Project Proposal

COMP 7118

Data Mining

FALL 2022

Sharat Chandra Kommineni

U00870249

Contact Mail: skmmnni1@memphis.edu

Should be submitted and approved by

What is the problem/application?

Detecting Phishing website

What methods will be tested or implemented?

- Feature extraction: -
  - Extract the necessary features for the project
- Classification:-
  - For a problem like detecting a phishing website is a supervised machine learning task. This comes under classification problem as the url of the website should be classified as phishing or legit.
  - It involves classification techniques like Decision Tree, Random forest, XGBoost and many**.**

What data sets will be used:

- The dataset is downloaded from
  - https://www.unb.ca/cic/datasets/url-2016.html for phishing urls
  - https://www.phishtank.com/developer_info.php for legitimate urls.
- In addition to this, we also collected articles from various online sources.

What are the potential challenges for implementation?

- Larger datasets require machines with higher computational capabilities to be processed, so takes more time and GPU.
- When new phishing strategies are introduced, phishing detection solutions do have a low detection accuracy and high false alarm rates.

What are the expected deliverables?

Initial observations

Description of source datasets

- The dataset is University of New Brunswick, https://www.unb.ca/cic/datasets/url-2016.html has legitimate URLs.
- The latest data from https://www.phishtank.com/developer_info.php has phishing URLs.

Previous/similar works

- Detecting phishing websites using machine learning technique
  https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0258361

- Phishing Websites Detection using Machine Learning Arun D. Kulkarni akulkarni@uttyler.edu Leonard L. Brown, III
  https://scholarworks.uttyler.edu/cgi/viewcontent.cgi?article=1022&context=compsci_fac
- DATASET ANALYSIS USING WEKA: "PHISHING WEBSITES DATASET"
  https://local.cis.strath.ac.uk/wp/extras/msctheses/papers/strath_cis_publication_2707.pdf

How the solution is implemented:

Discussion of algorithms used

Decision tree:-Decision tree learning is a process of finding the optimal rules in each internal tree node according to the selected metric. **Decision Tree** is a supervised algorithm used in machine learning.

Random Forests:-Multiple decision trees are built during training via the ensemble learning technique known as Random Forest. For classification problems, it forecasts the mode of the classes, and for regression tasks, it forecasts the mean of the trees.

XGBoost: **XGBoost**, which stands for Extreme Gradient Boosting, is a scalable, distributed **gradient-boosted** decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine learning library for regression, classification, and ranking problems.

Temporal/spatial efficiency of algorithm

The efficiency of the algorithm is improved using the above-mentioned algorithms.

Results and analysis:

Document label discovery (i.e. topics, attitude, type, quality)

We are going to build the model a d then evaluate the performance of the described models(Accuracy) .Later compare the models based on their accuracy and suggest the best one.

Time and memory calculations

It may require 4GB GPU, 4GB CPU for getting accurate results.

Prediction accuracy

The accuracy of the project will be more than 90% if we use the above-mentioned algorithms and with proper data preprocessing.

Conclusions

- Thoughts on potential future works

  The goal of this report's future work is to create an unsupervised deep learning technique that can extract knowledge from URLs. The research can also be expanded in order to obtain results for a bigger network and preserve an individual's privacy.

- Potential bias or discrepancies in findings

  There are no such discrepancies yet.