RAJEEV GANDHI MEMORIAL COLLEGE OF ENGINEERING & TECHNOLOGY
APPROVED BY AICTE - NEW DELHI AFFILIATED TO JNTU ANANTAPURAMU
NANDYAL - 518501, ANDHRA PRADESH
(AUTONOMOUS)

Detecting Malicious URLS

DETECTING MALICIOUS URLS USING MACHINE LEARNING TECHNIQUES

Observations by

S. Abdul Hameed – 22091A3201

S. Arshiya Parveen – 22091A3210

Under Mentorship of

Dr. B. Bhaskara Rao

Overview

Introduction

- The digital world is advancing rapidly, with increased online activity.
- Cyberattack risks are rising due to attackers' inventive techniques.
- Malicious URLs are a critical attack vector, used to extract information and trick users.
- This review examines machine learning (ML) techniques for detecting malicious URLs.
- It addresses the lack of research on detecting malicious Arabic websites

Why it is Important?

- Malicious URLs can lead to system hacks and sensitive data breaches.
- They cause billions of dollars in losses each year.
- Traditional blacklisting methods have limitations and can be bypassed.
- ML offers a promising solution by learning from experience and improving self-learning without human intervention.
- Securing websites is crucial for online activities like e-commerce, business, social networking, and banking

Datasets used

Priority	Datasets used
Most common dataset	PhishTank
Second most common	Datasets built by study authors.
Third most common	Alexa
Other datasets	Malware Domain List, UCI Machine Learning Repository, Kaggle, OpenPhish, DMOZ, CommonCrawl

- Arabic studies rely on custom-built datasets due to the lack of public Arabic datasets

Feature Extraction

Features	Method of Extraction
Lexical Features	Elements of the URL string (length, special characters, digits).
Content-Based Features	Actual content on the page (HTML tags, scripts).HTML tag count, Iframe count, hyperlink count, number of scripts
Network Features	DNS, network, and host information (IP count, latency, redirection).Resolved IP count, latency, redirection count, domain lookup time

Machine Learning Models Used

- Algorithms achieving high accuracy (>=99%): CNN, XGBoost, LSTM, SVM, CW, Majority Voting Classifier, RF, K-means, Ara-means, DT, NB.
- Frequently used algorithms with good performance: SVM, RF, DT, NB, and LR.
- Ensemble techniques often provide high accuracy.
- Algorithms with lower performance: BN, NN, DBN.

Model Training and Evaluation

- Studies reviewed used supervised, unsupervised, and semi-supervised learning.
- Most studies used binary classification (malicious or benign).
- Evaluation metrics varied across studies, but accuracy, precision, and recall were common.
- K-fold cross-validation was used in some studies
- Accuracy scores varied from model to model by achieving 90-99.7%

Applications

Malicious URL detection is crucial for various applications:

- Web security.
- Email security.
- Network security.
- Protecting online transactions and user data.
- Combating phishing, spam, and malware attacks.

Findings and Results

- Lexical features are frequently used in both Arabic and English content analysis.
- Content-based features are more common in Arabic studies.
- Network-based features are not used in Arabic content analysis.
- The highest accuracy in English studies was 99.98% (CNN).
- The highest accuracy in Arabic studies was 99.521% (DT).
- Key network-based features include domain country code, domain update time, and contract expiration time.

Challenges and Our Extensions

Challenges	Our Extension
Should manually enter the URL	Creating a chrome extension that automatically detects
Limited to Binary Classification	Extension to Multi class Classification
Imbalanced Datasets	Data Augmentation for imbalanced Datasets
Single Model usage	Using Hybrid Models (ML+ Rule based)

Conclusion

- Lexical features are frequently used, and SVM, RF, CNN, and XGBoost are effective algorithms.
- Future research should focus on addressing challenges like dataset size and feature selection and detector sustainability.
- Further research is needed to enhance the accuracy and robustness of malicious URL detection techniques.

Thank you