MỘT CÁCH TIẾP CẬN SỬ DỤNG HỌC SÂU TRONG PHÁT HIỆN MÃ ĐỘC POWERSHELL

Ngô Đức Hoàng Sơn - 230202030[1]

[1] Trường Đại học Công nghệ Thông tin, ĐHQG TPHCM

What ?

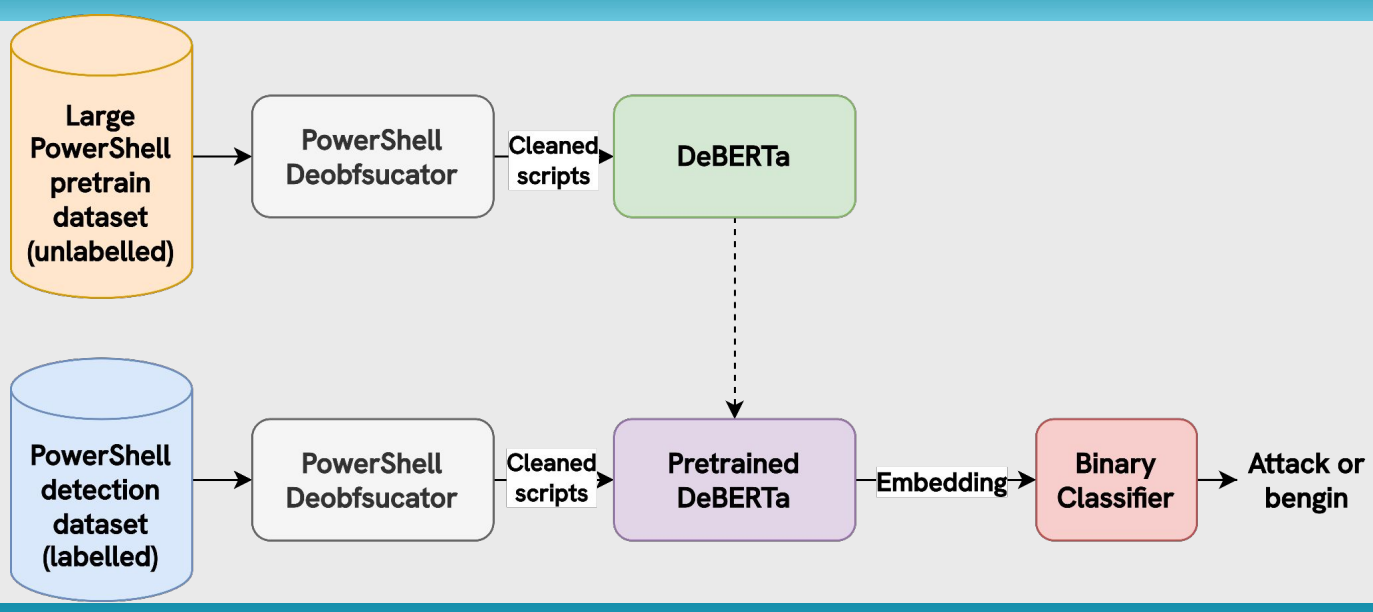
Một mô hình học sâu phát hiện mã độc PowerShell có khả năng trốn tránh phát hiện một cách tinh vi.

- Đề xuất mô hình phát hiện sử dụng mô hình ngôn ngữ DeBERTa có hiệu suất cao.
- Mô hinh có thể phát hiện các mã độc đã bị làm rối mã thông qua quá trình tiền xử lý.
-

Why ?

- Mã độc PowerShell không bị phát hiện bởi các công cụ truyền thống.
- Các phương pháp sử dụng học sâu hiện nay tập trung vào việc sử dụng các mô hình ngôn ngữ ở mức độ token.
- Các mô hình ngôn ngữ hiện đại đang ngày càng phát triển như transformer-based. DeBERTa được xem là một trong những mô hình có hiệu năng tốt, tuy vậy vẫn chưa được sử dụng rộng rãi trong phát hiện mã độc

Overview



Description

1. Xây dựng bộ dữ liệu

- Xây dựng 2 bộ dữ liệu về mã nguồn PowerShell.
- Bộ dữ liệu không nhãn sẽ được sử dụng cho việc tiền huấn luyện mô hình.
- Bộ dữ liệu có nhãn sẽ được sử dụng cho việc huấn luyện mô hình phát hiện.
- Các bộ dữ liệu sẽ được thu thập từ các nguồn chia sẻ, mã nguồn mở

```
(New-Object System.IO.StreamReader @((New-Object IO.Compression.
DeflateStream @([System.IO.MemoryStream][Convert]::FromBase64String
('NZBRa8IwFEb/Sh8CUZzpsDidoeA2HbiissnoBntp4×2JTZM2TVuL+N9Xy/p6z
+HAd1EuA19BPdbxCZh1dmBJCPGLFKAsRfvtwcfc2mzhukwomN8Tpkli3Lqq3ZwH66mRycfy35Ci
sI0uDdcpFIRFrp7mqyJ8hl7gwsC4ilQbSd2HY1k0/
LH2eirY2ZYmgabDaj353j6dPoMeF5mWkRGMd7iSbxPtJRM+w+SQSWEHeImHFH3FR8d3sDebY4pM
+O4jUNXCQpqN8A8e3fgIEzgDpr/aQMT4AO1D5gjl3NYOL9Y0F9R
+hax0raSOjq9CQufcObfgkG5UpRMYb9pod6Fx20nolUWW8cv1+gc='), [IO.Compression.
CompressionMode]::Decompress)), [Text.Encoding]::ASCII)).ReadToEnd() |
($PsHome[21] + $PsHome[34] + 'X');
| Powershell $pC695I = [Type] 'eNvIRONMent';
```

Hình 2. Một đoạn mã độc PowerShell được làm rối

2. Gỡ rối mã

- Các đoạn mã độc thường được các kẻ tấn công làm rối.
- Các đoạn mã khác nhau cũng có quy cách đặt tên biến khác nhau.
- Vì vậy chúng cần được gỡ rối và làm sạch về một định dạng nhất định.
- Hình 2 biểu diễn một đoạn mã PowerShell đã bị làm rối

3. Tiền huấn luyện

- Bộ dữ liệu mã nguồn PowerShell không có nhãn được sử dụng nhằm tiền huấn luyện mô hình DeBERTa.
- Mô hình sau khi được tiền huấn luyện có thể tạo nhúng với đầu vào là các đoạn mã PowerShell.
- Mô hình sau khi được tiền huấn luyện sẽ được tinh chỉnh nhằm phát hiện mã độc

3. Tiền huấn luyện

- Mô hình tiền huấn luyện DeBERTa với mã nguồn sẽ được tinh chỉnh với bộ dữ liệu có nhãn.
- Mô hình phát hiện sẽ thực hiện phân loại nhị phân. Đầu ra sẽ là đoạn mã đầu vào có phải là mã độc hay không.

NII

Tên tác giả chính– Trường Đại học
TEL : ******* Email : *********