

Identificando URLs Maliciosas

Uma Abordagem Puramente Léxica

Julio Cesar da Silva Rodrigues¹

¹Universidade Federal de São João del-Rei
Curso de Ciência da Computação
julio.csr.271@aluno.ufsj.edu.br



- ① Introdução
- ② Trabalhos Relacionados
- ③ Metodologia
- ④ Resultados
- ⑤ Conclusão

1 Introdução

2 Trabalhos Relacionados

3 Metodologia

4 Resultados

5 Conclusão

Contextualização

- Via rápida e direta para aplicar crimes cibernéticos;
- Potencial de infecção exponencial;
- Brasil no Top 15 com maior número de vítimas, segundo relatório¹ da Abranet;
- Evolução nas técnicas de camuflagem e detecção.

¹Disponível em: <https://www.abranet.org.br/Noticias/>

Relatorio-aponta-que-cada-URL-maliciosa-no-Brasil-afeta-18-usuarios-2585.html?

UserActiveTemplate=site

Objetivos

- Principais características que definem a natureza de uma URL;
- Capacidade de predição analisando somente a estrutura léxica;
- Modelo competitivo com classificação multiclasse.

1 Introdução

2 Trabalhos Relacionados

3 Metodologia

4 Resultados

5 Conclusão

Classificação Multiclasse

- Detecção de URLs baseada somente em atributos léxicos [Saleem Raja et al., 2021];
- Base de dados da UNB²;
- Extração de 27 atributos;
- Cinco algoritmos de *machine learning* selecionados;
- 99% de acurácia com *random forest*.

²University of New Brunswick

Classificação Binária

- Detecção de URLs de *phishing* direcionadas a brasileiros [Ayres et al., 2019];
- Aprendizado baseado em atributos léxicos e relacionados à rede;
- Bases de dados nacionais:
 - ① CaUMa³: URLs maliciosas;
 - ② UFBA⁴: URLs seguras.

³Serviço associado ao Catálogo de Fraudes da RNP

⁴Universidade Federal da Bahia

Classificação Binária

- Extração de 117 características;
- Quatro algoritmos de *machine learning* selecionados;
- Empate técnico com *F1 Score* média de 95,85%:
 - ① KNN;
 - ② SVM;
 - ③ J48.

1 Introdução

2 Trabalhos Relacionados

3 Metodologia

4 Resultados

5 Conclusão

Base de Dados

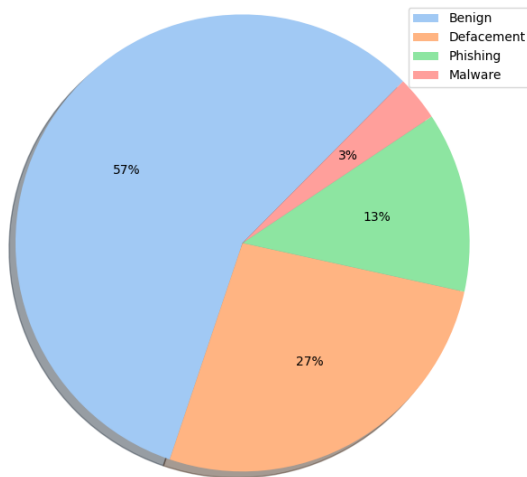
- Kaggle⁵:
 - ➊ Mais de 600 mil instâncias;
 - ➋ 4 tipos de URL.
- PhishTank⁶:
 - ➊ Mais de 100 mil instâncias;
 - ➋ Somente URLs de *phishing*.

⁵Disponível em: <https://www.kaggle.com/datasets/sid321axn/malicious-urls-dataset>

⁶Disponível em: https://phishtank.org/phish_archive.php

Base de Dados

Distribuição de Classes



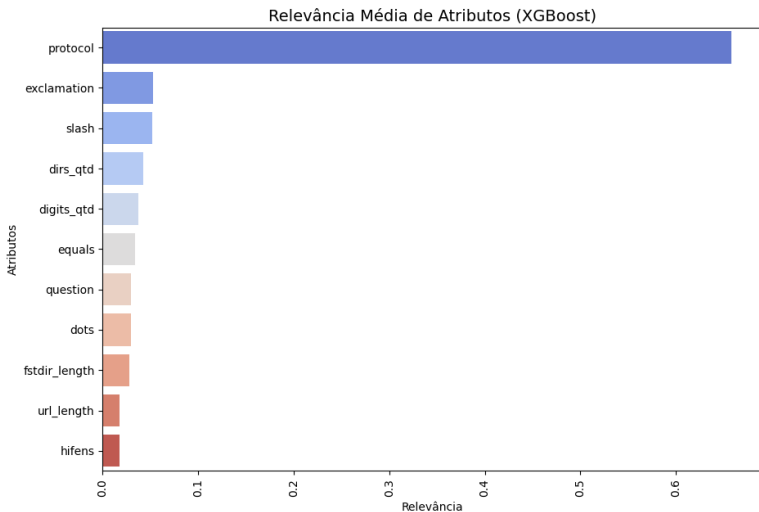
Balanceamento

- *Undersampling* aleatório para instâncias *benign*;
- Redução de 428081 para 200 mil instâncias;
- *Scraper* para coletar mais de 100 mil URLs da PhishTank;
- *SMOTE* [Bowyer et al., 2011] para suprir o déficit de instâncias de *defacement* e *malware*.

Construção de atributos

- 21 atributos criados:
 - ① 7 com base em análise estatística da base de dados;
 - ② 14 criados utilizando trabalhos relacionados como base;
 - ③ Exemplos: {Protocolo de comunicação, Comprimento da URL, Quantidade de *tokens*, ...}.
- Aplicação de *feature selection* com **RFECV**.

Construção de atributos



Base de Dados Final

- 800 mil instâncias;
- Perfeitamente balanceada;
- 11 atributos e 1 classe (4 valores distintos);
- Composta por URLs de bases de dados do *Kaggle* e *PhishTank*.

Modelos de Aprendizado Supervisionado

- Algoritmos selecionados:
 - ① XGBoost;
 - ② KNN;
 - ③ Regressão Logística.
- Validação cruzada com amostragem estratificada (*10-fold*);
- Métrica *Macro F1* e desvio padrão;
- Ajuste fino - *Tripartite (5-fold)*;
- Teste t de dupla cauda.

- 1 Introdução
- 2 Trabalhos Relacionados
- 3 Metodologia
- 4 Resultados**
- 5 Conclusão

Comparativo de Algoritmos

Tabela 1: *Macro F1 Scores* alcançadas pelos algoritmos

Resultados		
Algoritmo	Média	Desvio Padrão
Regressão Logística	0,7339	0,0039
XGBoost	0,9476	0,0006
KNN	0,9229	0,0008

- XGBoost e KNN com desempenho similar inicialmente;
- Desvio padrão constantemente baixo é um bom indicativo.

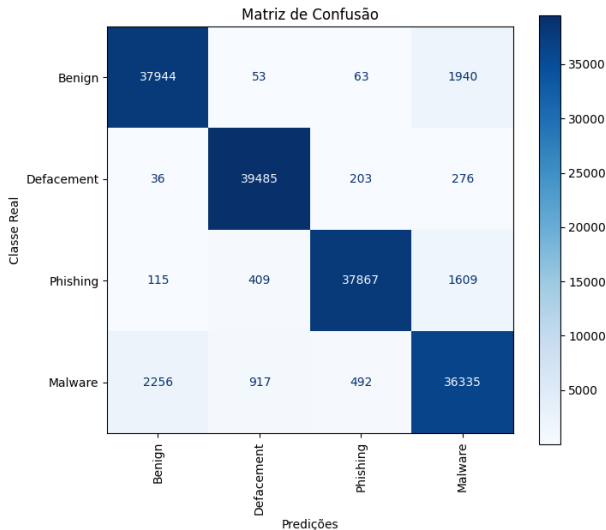
Teste Estatístico

- Nível de significância $\rightarrow \alpha = 0,05$;
- XGBoost x KNN:
 - ① Hipótese nula rejeitada;
 - ② Modelos estatisticamente distintos.
- XGBoost x Regressão Logística:
 - ① Hipótese nula rejeitada;
 - ② Modelos estatisticamente distintos.

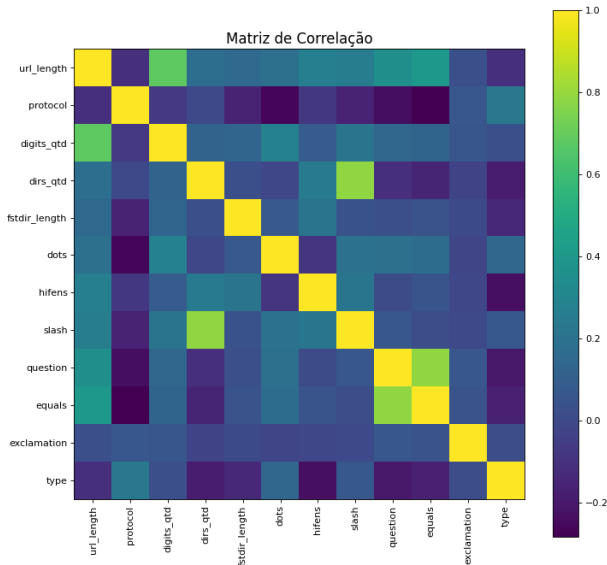
Teste Estatístico

- KNN x Regressão Logística:
 - ① Hipótese nula rejeitada;
 - ② Modelos estatisticamente distintos.
- XGBoost na dianteira como o melhor desempenho;
- Alta disparidade em custo computacional.

Limitações



Limitações



- 1 Introdução
- 2 Trabalhos Relacionados
- 3 Metodologia
- 4 Resultados
- 5 Conclusão**

Considerações Finais

- Distinção entre URLs maliciosas;
- Potencial para competir com modelos que utilizam outros tipos de características;
- Generalização e nível de confiabilidade na classificação.

Passos Futuros

- Exploração mais profunda na construção de atributos;
- Investigação minuciosa sobre o que separa *malware* de outros tipos de URL;
- Possíveis reduções no volume da base de dados (*instance selection?*).

Referências

- [Ayres et al., 2019] Ayres, L., Brito, I. V. S., and e Souza, R. G. (2019).
Utilizando aprendizado de máquina para detecção automática de urls maliciosas
brasileiras.
*In Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas
Distribuídos*, pages 972–985, Porto Alegre, RS, Brasil. SBC.
- [Bowyer et al., 2011] Bowyer, K. W., Chawla, N. V., Hall, L. O., and Kegelmeyer,
W. P. (2011).
SMOTE: synthetic minority over-sampling technique.
CoRR, abs/1106.1813.
- [Saleem Raja et al., 2021] Saleem Raja, A., Vinodini, R., and Kavitha, A. (2021).
Lexical features based malicious url detection using machine learning techniques.
Materials Today: Proceedings, 47:163–166.
NCRABE.