Universidad del Valle de Guatemala
Security Data Science
Jorge Yass
Sección 10

Fase 1

Procesamiento del lenguaje natural (NLP) para la detección de ataques de ingeniería social

Roberto Castillo 18546
Hugo Roman 19199
Oscar de León 19298
Mirka Monzón 18139
Josué Sagastume 18173

Motivación

El objetivo principal del proyecto es utilizar técnicas de procesamiento del lenguaje natural (NLP) para construir un modelo de aprendizaje automático que pueda detectar y prevenir ataques de ingeniería social en comunicaciones textuales como correos electrónicos y mensajería instantánea. El proyecto busca identificar los rasgos lingüísticos que indican un ataque de ingeniería social, entrenar modelos de machine learning para reconocer estos rasgos y utilizar el modelo para detectar futuros ataques de ingeniería social

El proyecto tiene como propósito proteger la información confidencial y los sistemas de las organizaciones contra las amenazas de ingeniería social. Al detectar los ataques de ingeniería social de manera temprana, se pueden prevenir robos de información, intrusiones en sistemas y otros daños a la empresa. Además, el proyecto tiene el propósito de reducir el número de falsos positivos y garantizar que el modelo de aprendizaje automático tenga una alta precisión en la detección de ataques de ingeniería social.

En resumen, el objetivo del proyecto es construir un sistema de detección de ataques de ingeniería social utilizando técnicas de procesamiento del lenguaje natural (NLP) y aprendizaje automático para proteger la información confidencial y los sistemas de las organizaciones contra las amenazas de ingeniería social.

Preguntas clave

- ¿Qué rasgos lingüísticos de las comunicaciones escritas apuntan a un ataque de ingeniería social?
- ¿Cómo pueden los algoritmos de aprendizaje automático reconocer con fiabilidad estos rasgos lingüísticos para detectar futuros ataques de ingeniería social?
- ¿Qué medidas pueden adoptarse para reducir las falsas alarmas y, al mismo tiempo, identificar con gran precisión los ataques de ingeniería social?

Revisión de la literatura

- Poudyal, S., Dasgupta, D., Akhtar, Z., & Datta Gupta. (s. f.). *A Multi-Level Ransomware Detection Framework using Natural Language Processing and Machine Learning*. ResearchGate. https://www.researchgate.net/profile/Subash-Poudyal/publication/336251881_A_Multi-Level_Ransomware_Detection_Framework_using_Natural_Language_Processin

g_and_Machine_Learning/links/5dd57d9f458515cd48afd862/A-Multi-Level-Ransomware-Detection-Framework-using-Natural-Language-Processing-and-Machine-Learning.pdf

Este artículo explica las razones por las cuales la detección de ransomware tiende a ser difícil, poniendo énfasis en la posibilidad de detener estos en general debido a su constante evolución. En este artículo, se proponen distintas perspectivas con las cuales se podría abordar el problema empleando procesamiento de lenguaje natural (NLP) y machine learning.

- Billah Karbab, E., & Debbabi, M. (2019, 24 abril). *MalDy: Portable, data-driven malware detection using natural language processing and machine learning techniques on behavioral analysis reports*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1742287619300271

- Lansley, M., Kapetanakis, S., & Polatidis, N. (2020). SEADer++ v2: Detecting Social Engineering Attacks using Natural Language Processing and Machine Learning. 2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA). https://doi.org/10.1109/inista49547.2020.9194623

El artículo utiliza varios métodos basados en NPL y machine learnig, y realiza testeos en datasets para comprobar su efectividad y validación, los cuales se comprueba que son efectivos.
Esto nos puede ayudar a determinar un modelo basado en NPL para la detección de ataques de ingeniería social.

- Peng, T., Harris, I., & Sawa, Y. (2018). Detecting Phishing Attacks Using Natural Language Processing and Machine Learning. 2018 IEEE 12th International Conference on Semantic Computing (ICSC). https://doi.org/10.1109/icsc.2018.00056

En este artículo se basa en realizar un análisis semántico del texto para detectar intenciones maliciosas dentro de referencias de correos electrónicos.
Estos nos puede aportar información para crear modelos y validar nuestras soluciones en nuestros data sets.

Recolección de datos

Los datos que utilizaremos la obtendremos de set de datos públicos

https://www.kaggle.com/code/stevenli1/ransomware-dataset-ransomwaredataset2016-analyse/data

Este set de datos contiene información sobre 582 muestras de ransomware y 942 de aplicaciones benignas (“goodware”), lo cual no dejaría con un total de 1524 muestras para el análisis y poder verificar si un archivo es un ransomware o es software benigno. Este set de datos cuenta con varias familias de ransomware, como “CryptLocker”, “CryptoWall”, “Trojan-Ransom”, entre otras. También se cuenta con un set de características para cada ransomware, tales como invocaciones a la API, extensiones de los archivos eliminados, operaciones de archivos y las extensiones de los archivos involucrados, etc.

También se cuenta con estas clasificaciones que se utilizarán en el conjunto de datos con los SHA1 y MD5 del software analizado (tanto goodware como ransomware).

Además de este set de datos para el análisis de ransomware, también se cuenta con otro set de datos para la detección de ataques en los que se pueden ver involucrados los ransomware.

https://github.com/npolatidis/seader

En esta fuente se pueden encontrar varios sets de datos que cuentan con la información del diálogo en texto del ataque, así como también si es un ataque o no. Este cuenta con 150 muestras de ejemplos de ataques y de no ataques en formato de texto.