Data Collection in Phishing Detection Using Deep Learning

Introduction

Phishing attacks are a significant cybersecurity threat, targeting individuals and organizations to steal sensitive information such as usernames, passwords, and financial details.

Deep learning techniques have been widely adopted to detect phishing attempts by analyzing large datasets of phishing and legitimate samples.

However, the success of deep learning models heavily depends on the quality and diversity of the collected data. This document explores the methods and sources for data collection in phishing detection using deep learning.

1. Sources of Data Collection

To effectively train deep learning models for phishing detection, data must be collected from various reliable sources. These sources can be categorized into the following:

1.1 Phishing Website Databases

Phishing websites are often short-lived, making real-time data collection essential. Several open-source repositories provide regularly updated lists of phishing URLs, including:

- PhishTank (https://www.phishtank.com/)

- OpenPhish (https://openphish.com/)

- APWG (https://apwg.org/)

1.2 Legitimate Website Lists

To differentiate phishing websites from legitimate ones, datasets must also contain valid websites. Trusted sources include:

- Alexa Top Sites (https://www.alexa.com/topsites)

- Common Crawl (https://commoncrawl.org/)

- Tranco List (https://tranco-list.eu/)

1.3 Email Datasets

Since phishing is commonly executed via email, email-based datasets are crucial for training deep

learning models. Some widely used email datasets include:

- Enron Email Dataset
- SpamAssassin Public Corpus
- UCEPROTECT Spam Database

1.4 Real-Time Data Scraping

Real-time data scraping helps in dynamically updating the dataset to keep up with evolving phishing techniques. Common methods include:

- Web Crawlers
- Email Parsing Scripts
- Network Traffic Analysis

1.5 User Reports & Threat Intelligence

Data reported by users and cybersecurity firms play a vital role in identifying new phishing threats. Security organizations like Google Safe Browsing and Microsoft SmartScreen provide updated phishing intelligence.

2. Feature Extraction from Collected Data

Once data is collected, it must be preprocessed and relevant features must be extracted for deep learning models to use.

2.1 URL-Based Features

- Length of URL
- Presence of special characters
- Domain age and WHOIS information
- Use of HTTPS vs. HTTP

2.2 Website Content-Based Features

- Presence of deceptive login forms
- Suspicious JavaScript functions
- Hidden iframes or redirections

2.3 Email-Based Features

- Sender domain reputation
- Presence of phishing keywords
- Email header anomalies

3. Data Preprocessing & Augmentation

Raw collected data requires preprocessing and augmentation to improve model performance.

- Data Cleaning
- Tokenization
- Feature Normalization
- Data Augmentation

4. Conclusion

Data collection is the foundation of phishing detection using deep learning. A well-curated dataset ensures better model accuracy and robustness.

Regular updates and continuous data refinement enhance AI-driven cybersecurity solutions.