Website Detective

Supervisors:

João Paulo Barraca jpbarraca@ua.pt

Mário Antunes mario.antunes@ua.pt

Group size: 4

Tags: Machine Learning, Cybersecurity, Fraud Detection

Context

The Internet provides access to vast amounts of services and information, but not all is correct and to be trusted. Several websites appear with the sole purpose of defrauding users, with fake shops, fake servers or by providing fake news. Similar trends are being followed in email communications, with frequent campaigns with fake websites to phish user credentials and other sensitive data.

Detecting these fake websites and scams is vital for individuals and organizations, and new technologies can greatly help in this task, keeping everyone safe. This can be done by analyzing fingerprints on the website, such as its content, creation date, technology stack, hosting provider, reputation track, and even user feedback.

Objectives

This project will address the issue of fake websites, aiming to create a portal that can be used by the academic community, or even the entire public in determining if a website is potentially harmful. The approach proposed for the work will make use of OSINT from external sources and Machine Learning techniques to analyze the website content and extract relevant indicators (from the semantic content using word embeddings to metadata statistical analysis). We will consider the development of a front facing application, where users can specify an URL/QR code and check the consensus about it. They can also vote on the URL (Up or Down), allowing the community to provide further feedback.

Workplan

1. Product definition and requirements
2. Use case definition
3. Identification of data sets and external data sources
4. Identification of relevant indicators
5. First iteration of the frontend and backend components
6. Obtaining feedback from users
7. Second iteration of the frontend and backend components
8. Final delivery and demonstration

Remarks

The project will make use of existing open source information from other services, existing datasets and well established Machine Learning methods and technologies. Development will focus on the product requirements, backend to check the website correctness and the frontend presented to users.

The datasets used can be of real data if that is deemed as relevant, in coordination with the Cybersecurity Office of the University of Aveiro.