

AVALIAÇÃO DE QUALIDADE DE CONJUNTOS DE DADOS DE MALWARE PARA APRENDIZADO DE MÁQUINA

**Thaís da Silva Hernandez, Caroline Duarte
Gandolfi**

Instituto Militar de Engenharia
thaísa.silva@ime.eb.br,
caroline.gandolfi29@ime.eb.br

Lucila M. S. Bento

Instituto de Matemática e Estatística - Universidade do
Estado do Rio de Janeiro
lucila.bento@ime.uerj.br

Luiz Bonino da Silva Santos

Faculty of Electrical Engineering, Mathematics and
Computer Science, University of Twente
l.o.boninodasilvasantos@utwente.nl

Augusto Parisot de Gusmão Neto

Centro de Análises de Sistemas Navais
aparisot@id.uff.br

Raphael C. S. Machado

Instituto de Computação, Universidade Federal
Fluminense
raphaelmachado@ic.uff.br

Anderson Fernandes Pereira dos Santos, Maria

Cláudia Reis Cavalcanti

Instituto Militar de Engenharia
anderson@ime.eb.br, yoko@ime.eb.br

RESUMO

Com o crescimento do ciberespaço, crescem também os danos causados por *malwares*, que são uma das principais ferramentas utilizadas por agentes maliciosos. Algoritmos de aprendizado de máquina vêm se consolidando como ferramentas importantes na detecção de ameaças. Modelos usados por esses algoritmos dependem de dados para treinamento e teste. Nesse sentido, os conjuntos de dados de *malware* tornaram-se valiosos na implantação de sistemas anti-malwares modernos. No entanto, esses conjuntos de dados enfrentam problemas de qualidade das amostras, além de não acompanharem a velocidade da evolução tecnológica, tornando-se obsoletos. Além disso, muitos dos conjuntos de dados utilizados em pesquisas não são de acesso público. Este trabalho propõe uma estrutura de avaliação de qualidade baseada em métricas focada na amostragem e na temporalidade dos dados. Ademais, incorpora critérios alinhados aos princípios FAIR, com o objetivo de incentivar a publicação de conjuntos de dados mais confiáveis e reutilizáveis.

PALAVRAS CHAVE. conjuntos de dados, análise de malware, FAIR.

Tópicos (Qualidade de Dados, Aprendizado de Máquina, Defesa e Segurança Cibernética)

ABSTRACT

As cyberspace grows, so does the damage caused by *malware*, which is one of the main tools used by malicious agents. Machine learning algorithms have been consolidated as important tools for detecting threats. Models used by these algorithms depend on data for training and testing. In this sense, *malware* datasets have become valuable in the deployment of modern anti-malware systems. However, these datasets face problems with the quality of the samples, as well as not keeping up with the speed of technological evolution and becoming obsolete. In addition, many of the datasets used in research are not publicly accessible. This paper proposes a quality assessment framework based on metrics focused on sampling and data temporality. It also incorporates criteria aligned with the FAIR principles, with the aim of encouraging the publication of more reliable and reusable datasets.

KEYWORDS. datasets, malware analysis, FAIR.

Paper topics (Data Quality, Machine Learning, Cyber Defense and Security)