UNIVALI

Escola do Mar, Ciências e Tecnologias / Escola de Artes, Comunicação e Hospitalidade

Curso de Sistemas para Internet

Disciplina Internet das Coisas

Prof.Felipe Viel

Detecção de Malicious and Non Malicious URLs

Machine Learning

Link do Github: https://github.com/nicoletenorio/TrabalhoIOT_M3

Acadêmicos:

Dryan Jhonatas Dumke - dryan@edu.univali.br
Marcos Henrique Baumgartel Comper - marcos.comper@edu.univali.br
Nicole Ewellin Tenorio de Oliveira - oliveira.nicole@edu.univali.br

Itajaí, Abril, 2022

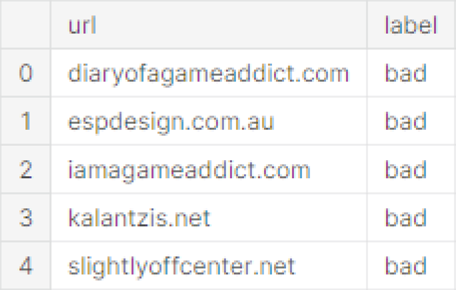
Introdução (Explicação do Tema)

Um site malicioso é composto por um software que tenta instalar um malware (Malware é a abreviação de "software malicioso" (em inglês, malicious software) e se refere a um tipo de programa de computador desenvolvido para infectar o computador de um usuário legítimo e prejudicá-lo de diversas formas, como coletar suas informações pessoais, ou na pior das hipóteses obter acesso total ao computador). Essas ações geralmente requerem alguma ação da sua parte, porém no caso de um download de drive-by o site tentará instalar o software no computador sem requisitar permissão ao dono.

Essa aplicação irá criar um modelo que pode detectar sites maliciosos, a URL do site é usada como um recurso e a Rede Neural Convolucional é usada como algoritmo para detecção de sites maliciosos. O modelo será validado pelo método holdout.

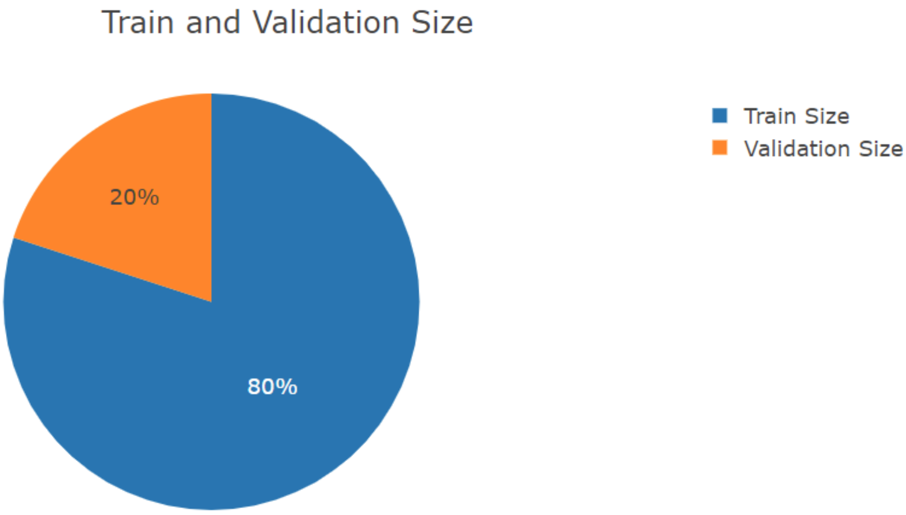
Demonstração da Aplicação

Carregando os dados dos sites maliciosos:



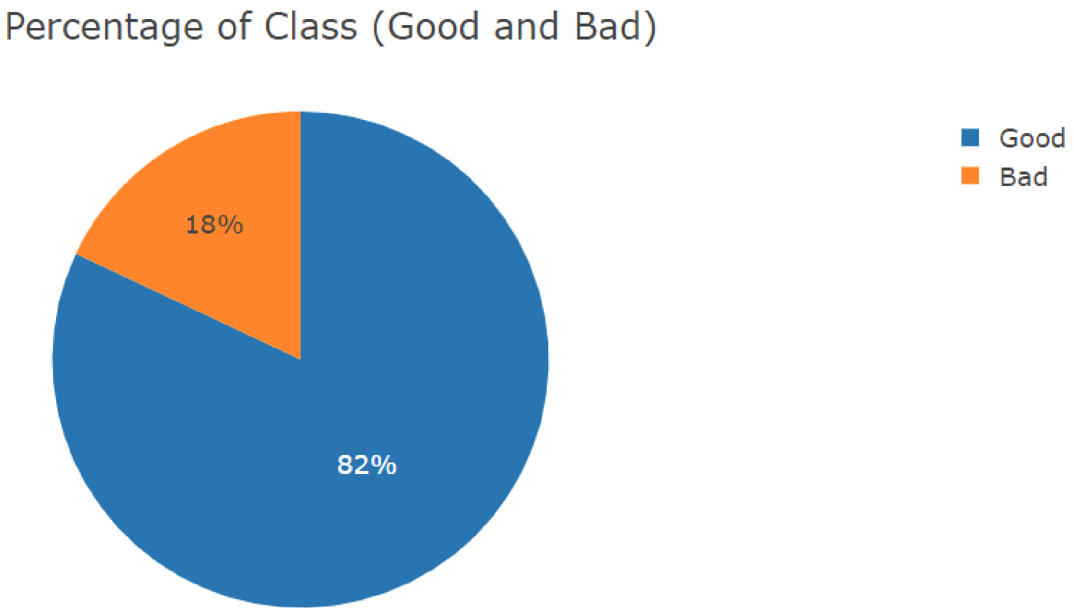
	url	label
0	diaryofagameaddict.com	bad
1	espdesign.com.au	bad
2	iamagameaddict.com	bad
3	kalantzis.net	bad
4	slightlyoffcenter.net	bad

Neste bloco de notas, o método de validação utilizado é o método de validação. O método holdout é um método que separa os dados de treinamento e teste em 80% e 20%.

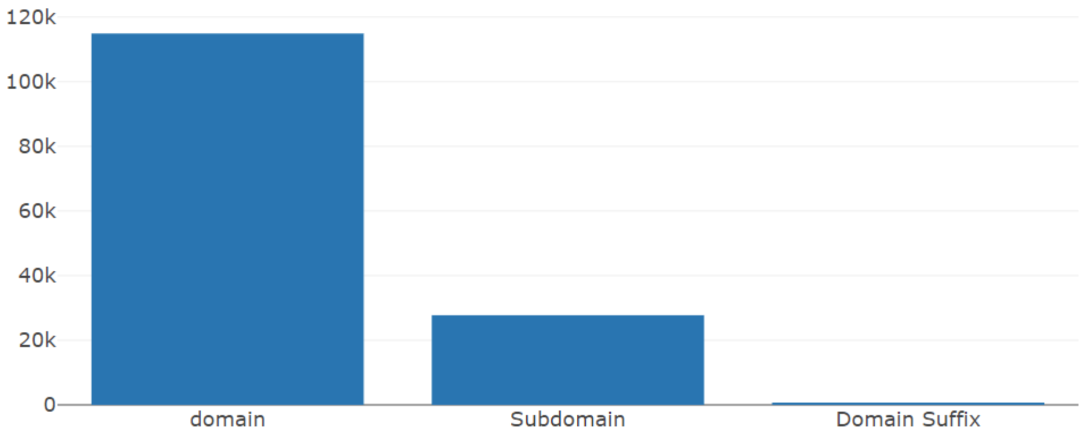


Sobre a análise de dados e engenharia dos recursos:

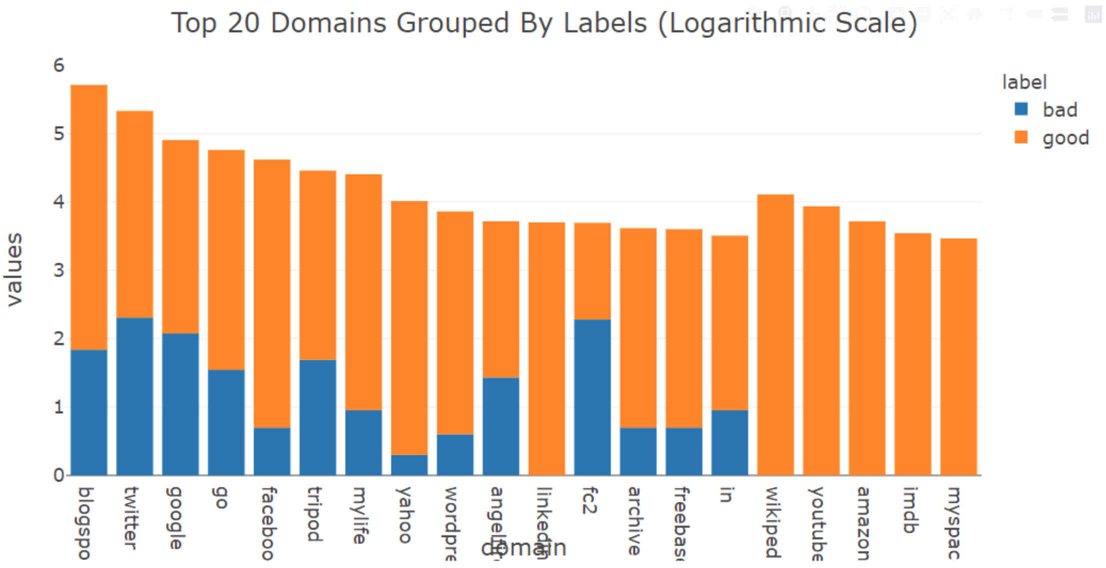
Análises de dados para expandir o conhecimento sobre esses dados e fazer uma engenharia de recursos. Primeiro queremos descobrir se os dados estão desequilibrados.



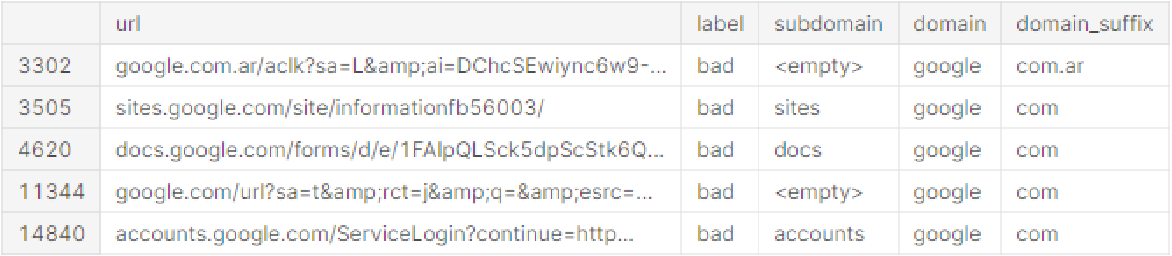
Como você pode ver no gráfico acima, 82% têm rótulos bons, enquanto 18% têm rótulos ruins. Em seguida, vamos descobrir o domínio de sufixo, domínio e subdomínio mais usados. Precisamos extrair subdomínios, domínios e sufixos de domínio para poder fazer a análise. Vamos verificar quantos domínios, subdomínios e sufixos de domínio únicos que extraímos.



Em seguida é traçado o recurso de domínio;

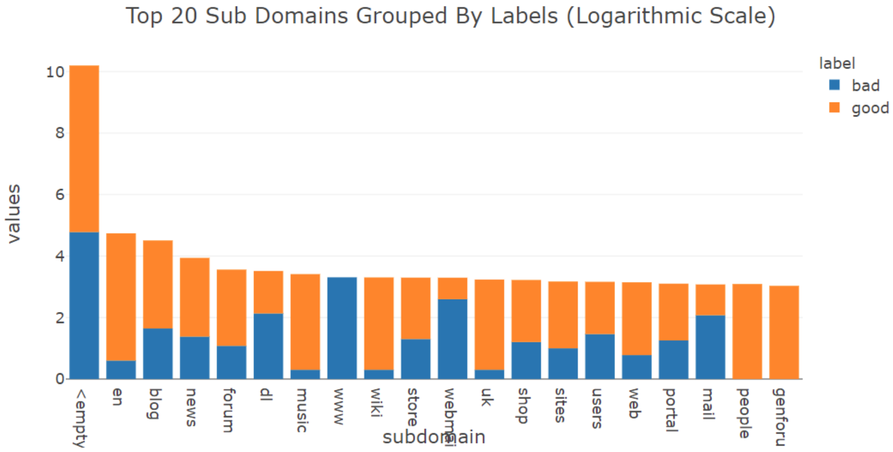


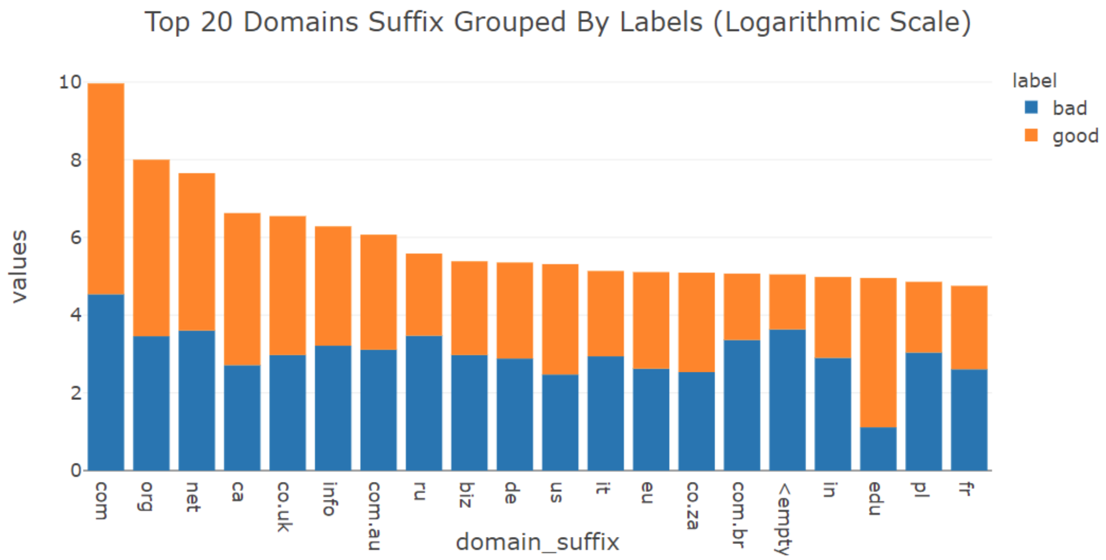
Com base no gráfico acima, há coisas interessantes a serem observadas, existem alguns domínios famosos que não contêm rótulos ruins como linkedin, wikipedia, youtube, amazon, imdb e myspace e há muitos domínios famosos que contêm rótulos ruins, por exemplo, google. Vamos dar uma amostra no domínio do google para ver os rótulos ruins.



	url	label	subdomain	domain	domain_suffix
3302	google.com.ar/aclk?sa=L&ai=DChcSEwiync6w9-...	bad	<empty>	google	com.ar
3505	sites.google.com/site/informationfb56003/	bad	sites	google	com
4620	docs.google.com/forms/d/e/1FAIpQLSck5dpScStk6Q...	bad	docs	google	com
11344	google.com/url?sa=t&rct=j&q=&esrc=...	bad	<empty>	google	com
14840	accounts.google.com/ServiceLogin?continue=http...	bad	accounts	google	com

Talvez algumas dessas urls contenham malware, quem sabe…





Como você pode ver no gráfico acima, a maioria deles contém rótulos ruins, mesmo que existam alguns que não são, o próximo passo, precisamos fazer tokenização na url para que ela possa ser usada como entrada para o modelo CNN.

1 199 / 5 000
Resultados da tradução
Antes da tokenização:
mister-ed.com/welcome/file/update/rbc/login.php

Após a tokenização:
[12, 5, 9, 7, 2, 10, 15, 2, 16, 13, 8, 3, 12, 6, 26, 2, 14, 8, 3, 12, 2, 6, 25, 5, 14 , 2, 6, 19, 17, 16, 4, 7, 2, 6, 10, 21, 8, 6, 14, 3, 20, 5, 11, 13, 17, 18, 17]

Cada URL tem um comprimento diferente, portanto, o preenchimento é necessário para equalizar cada comprimento de URL. Próximo passo faremos o preenchimento na url que já temos tokenize

Antes do preenchimento:
[12, 5, 9, 7, 2, 10, 15, 2, 16, 13, 8, 3, 12, 6, 26, 2, 14, 8, 3, 12, 2, 6, 25, 5, 14 , 2, 6, 19, 17, 16, 4, 7, 2, 6, 10, 21, 8, 6, 14, 3, 20, 5, 11, 13, 17, 18, 17]
Após o preenchimento:
[12 5 9 7 2 10 15 2 16 13 8 3 12 6 26 2 14 8 3 12 2 6 25 5
14 2 6 19 17 16 4 7 2 6 10 21 8 6 14 3 20 5 11 13 17 18 17 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

```
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]
```

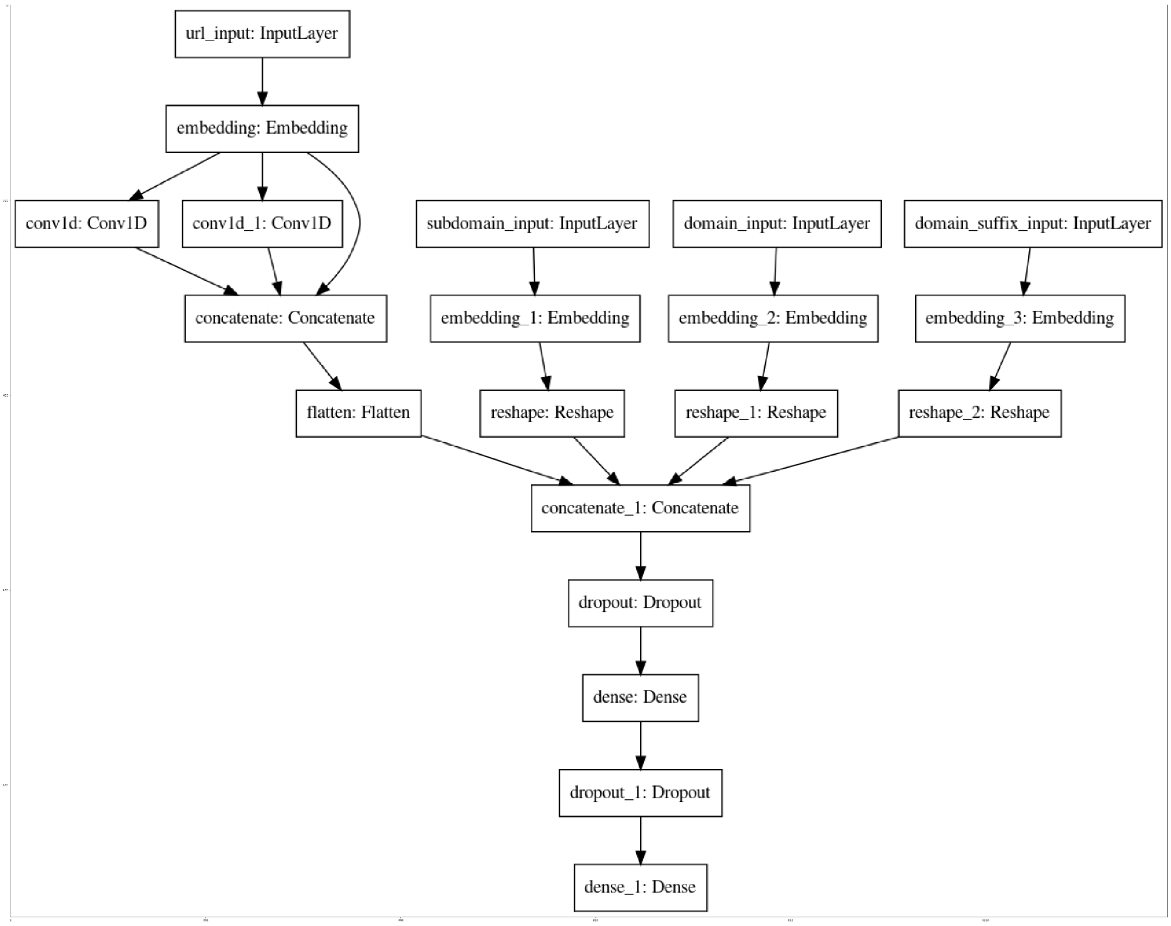
Também codificamos subdomínio, domínio, domínios de sufixo e rotulamos em variáveis numéricas

	url	label	subdomain	domain	domain_suffix
0	mister-ed.com/welcome/file/update/rbc/login.php	bad	0	0	0
1	ip-23-229-147-12.ip.secureserver.net/public/fi...	bad	1	1	1
2	facebok-info.com/unitedkingdom/log.php	bad	0	2	0
3	independent.co.uk/news/obituaries/john-gross-g...	good	0	3	2
4	facebook.com/geoffrey.gray	good	0	4	0

O próximo passo é codificar a variável de destino (rótulo) para numérico, por exemplo, o rótulo ruim se torna 1 e o rótulo bom se torna 0.

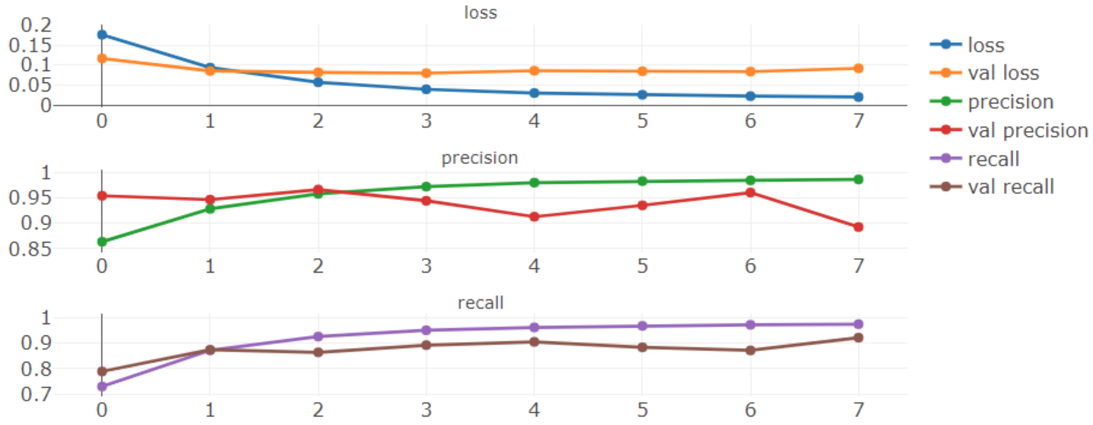
	url	label	subdomain	domain	domain_suffix
0	mister-ed.com/welcome/file/update/rbc/login.php	1	0	0	0
1	ip-23-229-147-12.ip.secureserver.net/public/fi...	1	1	1	1
2	facebok-info.com/unitedkingdom/log.php	1	0	2	0
3	independent.co.uk/news/obituaries/john-gross-g...	0	0	3	2
4	facebook.com/geoffrey.gray	0	0	4	0

Criando o modelo CNN



O modelo recebeu 4 entradas, a primeira entrada veio da URL que foi feita a tokenização e preenchimento. Outras entradas são subdomínios, domínios e domínios de sufixo que foram codificados. A entrada de URL passará pela camada de incorporação e pela camada de convolução, enquanto outras entradas passarão pela camada de incorporação. Em seguida, os resultados de cada entrada serão concatenados.

Treinamento de modelo;



Resultados Obtidos (Explicação)

```
Validation Data:
0    68964
1    15129
Name: label, dtype: int64
```

```
Confusion Matrix:
[[68408   556]
 [ 1569 13560]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.98      0.99      0.98     68964
           1       0.96      0.90      0.93     15129

    accuracy                           0.97     84093
   macro avg       0.97      0.94      0.96     84093
weighted avg       0.97      0.97      0.97     84093
```

Conclusão obtidas com a Pesquisa

Em conclusão, o modelo que foi treinado tem um valor de alta precisão e revocação, mas o que deve ser considerado é o valor de precisão. O valor de precisão deve ser alto porque, se for baixo, um site que não seja malicioso tem a possibilidade de ser classificado como malicioso

Referências:

KASPERSKY. **Malware**. Disponível em: https://www.kaspersky.com.br/resource-center/preemptive-safety/what-is-malware-and-how-to-protect-against-it. Acesso em: 28 jun. 2022.

NORTON. **Malware**. Disponível em: ttps://us.norton.com/internetsecurity-malware-what-are-malicious-websites.html. Acesso em: 28 jun. 2022.