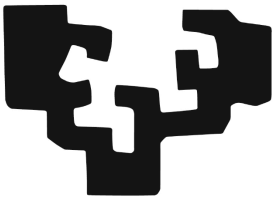
eman ta zabal zazu



Universidad del País Vasco

Euskal Herriko Unibertsitatea

Ciberseguridad

Carla Carbonell Canseco - ccarbonell001@ikasle.ehu.eus / a00573951@tec.mx

Phishing + Web Scraping + Machine Learning

Profesora:

Goizalde Badiola Zabala

23 de Diciembre de 2025

Phishing + Web Scraping + Machine Learning

Resumen

El phishing es uno de los ataques cibernéticos más comunes y efectivos en la actualidad, dirigido principalmente a usuarios finales mediante sitios web fraudulentos que imitan servicios legítimos con el objetivo de robar información sensible como credenciales o datos financieros. Los métodos tradicionales de detección, como listas negras o reglas estáticas, presentan importantes limitaciones frente a ataques nuevos y desconocidos. En este estudio se propone un enfoque de detección de phishing basado en contenido HTML utilizando técnicas de aprendizaje automático. Para ello, se construyó un conjunto de datos de sitios web legítimos y de phishing mediante técnicas de web scraping. A partir del código HTML de cada sitio se extrajeron características relevantes, que posteriormente se utilizaron para entrenar un modelo de clasificación supervisado. Los resultados obtenidos demuestran que las características basadas en contenido HTML permiten diferenciar eficazmente entre sitios legítimos y maliciosos.

1. Introducción

Los ataques de phishing representan una de las principales amenazas en el ámbito de la ciberseguridad. Este tipo de ataque se basa en técnicas de ingeniería social, donde el atacante engaña al usuario para que interactúe con un sitio web fraudulento y proporcione información confidencial. Debido a la facilidad con la que pueden crearse nuevos sitios de phishing, los métodos tradicionales de detección resultan insuficientes.

El aprendizaje automático ha emergido como una solución eficaz para abordar este problema, ya que permite identificar patrones complejos y generalizar a partir de datos previamente observados. En este trabajo se adopta un enfoque basado en contenido, analizando directamente la estructura HTML de las páginas web. El objetivo principal es diseñar un pipeline completo que incluya la recolección de datos mediante web scraping, la extracción de características, la construcción del dataset y el entrenamiento de un modelo de aprendizaje automático para la detección de phishing.

2. Trabajos relacionados

La literatura sobre detección de phishing con aprendizaje automático clasifica los enfoques en varias categorías: basados en URL, basados en contenido, basados en similitud visual y enfoques híbridos. Los métodos basados en URL analizan características léxicas como la longitud de la URL o la presencia de caracteres especiales, aunque pueden ser fácilmente evadidos.

Los enfoques basados en contenido analizan elementos internos del sitio web, como formularios, scripts, enlaces externos o campos de contraseña. Diversos estudios han demostrado que estas características son más difíciles de falsificar y ofrecen mejores resultados de generalización. Los enfoques híbridos combinan múltiples tipos de características para mejorar la precisión. Sin embargo, los principales desafíos señalados en la literatura incluyen la evolución constante de las técnicas de phishing, el desbalanceo de clases en los datasets y la necesidad de datos actualizados.

3. Metodología

3.1 Construcción del dataset

Para la construcción del conjunto de datos se utilizaron dos fuentes principales:

- **Sitios web legítimos**, obtenidos de la lista Tranco.
- **Sitios web de phishing**, obtenidos del dataset de PhishTank.

Se aplicaron técnicas de web scraping utilizando la librería *requests* para descargar el contenido HTML de cada URL. Para limitar el tiempo de ejecución, se trabajó con un subconjunto de URLs. Cada instancia fue etiquetada como legítima (0) o phishing (1) y almacenada en archivos CSV estructurados.

3.2 Extracción de características

A partir del HTML de cada sitio web, se extrajeron características basadas en contenido utilizando la librería BeautifulSoup. Entre las características consideradas se incluyen:

- Número de enlaces

- Número de formularios
- Número de campos de entrada
- Presencia de campos de contraseña
- Número de scripts
- Número de iframes
- Número de imágenes
- Número de enlaces externos

Estas características fueron seleccionadas por su relevancia en la detección de comportamientos típicos de sitios de phishing.

3.3 Modelado y evaluación

Se empleó aprendizaje automático supervisado para clasificar los sitios web. Los datos fueron divididos en conjuntos de entrenamiento y prueba. Se utilizó un modelo de Random Forest debido a su capacidad de generalización y robustez frente a ruido. Para evaluar el rendimiento del modelo se calcularon métricas estándar como accuracy, precision y recall, así como la matriz de confusión.

4. Resultados

Los experimentos realizados muestran que el modelo es capaz de distinguir eficazmente entre sitios legítimos y de phishing utilizando únicamente características basadas en contenido HTML. Las métricas obtenidas indican un buen equilibrio entre precisión y capacidad de detección, lo que confirma la utilidad de este enfoque frente a métodos tradicionales.

5. Conclusiones

En este trabajo se ha presentado un sistema completo de detección de phishing basado en contenido HTML mediante aprendizaje automático. Los resultados obtenidos demuestran que el análisis del código HTML proporciona señales relevantes para la identificación de sitios fraudulentos. Como trabajo futuro, se propone ampliar el dataset, incorporar nuevas características y evaluar otros algoritmos de aprendizaje automático para mejorar el rendimiento del sistema.