLP Dataset. https://drive.google.com/file/d/1Lp3ueOd7A xmAl2Y0jJ2U2XlEFa6q8AcT/view?usp=drive link

Aljofey’s Dataset: - Aljofey, A.e.a.: An effective detection approach for phishing websites using url and html features. Scientific Reports 12(1), 8842 (2022). https://doi.org/10.103 8/s41598-022-10841-5

**Methodologies Used**

The paper introduces the MultiText-LP model, a hybrid framework that integrates:

* **Multilayer Perceptron (MLP):** Used to process numeric and categorical features extracted from HTML (e.g., hyperlink counts, CSS/JavaScript file counts, and other structural attributes).
* **Pretrained Natural Language Processing (NLP) Models:** Two distinct models are deployed:
  + One model processes the “page title” (NLP-1).
  + Another analyzes the “page content” (NLP-2).

The secret sauce of their approach is the combination of both MLP and NLP processing types. By putting the MLP and two NLP models into a single representation, which is then passed through a linear classifier, greatly utilizes the strengths of both structured and unstructured data analysis.

**Findings and Contributions**

* **High Detection Performance:** The fused MultiText-LP model achieved an impressive F1 score of 96.80% and an accuracy of 97.18% on their research dataset, outperforming both standalone models and existing approaches (e.g., Aljofey’s method).
* **Innovative Fusion Mechanism:** By combining features from two pretrained NLP models and an MLP, the study demonstrates that integrating multiple data modalities can greatly enhance their phishing detection performance. This conclusion is backed by their F1 score and accuracy, which are impressive and competitive compared to other approaches on the market.

**Relevance to Project**

The findings of this paper greatly helps us with our machine learning phishing detection application. The success and effectiveness with using the hybrid approach makes us very interested in their approach and further research is needed. Also, the open sharing of the dataset and detailed methodology used is an amazing reference, enabling us to replicate, compare, or build upon their approach in our own research.