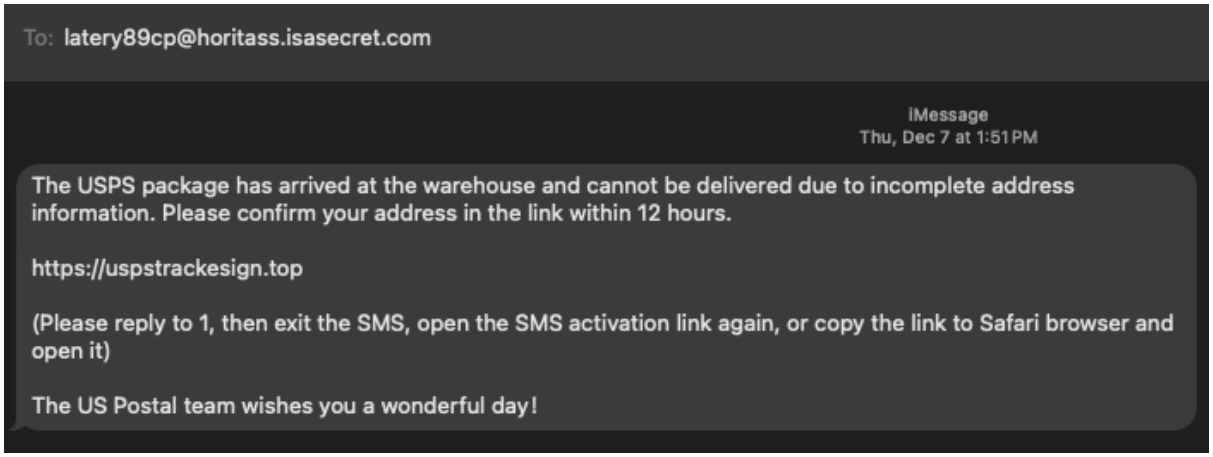
1 Objetivos

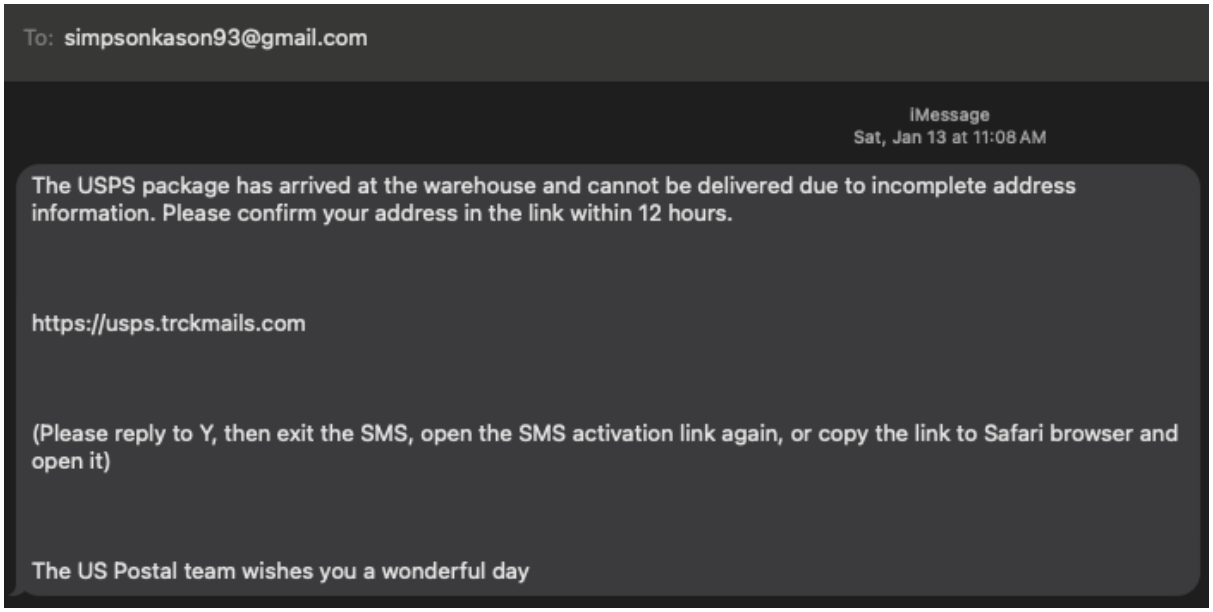
- Aplicar el uso de técnicas de pre-procesamiento para lenguaje natural
- Utilizar y evaluar método clásicos de representación numérica
- Implementar un modelo que utilice NLP para la detección de spam en SMS

2 Preámbulo

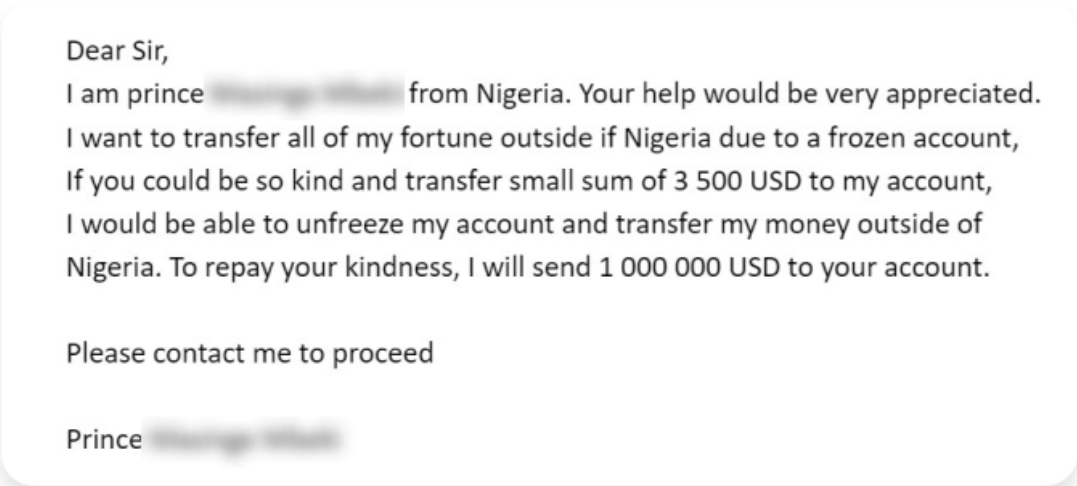
SPAM

El SPAM se refiere a mensajes no deseados o solicitados sobre publicidad de productos o servicios, que utilizan los correos electrónicos y/o SMS para ser distribuidos. Por lo general solo son molestos, pero algunos son peligrosos al combinarse con phishing, fraude (SCAM) o incluir malware. Ejemplos de phishing:





SCAM:



3 Desarrollo

El laboratorio será desarrollado en parejas. Se debe entregar un enlace a un repositorio de Github con el código fuente del pre-procesamiento, los modelos de representación de texto y otras características, y implementación de los modelos de clasificación, así como la explicación de las métricas de evaluación.

Parte 1 – Ingeniería de características

Exploración de datos y Pre-procesamiento

Aplique las técnicas de pre – procesamiento de lenguaje natural que considere necesarias (conversión de minúsculas, mayúsculas, eliminación de acentos, expansión de contracciones, eliminación de stop words, etc.)

Puede generar otras características como las expuestas en el artículo “Phishing email detection using robust nlp techniques.”

Representación de texto

Utilice los modelos de BoG (para n = 1,2) y TF-IDF. Muestre algunos ejemplos de los mensajes en su representación numérica.

Parte 2 – Implementación del modelo

Separación de datos

- Datos de entrenamiento: 70%
- Datos de prueba: 30%

Implementación

Utilice un algoritmo de ML para entrenar el modelo con cada uno de los modelos de representación numérica. Muestre los valores obtenidos para las siguientes métricas:

- Matriz de confusión
- Precision
- Recall
- Curva ROC
- AUC

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

Rúbrica

Aspecto	**Punteo (sobre 100 pts)**
Pre-procesamiento Ingeniería de características	20
Implementación completa de los dos modelos (10 pts c/u)	20
Explicación de las métricas de rendimiento para cada modelo (10 pts c/u)	20
Preguntas (10 pts c/u)	40