PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR

FACULTAD DE HÁBITAT, INFRAESTRUCTURA Y CREATIVIDAD

INGENIERÍA EN SISTEMAS

MAESTRÍA EN SISTEMA DE INFORMACIÓN, MENCIÓN DATA SCIENCE

TEMA

Análisis interpretable de patrones lingüísticos en correos electrónicos de phishing en inglés mediante aprendizaje automático supervisado

AUTORA: Gabriela Rodríguez

DIRECTOR: Mgtr. Sebastián Tamayo

QUITO – 2026

Plan Final del Anteproyecto

1. Título del proyecto

Análisis de patrones lingüísticos en correos electrónicos de phishing en inglés mediante aprendizaje automático supervisado.

2. Planteamiento del problema y pregunta de investigación

El phishing por correo electrónico continúa siendo una amenaza relevante para la seguridad de la información, debido a que utiliza mensajes diseñados para engañar al usuario y persuadirlo de realizar acciones que pueden comprometer datos personales, credenciales de acceso o información institucional. Estos correos no solo se apoyan en elementos técnicos, como enlaces o remitentes falsificados, sino también en el uso estratégico del lenguaje para generar confianza, urgencia, temor o presión en el destinatario.

A pesar de los avances en herramientas automáticas para detectar correos fraudulentos, gran parte de los enfoques existentes se ha concentrado principalmente en clasificar los mensajes como legítimos o phishing, no obstante, limitar el análisis al resultado de clasificación puede dejar sin explicar qué rasgos del lenguaje permiten diferenciar un correo legítimo de uno fraudulento. Esta situación representa una limitación para comprender cómo se construyen lingüísticamente los mensajes de phishing y qué patrones textuales pueden influir en su capacidad de engaño.

En el caso de los correos electrónicos en inglés, esta problemática resulta especialmente relevante debido a la amplia circulación internacional de mensajes fraudulentos en este idioma y a la diversidad de estrategias discursivas utilizadas por los atacantes, ya que los correos de phishing pueden recurrir a expresiones de urgencia, solicitudes de verificación, referencias a cuentas bloqueadas, promesas de beneficios, advertencias de seguridad o instrucciones aparentemente legítimas. Sin embargo, estos elementos no siempre son evidentes para los usuarios ni son analizados con suficiente profundidad desde el punto de vista lingüístico.

Por ello, el problema de investigación se centra en la necesidad de analizar qué características lingüísticas y textuales permiten diferenciar correos electrónicos legítimos y correos de phishing en inglés. Este enfoque busca ir más allá de la simple detección automática, aportando una comprensión más clara de los patrones del lenguaje asociados al phishing y de su utilidad para el análisis académico y aplicado de este tipo de amenaza.

Pregunta de investigación

¿Qué características lingüísticas y textuales permiten diferenciar los correos electrónicos legítimos de los correos de phishing en inglés mediante aprendizaje automático supervisado?

3. Objetivos generales y específicos

3.1. Objetivo general

Analizar los patrones lingüísticos y textuales que diferencian los correos electrónicos legítimos de los correos de phishing en inglés, mediante aprendizaje automático supervisado.

3.2. Objetivos específicos

- Caracterizar el conjunto de datos seleccionado.
- Preparar el corpus de correos electrónicos para el análisis.
- Identificar patrones lingüísticos y textuales en correos legítimos y de phishing.
- Implementar modelos de aprendizaje automático supervisado.
- Evaluar el desempeño de los modelos implementados.
- Analizar las características más relevantes que diferencian los correos legítimos de los correos de phishing.

4. Justificación del proyecto

El phishing por correo electrónico representa una amenaza relevante para las organizaciones, debido a que utiliza mensajes engañosos orientados a obtener credenciales, inducir pagos indebidos, comprometer cuentas o facilitar accesos no autorizados, sin embargo, estos mensajes no solo dependen de elementos técnicos, como enlaces, dominios o remitentes falsificados, sino también de recursos lingüísticos diseñados para generar confianza, urgencia, temor o presión en el destinatario.

Aunque existen filtros de correo, listas de bloqueo y controles tradicionales, los atacantes modifican constantemente el contenido de los mensajes y las estrategias de ingeniería social utilizadas, por esta razón, resulta necesario estudiar no solo si un correo puede ser clasificado como legítimo o phishing, sino también qué patrones lingüísticos y textuales permiten establecer esa diferenciación.

La presente investigación es importante porque propone analizar correos electrónicos de phishing en inglés mediante técnicas de procesamiento de lenguaje natural y aprendizaje automático supervisado. Este enfoque permitirá identificar características textuales asociadas a los mensajes fraudulentos, tales como expresiones de urgencia, solicitudes de información, estructuras persuasivas, referencias a cuentas o credenciales y otros patrones presentes en el contenido del correo.

Desde el punto de vista académico, el estudio contribuye al análisis del phishing desde una perspectiva lingüística y computacional, evitando limitarse únicamente al rendimiento de los modelos de clasificación. De esta manera, el aprendizaje automático se utiliza como una herramienta para apoyar la identificación de patrones diferenciadores entre correos legítimos y correos fraudulentos.

Desde el punto de vista aplicado, los resultados pueden servir como base para futuras investigaciones o soluciones orientadas a mejorar la comprensión y detección temprana de correos de phishing. Además, el proyecto fortalece la relación entre ciencia de datos, procesamiento de lenguaje natural y ciberseguridad, al mostrar cómo los datos textuales pueden transformarse en evidencia útil para el análisis de amenazas digitales.

5. Marco teórico

5.1. Phishing y riesgo organizacional

El phishing es una técnica de ingeniería social en la que un atacante envía mensajes que aparentan provenir de una entidad confiable para persuadir al usuario a entregar información sensible, acceder a enlaces falsos, descargar archivos maliciosos o realizar acciones que comprometan la seguridad. En un entorno organizacional, este tipo de ataque puede afectar la confidencialidad, integridad y disponibilidad

de la información, debido a que el correo electrónico sigue siendo uno de los principales canales de comunicación institucional.

Los correos phishing suelen presentar señales como lenguaje de urgencia, solicitudes de verificación de cuenta, enlaces acortados o sospechosos, remitentes similares a dominios legítimos, errores de redacción y uso de marcas conocidas para generar confianza. Sin embargo, estas señales no siempre son evidentes, por lo que resulta conveniente utilizar métodos automáticos que aprendan patrones a partir de datos históricos.

5.2. Procesamiento de lenguaje natural aplicado a correos electrónicos

El procesamiento de lenguaje natural, conocido como NLP, permite tratar datos textuales para que puedan ser analizados por algoritmos computacionales. En el caso de correos electrónicos, las tareas básicas incluyen limpieza de texto, normalización, eliminación de caracteres especiales, conversión a minúsculas, tokenización, eliminación de palabras vacías y revisión de términos frecuentes. Estas técnicas ayudan a reducir ruido y a conservar información relevante para el análisis.

En esta investigación, el NLP se utilizará para convertir el contenido de los correos en una representación estructurada. Además del cuerpo del mensaje, se podrán considerar elementos como asunto del correo, longitud del texto, presencia de enlaces, palabras asociadas a urgencia y frecuencia de términos. Estas características pueden contribuir a diferenciar mensajes legítimos de mensajes phishing.

5.3. TF-IDF como técnica de representación textual

TF-IDF, por sus siglas en inglés Term Frequency-Inverse Document Frequency, es una técnica de vectorización que asigna un peso a cada término de acuerdo con su frecuencia dentro de un documento y su relevancia en relación con el conjunto completo de documentos. En términos simples, una palabra tendrá mayor importancia si aparece con frecuencia en un correo, pero no aparece de manera común en todos los correos del dataset.

Esta técnica es útil para la clasificación de correos porque permite transformar texto en variables numéricas que pueden ser procesadas por modelos de Machine Learning. En un problema de phishing, TF-IDF puede resaltar términos relacionados con verificación, contraseña, bloqueo de cuenta, pago, enlace, urgencia u otras expresiones que aparecen con mayor frecuencia en mensajes maliciosos.

5.4. Clasificación supervisada

La clasificación supervisada es una técnica de Machine Learning que utiliza datos previamente etiquetados para entrenar un modelo capaz de asignar una clase a nuevos registros. En este proyecto, la clasificación será binaria, ya que cada correo será clasificado como legítimo o phishing. Para ello, se propone comparar algoritmos como Naive Bayes, Regresión Logística, Random Forest y Support Vector Machine.

Naive Bayes suele ser utilizado en clasificación de texto por su eficiencia y simplicidad. La Regresión Logística permite estimar la probabilidad de pertenencia a una clase. Random Forest combina múltiples árboles de decisión para mejorar la estabilidad del modelo. Support Vector Machine puede ser útil cuando se trabaja con espacios de alta dimensionalidad, como ocurre con los vectores generados por TF-IDF.

6. Metodología de investigación

6.1. Enfoque, alcance y diseño

La investigación tendrá un enfoque cuantitativo, una finalidad aplicada, un diseño no experimental y un alcance exploratorio-comparativo con componente interpretativo. Será cuantitativa porque utilizará datos textuales transformados en variables analizables y métricas numéricas para evaluar el desempeño de modelos de aprendizaje automático supervisado. Tendrá una finalidad aplicada porque aborda un problema vinculado con el análisis de correos electrónicos de phishing y la seguridad de la información.

El diseño será no experimental, debido a que no se manipularán usuarios, organizaciones ni entornos reales de correo electrónico; el análisis se realizará a partir de un conjunto de datos previamente disponible. El alcance será exploratorio-comparativo porque permitirá identificar patrones lingüísticos y textuales en correos legítimos y de phishing, así como comparar el desempeño de distintos modelos supervisados. Finalmente, incorporará un componente interpretativo porque no se limitará a reportar métricas de clasificación, sino que buscará analizar qué características textuales contribuyen a diferenciar ambos tipos de correos.

6.2. Fuente de datos

La fuente principal será el dataset público Human-LLM generated phishing-legitimate emails, disponible en Kaggle.

6.7. Consideraciones éticas y de seguridad

El estudio utilizará datos públicos y no intervendrá en usuarios reales. No se publicarán ejemplos completos de correos que contengan posible información personal, credenciales, enlaces activos o datos sensibles. Los resultados se presentarán con fines académicos y no como una herramienta lista para producción.

7. Cronograma de actividades

El cronograma se organiza en ocho semanas de trabajo, siguiendo una secuencia lógica basada en las etapas de comprensión del negocio, comprensión de los datos, preparación de datos, modelado, evaluación y presentación de resultados. Esta estructura permite visualizar el avance del proyecto desde junio S3 hasta agosto S2, según las fechas indicadas para el desarrollo de la investigación.

Fase	Junio S3	Junio S4	Julio S1	Julio S2	Julio S3	Julio S4	Agosto S1	Agosto S2
Comprensión del negocio	x							
Comprensión de datos		x	x					
Preparación de datos			x	x				
Modelado				x	x	x		
Evaluación							x	
Presentación de resultados								x

8. Recursos

Categoría	Recurso	Descripción
Humanos	Tesista	Responsable de descarga de datos, preparación, modelado, análisis y redacción.
Humanos	Director de tesis	Orientación metodológica, revisión de avances y validación académica.
Técnicos	Python	Lenguaje principal para procesamiento y modelado.
Técnicos	Jupyter Notebook o Google Colab	Ambiente de experimentación reproducible.
Técnicos	Pandas y NumPy	Manipulación y análisis de datos.
Técnicos	Scikit-learn	Pipelines, TF-IDF, modelos y métricas.
Técnicos	NLTK o spaCy	Procesamiento de texto y apoyo lingüístico.
Técnicos	Matplotlib o Seaborn	Visualización de resultados.
Técnicos	Zotero	Gestión bibliográfica.
Financieros	Internet y equipo informático	Descarga de dataset, consulta bibliográfica y desarrollo experimental.
Financieros	Impresión y empaste	Entrega final del documento, si aplica.