Universidad del Valle de Guatemala

Facultad de Ingeniería

Departamento de Ciencias de la Computación

Yongbum Park (20117)

Santiago Taracena Puga (20017)

Laboratorio 02

Detección de spam

Security Data Science

Catedrático: Jorge Yass

Introducción

El fenómeno del SPAM, caracterizado por el envío masivo de mensajes no deseados, ya sea por correo electrónico o SMS, constituye una problemática persistente en el ámbito digital. Estos mensajes no solo resultan molestos para los usuarios, sino que también pueden representar una amenaza potencial al ser utilizados como vectores para la distribución de malware, fraudes (SCAM) o para perpetrar ataques de phishing. Por lo tanto, la detección eficiente de SPAM se convierte en un desafío crucial para garantizar la seguridad y privacidad de los usuarios en el mundo digital.

En este laboratorio, nos proponemos explorar técnicas de procesamiento de lenguaje natural (NLP) y métodos clásicos de representación numérica para abordar la detección de SPAM en mensajes de texto, específicamente en el contexto de los SMS. A través de la aplicación de estas técnicas, buscamos desarrollar modelos de clasificación capaces de distinguir entre mensajes legítimos y aquellos que constituyen SPAM, con el objetivo de proporcionar herramientas efectivas para mitigar esta problemática.

En la primera parte del laboratorio, nos enfocaremos en la ingeniería de características, donde aplicaremos diversas técnicas de preprocesamiento de texto, como la conversión de minúsculas, eliminación de acentos, expansión de contracciones y eliminación de palabras comunes (stop words), entre otras. Además, exploraremos la generación de características adicionales inspiradas en técnicas avanzadas de NLP usadas en la actualidad, con el fin de enriquecer la representación de los mensajes de texto y mejorar la capacidad de los modelos para detectar patrones asociados con el SPAM.

Posteriormente, en la segunda parte del laboratorio, implementaremos modelos de aprendizaje automático utilizando algoritmos clásicos de clasificación. Utilizaremos tanto el modelo de Bag of Words (BoG) para n = 1 y n = 2, como el modelo de TF-IDF (Term Frequency-Inverse Document Frequency) para representar numéricamente los mensajes de texto. Evaluaremos el rendimiento de cada modelo en términos de métricas como la precisión, el recall y el área bajo la curva ROC, con el objetivo de identificar el enfoque más efectivo para la detección de SPAM en SMS.

A través de este laboratorio, esperamos proporcionar una comprensión más profunda de las técnicas de procesamiento de lenguaje natural y su aplicación en la detección de SPAM, así como ofrecer herramientas prácticas para abordar esta problemática en el ámbito de la seguridad de datos. Otro objetivo considerablemente importante consiste en lograr representar textos de una longitud alta como los que se encuentran presentes en el dataset de forma eficiente y fácil de aplicar para el procedimiento en el que consiste entrenar un modelo de inteligencia artificial como los que se deben utilizar.

Discusión

1. ¿Qué error es más “aceptable”: dejar pasar un SMS de SPAM (falso negativo) o bloquear un SMS legítimo (falso positivo)? Justifique su respuesta.

En términos de qué error es más "aceptable" entre dejar pasar un SMS de SPAM (falso negativo) o bloquear un SMS legítimo (falso positivo), la respuesta depende del contexto y las consecuencias asociadas con cada tipo de error. En general, ambos errores tienen implicaciones negativas, pero en el caso de la detección de SPAM, es más crítico minimizar los falsos negativos, es decir, dejar pasar un mensaje de SPAM como legítimo. Esto se debe a que los usuarios podrían verse expuestos a riesgos de seguridad, como phishing o malware, si reciben mensajes de SPAM no detectados. Por otro lado, bloquear un SMS legítimo podría generar molestias para el usuario, pero puede corregirse fácilmente y no conlleva riesgos de seguridad significativos.

2. Compare los valores para cada modelo de representación numérico. En base a la respuesta de la primera pregunta ¿Qué modelo de representación numérica produjo el mejor resultado, BoG o TF-IDF? ¿Cuál o cuáles son las razones por las que dicho modelo se comportó de mejor manera?

Al comparar los valores para cada modelo de clasificación (Random Forest y SVM), ambos muestran un buen rendimiento en términos de precisión y recall, con puntajes de precisión cercanos al 95% y puntajes de recall superiores al 75%. Sin embargo, el modelo SVM parece tener un rendimiento ligeramente mejor en términos de recall, lo que significa que es más efectivo para identificar correctamente los mensajes de SPAM. En base a la respuesta de la primera pregunta, donde se prioriza minimizar los falsos negativos, el modelo SVM podría considerarse más efectivo en este contexto.

3. En base a la exploración de datos e ingeniería de características que realizó en el primer y este laboratorio, ¿qué consejos le daría a un familiar que le solicita ayuda para detectar si un email o SMS es phishing o no? ¿En qué características de una URL/email podría fijarse su familiar para ayudarlo a detectar un potencial phishing?

Basándonos en la exploración de datos e ingeniería de características, los consejos que podríamos dar a un familiar para detectar si un email o SMS es phishing incluyen verificar la autenticidad de los remitentes, prestar atención a la ortografía y gramática deficientes, y evitar hacer clic en enlaces sospechosos o adjuntos de origen desconocido. En el caso específico de las URLs, características como dominios engañosos, subdominios inusuales, solicitudes de información personal o financiera a través de formularios en línea y enlaces que redirigen a sitios web desconocidos pueden ser indicadores de phishing.

4. Si detectamos una URL o email/SMS de phishing, ¿qué podemos hacer para detener su distribución?

Si detectamos una URL o email/SMS de phishing, es importante tomar medidas para detener su distribución y proteger a los usuarios. Esto puede incluir informar a las autoridades pertinentes, como proveedores de servicios de correo electrónico o SMS, así como a organizaciones especializadas en la detección y prevención de actividades maliciosas en línea. Además, es fundamental educar a los usuarios sobre las prácticas de seguridad digital y proporcionar herramientas efectivas de filtrado y bloqueo de contenido malicioso en sus dispositivos y plataformas de comunicación. Estas medidas ayudarán a prevenir la propagación de phishing y proteger a los usuarios contra posibles amenazas en línea.