Ciberseguridad con Inteligencia Artificial

Dr. Vitali Herrera Semenets – CENATAV, La Habana, Cuba (vherrera@cenatav.co.cu)
MSc. Felipe Antonio Trujillo Fernández – IBERO, Ciudad de México, México (felipe.trujillo@ibero.mx)
MSc. Joshua Ismael Haase Hernández – IBERO, Ciudad de México, México (joshua.haase@ibero.mx)
Dr. Lázaro Bustio Martínez – IBERO, Ciudad de México, México (lazaro.bustio@ibero.mx)
Coordinación de Ciencia de Datos - Departamento de Estudios en Ingeniería para la Innovación – Ibero
Primavera 2024

Sesión 2

1. Introducción

En el panorama digital actual, la detección y prevención de ataques de phishing son cruciales para proteger la integridad de la información en línea. La aplicación del Procesamiento de Lenguaje Natural (PLN) permite analizar correos electrónicos en busca de indicios de actividad maliciosa, fortaleciendo así las defensas contra el fraude cibernético y salvaguardando los datos confidenciales. En esta actividad práctica, se explorará cómo emplear técnicas avanzadas de PLN para identificar patrones característicos de ataques de phishing, proporcionando una sólida defensa contra las amenazas en constante evolución del ciberespacio.

2. Objetivo

Utilizar técnicas de Procesamiento de Lenguaje Natural (PLN) para detectar y analizar patrones en correos electrónicos con el fin de identificar posibles intentos de phishing.

3. Indicaciones

a) Análisis exploratorio de datos:

- De la web del taller, descargar el dataset "*emails.zip*" que contiene correos electrónicos legítimos ("*legit.txt*") y maliciosos ("*phish.txt*") en español.

- Cargue los mensajes en un dataframe de `Pandas` cada uno. Trate de identificar los patrones que caracterizan a ambos tipos de correo electrónico mediante Análisis Exploratorio de Datos. ¿Qué se logró identificar?
- Una los dos dataframes creados anteriormente en uno y asigne las etiquetas correspondientes para identificar los mensajes de phishing de los legítimos.
- Análisis de las palabras más representativas mediante un gráfico WordCloud[1]:
    i. Utiliza la biblioteca `WordCloud` para generar el wordcloud a partir de las frecuencias de palabras.
- Interpretación y análisis:
    i. Examina el wordcloud para identificar las palabras más frecuentes en los correos electrónicos maliciosos y legítimos.
    ii. Analiza las palabras clave para inferir los temas o tópicos principales abordados en los correos electrónicos de ambos tipos.
- Aplica un algoritmo de agrupamiento (por ejemplo, KMeans) para agrupar los correos electrónicos.
    i. Visualiza los grupos obtenidos.
    ii. Entender la naturaleza de los grupos:
        1. Analiza las características de los grupos obtenidos.
        2. Identifica patrones comunes en cada grupo. Para ello, puede generar un wordcloud con los correos electrónicos de cada grupo obtenido.

b) Entrenar un modelo de clasificación:
- Divide el conjunto de datos en conjuntos de entrenamiento y prueba.
- Selecciona un algoritmo de clasificación basado en reglas (por ejemplo, RIPPER) y entrénalo utilizando el conjunto de entrenamiento.
- Evalúa el rendimiento del modelo utilizando el conjunto de prueba.
- Analice las reglas de clasificación obtenidas. ¿Qué se puede concluir de ellas?

c) Evaluación del modelo creado:
- Una vez entrenado el modelo, clasifique el conjunto de correos electrónicos "*unknown.zip*" que puede obtenerse de la web del taller.

d) Indica las conclusiones a las que has podido arribar después de realizar el ejercicio.

[1] Un WordCloud, también conocido como nube de palabras o nube de etiquetas, es una visualización que representa visualmente la frecuencia de palabras en un texto, donde las palabras más frecuentes aparecen más grandes y las menos frecuentes más pequeñas. Es una forma efectiva de resumir y visualizar la importancia relativa de diferentes términos en un conjunto de datos de texto. Los WordClouds se utilizan comúnmente en análisis de texto, minería de datos, visualización de datos y presentaciones para resaltar patrones, temas o palabras clave prominentes en un texto dado.