TRABAJO DE GRADO
Opción Seminario

Sistema de detección de Correos Phishing basado en big data

Corporación Universitaria Remington.

Facultad de Ingeniería

Especialización en Ciberseguridad

Cristian David Sánchez García

Tutor:

Juan Pablo Vélez

Opción de Trabajo de grado Seminario Big Data

2025

Resumen

Mediante técnicas de Big Data y el aprendizaje automático se desarrolló el siguiente proyecto, donde el objetivo principal es lograr hacer una eficiente identificación de correos electrónicos tipo Phising. El siguiente planteamiento tuvo como objetivo realizar una aplicación prototipo que permitió implementar una identificación manual de correos tipo Phishing, ya que esta sigue siendo una de las amenazas más frecuentes a los usuarios de internet.

El identificador de correos tipo Phising se realizó utilizando Python y Term Frequency-Inverse Document Frequency (TF-IDF) para la captura de características del contenido dentro de los correos electrónicos, transformando este contenido en una representación numérica para su análisis. Después, se entrenó un modelo Multinomial Naive Bayes, que se caracteriza por su eficiencia en tareas de clasificación de texto, ya sea en datasets pequeños o medianos y por último se diseñó una interfaz gráfica inicial con Streamlit, para que cualquier persona pueda hacer uso de la aplicación.

Este es un prototipo inicial, por lo que se plantea que pueda ser escalable en el tiempo utilizando técnicas de Big Data para capturar grandes cantidades de información, almacenarla y analizarla. Los resultados que se presentan en el siguiente documento son evidencia que incluso haciendo uso de modelos sencillos con el correcto uso de las herramientas, podemos identificar correos electrónicos tipo Phising y garantizar la seguridad de la información; ya sea personal o empresarial.

Palabras clave

Big data, Ciberseguridad, Machine Learning, Phishing, Clasificador de correos.

Marco conceptual

Fundamentos del Phishing y Big Data

Ingeniería social

La ingeniería social se conoce como cualquier acto que influye de forma positiva o negativa sobre una persona. En el ámbito informático, (Mitnick, K., & Simon, W. 2001), en su libro *El arte del engaño*, explica la utilización de técnicas implementadas para engañar a un usuario y que así este comparta información sensible, generando una brecha de seguridad insospechada. Existen en la actualidad diferentes tipos de técnicas de Ingeniería social en el ámbito informático; tales como: Baiting, Phishing, Quid Pro Quo, Pretexting.

Desde una perspectiva más humanista, (Hadnagy, C 2018) explica que la ingeniería social no se centra únicamente en el uso de herramientas tecnológicas, sino en las habilidades del atacante para influir en la toma de decisiones de las personas. Según el autor, estas técnicas combinan principios de comunicación, persuasión y psicología con el fin de manipular a los usuarios y lograr que realicen acciones que normalmente no harían, como revelar información confidencial o acceder a enlaces maliciosos

Phishing

El Phishing se define como la práctica en la cual se envían correos electrónicos que parecieran ser de fuentes confiables con el fin de obtener información confidencial (Hadnagy & Fincher, 2015). Esta técnica es usada desde hace más de 30 años, donde personas se hacían pasar por empleados de una compañía de Internet de Estados Unidos para obtener

credenciales de inicio de sesión y así mismo obtener Internet gratis y enviar correos electrónicos de spam desde la cuenta de la víctima (ESET, 2025).

Diversos informes de seguridad confirman que el phishing continúa siendo una de las técnicas más utilizadas por los ciberdelincuentes. De acuerdo con el *Data Breach Investigations Report* de Verizon (2023), una gran proporción de los incidentes de seguridad inicia con correos engañosos diseñados para hacer que el usuario dé acceso a información sensible, lo que refuerza la necesidad de contar con mecanismos automatizados de detección.

Big Data

(Según Erl, Khattak y Bühler 2016), cuando hablamos de Big Data, no solo hablamos de grandes volúmenes de datos, sino también de otro tipo de características como su variedad, velocidad, valor generado a partir del contenido que puede ser texto, imágenes, sensores o registros web. Por lo que esto también implica una transformación en la forma cómo las empresas realizan el manejo de la información, para garantizar la seguridad y obtener resultados eficientes.

Por otra parte, (López Fandiño 2023) describe Big Data como un conjunto de diversas tecnologías y herramientas que garantizan el manejo de volúmenes masivos de información para así convertirlos en conocimiento útil por medio del análisis en tiempo real.

De acuerdo a (Mayer-Schönberger y Cukier 2013), el Big Data genera un cambio en la manera de capturar y analizar la información, ya que brinda la oportunidad de trabajar con grandes volúmenes de datos para identificar tendencias que quizás antes pasaban desapercibidas. Ellos indican que el valor de Big Data no radica únicamente en la precisión de los datos, sino que va más allá, ya que también es indispensable poder aportar a la toma de decisiones por medio de su análisis y así anticipar comportamientos, incluso cuando no se conoce la causa exacta de los problemas analizados.

Ciberseguridad

La ciberseguridad, según (Guttman, B. and Roback, E. 1995) en *An Introduction to Computer Security: The NIST Handbook*, hacen referencia al conjunto de acciones y medidas que buscan proteger los sistemas y la información frente a accesos no permitidos. Allí se menciona que la seguridad empieza por algo tan básico como identificar esas posibles fallas de seguridad y por qué. El NIST insiste en que la seguridad no es un producto final que se instala, sino un proceso de mejoramiento continuo que necesita monitoreo, ajustes y una correcta gestión de los recursos tecnológicos y humanos.

Por otro lado, *Cybersecurity Essentials* brinda una visión más actual del problema, explicando que la ciberseguridad debe abarcar no solo los equipos y las redes, sino también a las personas que hacen uso de ellos. (Brooks, C., Grow, C., & Craig, P. 2018) ya que las amenazas actuales son más complejas y son debidas tanto a fallos técnicos como a errores humanos, por lo que las medidas de protección deben ser transversales. Allí mencionan medidas de protección donde la defensa en capas, la gestión de accesos, la detección de anomalías en el tráfico y la respuesta ante incidentes, garantizan la seguridad combinando buenas prácticas, conocimientos técnicos y una cultura organizacional orientada a minimizar riesgos.

Es así como la ciberseguridad puede entenderse como el conjunto de principios, prácticas y tecnologías orientadas a proteger los sistemas de información, las redes y los datos frente a accesos no autorizados e intencionales. Según (Stallings 2018), la ciberseguridad pretende defender los siguientes pilares fundamentales: la confidencialidad, la integridad y la disponibilidad de la información. Con esto se busca garantizar que los datos sólo sean accesibles por usuarios autorizados por una empresa.

Desde una perspectiva técnica, el National Institute of Standards and Technology explica un enfoque integral de la ciberseguridad en las cuales sus funciones son identificar, proteger, detectar, responder y recuperar (NIST, 2018).Entendiendo la ciberseguridad como un proceso continuo, en el cual la detección temprana de amenazas es fundamental.

Machine Learning

El Machine Learning se define como un conjunto de métodos que permiten que un sistema aprenda a partir de datos en lugar de depender únicamente de instrucciones programadas. (Mitchell, T. M. 1997) lo describe como la capacidad de un programa para optimizar su desempeño en tareas específicas como consecuencia de la experiencia acumulada. Con esta definición, podemos profundizar que el aprendizaje no surge por simple codificación, sino también es debido a la exposición del modelo a ejemplos reales que le permiten reconocer patrones, corregir errores y ajustar sus resultados.

Por otro lado, (Bishop, C. M. 2006), ahonda en el concepto desde una perspectiva más matemática, explicando que los algoritmos de Machine Learning buscan identificar regularidades dentro de grandes volúmenes de información, utilizando modelos probabilísticos y estadísticos que ayudan a interpretar datos complejos como texto, imágenes, señales o registros históricos. Para Bishop, el proceso de aprendizaje consiste en aproximar funciones que permitan predecir o clasificar información nueva basándose en lo aprendido previamente.

(Goodfellow, Bengio y Courville 2016) indican que el Machine Learning permite a los sistemas construir modelos capaces de aprender representaciones complejas a partir de los datos, especialmente en escenarios donde el volumen de información es elevado, ya que los

algoritmos ajustan sus parámetros de forma automatizada para mejorar su desempeño en tareas como clasificación, predicción o reconocimiento de patrones**.**

Objetivos

Objetivo general

Desarrollar una aplicación para la detección de correos electrónicos de Phishing utilizando técnicas de Big Data

Objetivos específicos

- Construir un dataset inicial de correos electrónicos por medio de un conjunto de mensajes etiquetados (phishing y seguros) para el entrenamiento del modelo, con el fin de disponer de datos adecuados para la fase de aprendizaje automático.
- Implementar técnicas de preprocesamiento de texto aplicando métodos como la vectorización TF-IDF para transformar el contenido de los correos en atributos numéricos que puedan ser analizados por el modelo.
- Entrenar un modelo de Machine Learning supervisado y ajustar el clasificador Naive Bayes u otro algoritmo similar para identificar patrones asociados al phishing dentro del dataset.
- Diseñar una interfaz gráfica funcional con Streamlit que permita a los usuarios ingresar o pegar el contenido de un correo electrónico y obtener la clasificación del sistema de forma clara y rápida.
- Validar el funcionamiento del sistema con diferentes textos, verificar su comportamiento y evaluar su capacidad para identificar correctamente correos maliciosos frente a correos legítimos.

Marco contextual

Planteamiento del problema

En Colombia, el Phishing se ha consolidado como una de las amenazas más frecuentes en los ciudadanos, organizaciones públicas y privadas, y es por eso que a medida que aumenta la digitalización en el país en cuanto a servicios financieros, trámites gubernamentales y actividades comerciales, también aumenta la posibilidad a posibles fraudes para hacer que los usuarios brinden sus contraseñas, datos bancarios o información sensible.

El Centro Cibernético Policial, a través del CAI Virtual, registró en su *Balance Anual 2024* un total de 11.866 incidentes atendidos, de los cuales 1.359 corresponden a casos de phishing, ubicándose entre las modalidades más recurrentes del cibercrimen en el país (Centro Cibernético Policial, 2024).

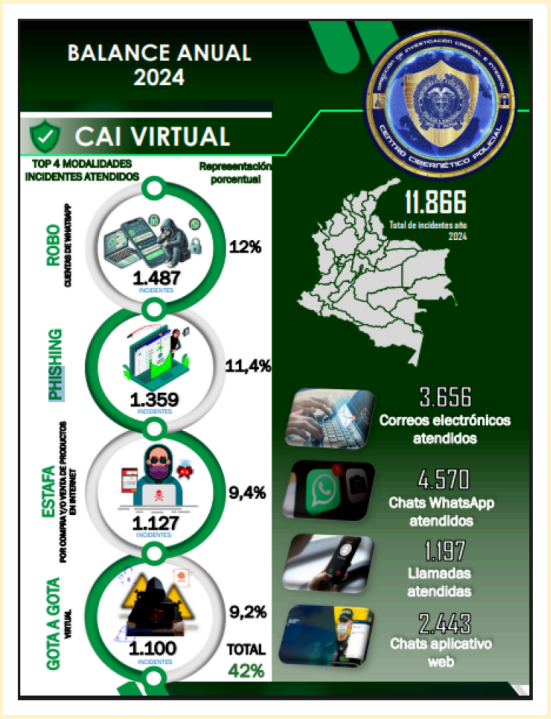

Figura 1: Infografía Balance Anual 2024

Fuente: Centro Cibernético Policial, 2024.

Además, las autoridades atendieron 3.656 correos electrónicos reportados por sospecha de fraude, lo que confirma que el correo sigue siendo el medio favorito para los ataques.

Medios nacionales como *El Colombiano* e *Infobae* han visibilizado el incremento sostenido de estas prácticas engañosas, destacando que Colombia se encuentra entre los países más afectados en Latinoamérica, según estudios de Kaspersky y ESET (Infobae, 2024; El Colombiano, 2024). Estos estudios demuestran que los delincuentes han perfeccionado sus métodos, creando mensajes cada vez más sofisticados que imitan bancos, entidades del Estado e incluso plataformas de comercio electrónico.

Hay un vacío crítico, ya que muchas personas no tienen a la mano herramientas que les permitan identificar este tipo de correos tipo Phishing, y en el caso de las empresas, la mayoría no cuentan con herramientas que les permitan garantizar la seguridad de la información o en este caso, analizar grandes volúmenes de correo de forma automatizada. Allí es donde las técnicas de Big Data y Machine Learning pueden unirse con la ciberseguridad para entrar a solucionar este tipo de falencias.

Por esa razón, nace la necesidad de diseñar una aplicación que permita clasificar correos electrónicos como phishing o seguros, aprovechando modelos de aprendizaje automático entrenados con grandes volúmenes de datos. Esta solución contribuiría a reducir el riesgo de fraudes, fortalecer la seguridad digital y apoyar los esfuerzos nacionales en la prevención del cibercrimen.

Construcción del clasificador

Dataset y procesamiento de datos

Para entrenar el sistema, se construyó un dataset pequeño compuesto por seis ejemplos de correos electrónicos. Tres de ellos fueron clasificados como phishing debido a su uso de urgencia, solicitudes de verificación o cambios de contraseña, mientras que los otros tres corresponden a comunicaciones laborales legítimas. Este conjunto inicial se utilizó con fines demostrativos, pero puede ampliarse fácilmente incorporando datasets públicos como Enron Email Dataset, Nazario Phishing Corpus, o repositorios de Kaggle.

El dataset fue procesado aplicando limpieza básica del texto, normalización y preparación para su vectorización. Cada correo fue asignado a una etiqueta binaria: 1 = phishing, 0 = seguro, permitiendo que el modelo aprenda patrones asociados a cada categoría.

Metodología utilizada

- **Se instala Python y se abre la terminal o PowerShell**
- En Windows, se presiona Win + R, y se escribe cmd o powershell
- Se navega hasta la carpeta donde se van a poner los archivos del proyecto (o crea una nueva carpeta para el proyecto).
- **Se crea la carpeta del proyecto**
- Se navega a la carpeta con: cd ruta\de\la\carpeta\phishing-detector ,
- Código de la aplicación:
    - import streamlit as st
    - from sklearn.feature_extraction.text import TfidfVectorizer
    - from sklearn.naive_bayes import MultinomialNB
    - correos = [

```
    "Urgente: verifique su cuenta ahora",
    "Su banco ha detectado actividad sospechosa",
    "Actualice su contraseña inmediatamente",
    "Reunión a las 10 AM",
    "Aquí está el informe solicitado",
    "La factura de este mes está adjunta"
]
labels = [1, 1, 1, 0, 0, 0]

vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(correos)
modelo = MultinomialNB()
modelo.fit(X, labels)
st.title("Detector Básico de Correos Phishing 🛡️")
st.write("Escribe un texto o pega un correo electrónico completo:")

texto = st.text_area("Contenido del correo:")

if st.button("Analizar"):
    X_input = vectorizer.transform([texto])
    pred = modelo.predict(X_input)[0]
    if pred == 1:
        st.error("⚠️ Este correo parece PHISHING.")
    else:
        st.success("✔️ Este correo parece SEGURO.")
```

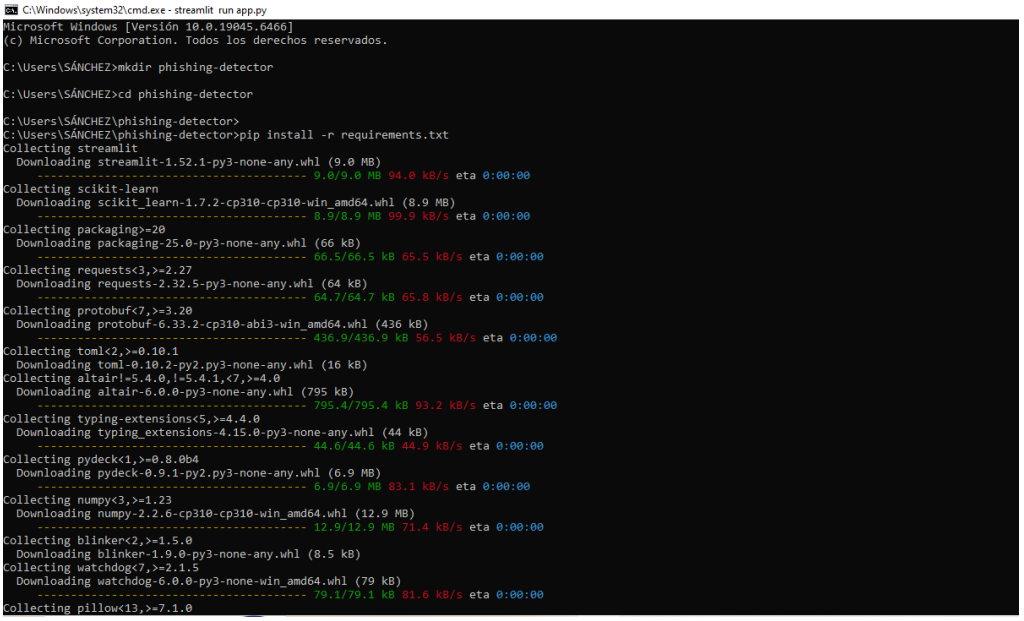

Figura 2: Imagen 1 Python

Fuente: Elaboración propia (2025)

La figura 2, corresponde a la instalación de los requerimientos exigidos por Python.

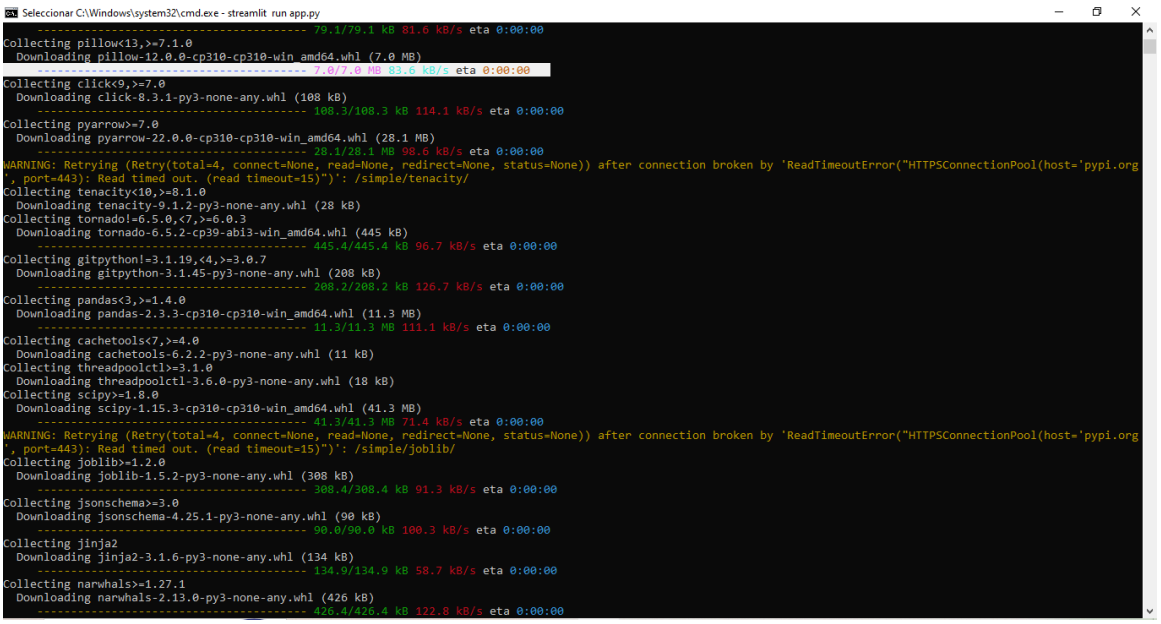

Figura 3: Imagen 2 Python

Fuente: Elaboración propia (2025)

La figura 3, corresponde a la instalación de los requerimientos exigidos por Python.

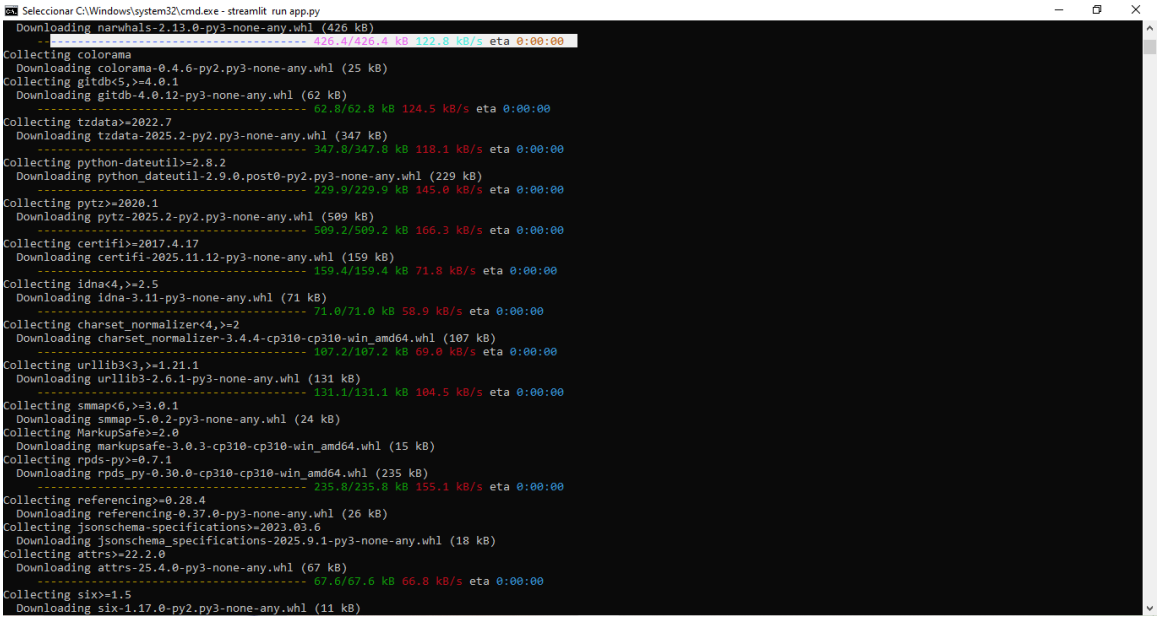

Figura 4: Imagen 3 Python

Fuente: Elaboración propia (2025)

La figura 4, corresponde a la instalación de los requerimientos exigidos por Python.

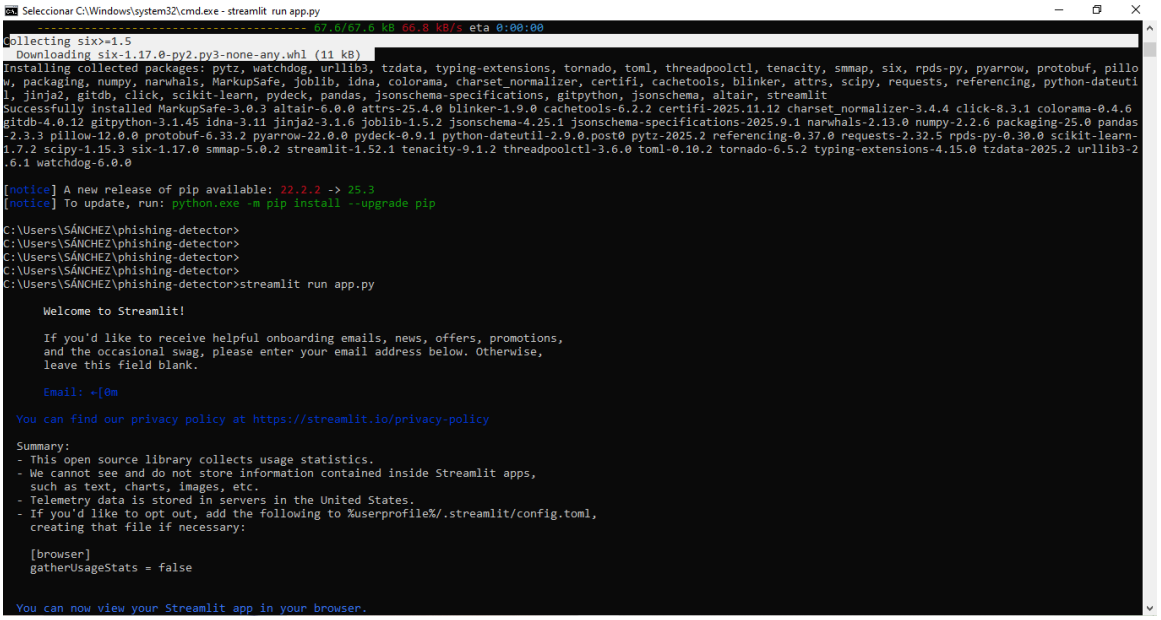

Figura 5: Imagen 4 Python

Fuente: Elaboración propia (2025)

La figura 5, corresponde a los permisos accedidos por Python para abrir Streamlit en el navegador.

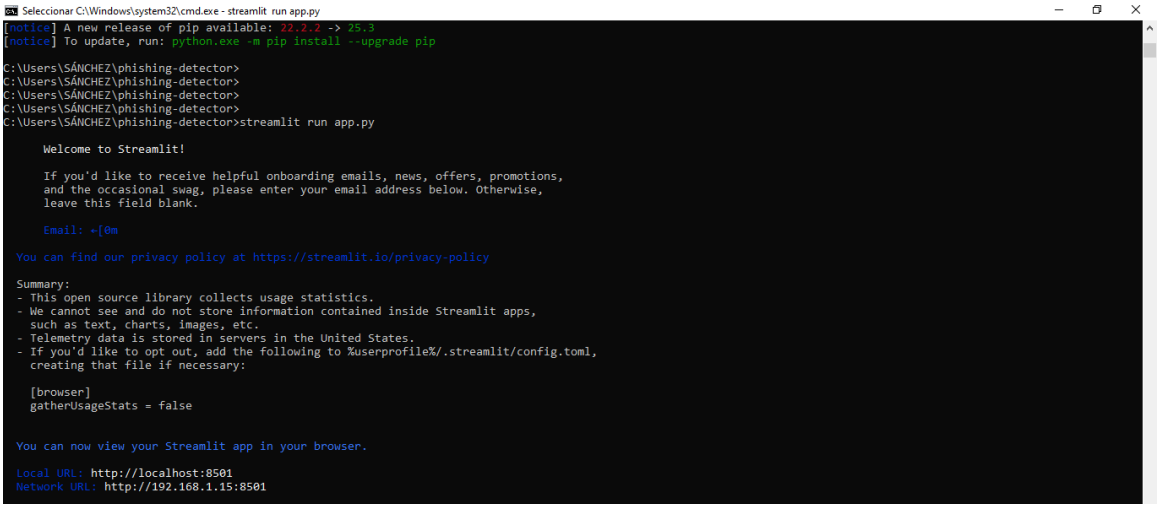

Figura 6: Imagen 5 Python

Fuente: Elaboración propia (2025)

La figura 6, corresponde a las formas posibles de acceder al detector de correos Phishing; desde el localhost o una dirección IP específica.

Extracción de características

Para representar el contenido textual de los correos, se utilizó el método *Term Frequency–Inverse Document Frequency* (TF-IDF), disponible en la librería *scikit-learn*. Este enfoque convierte las palabras en valores numéricos ponderados según su importancia dentro del conjunto de correos.

TF-IDF facilita que el modelo identifique términos frecuentes en mensajes fraudulentos, como *"verifique", "urgente", "actividad sospechosa", "actualice su contraseña"*, diferenciándolos de expresiones propias de correos formales o laborales.

El resultado es una matriz numérica que alimenta directamente al modelo de clasificación.

Metodología utilizada:

- requirements.txt
    - Se crea un archivo llamado requirements.txt con este contenido:

    - streamlit
    - scikit-learn

- **Se instalan las dependencias**
- En la terminal (con la carpeta del proyecto abierta), se ejecuta:
    - pip install -r requirements.txt
    - Esto instalará Streamlit y scikit-learn, que necesita el proyecto.

- **Se ejecuta la app**
- streamlit run app.py

Indicadores incluidos:

- **Urgencia:** “24 horas”, “suspensión”.
- **Suplantación:** Marca bancaria genérica.
- **URLs dudosas:** dominio “secure-login.example” (dominio reservado).
- **Autenticación fallida:** SPF/DKIM/DMARC simulados en fail.

Entrenamiento del modelo y métricas de evaluación

El algoritmo utilizado fue Multinomial Naive Bayes, una técnica ampliamente aplicada en clasificación de texto por su rapidez, bajo costo computacional y buen desempeño en datasets pequeños.

El modelo fue entrenado con las características generadas por TF-IDF y las etiquetas correspondientes.

En este prototipo no se calcularon métricas debido al tamaño reducido del dataset, pero en una versión ampliada se pueden emplear:

- Accuracy
- Matriz de confusión

Estas métricas permiten medir el rendimiento del sistema cuando se utilizan miles o millones de correos, tal como ocurre en escenarios reales de Big Data.

Metodología utilizada:

- Prueba la app
- Escribe o pega texto de un correo sospechoso y pulsa **Analizar**, el sistema dirá si parece phishing o no.

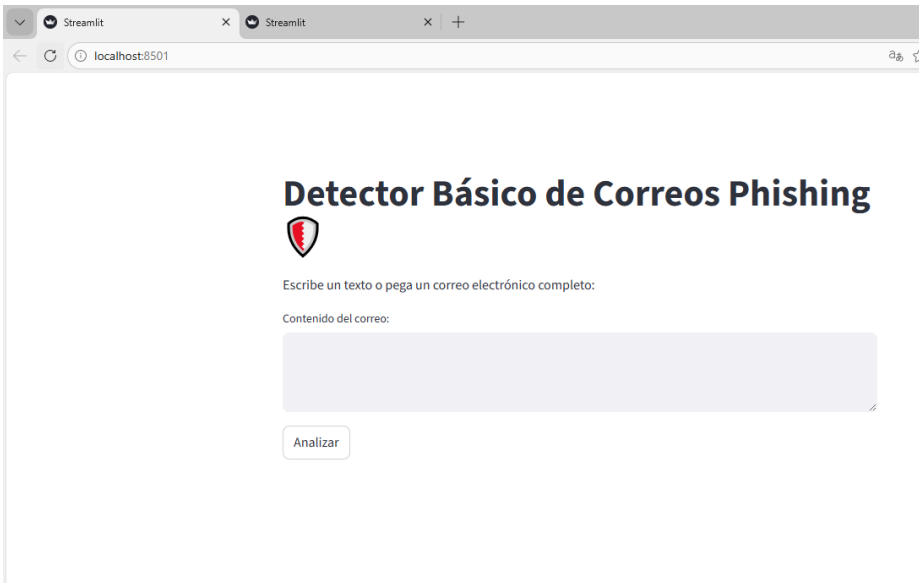

Figura 7: Imagen aplicación detector de correos electrónicos Phishing por medio de localhost

Fuente: Elaboración propia (2025)

La figura 7, corresponde a la visualización de la aplicación desde el navegador por medio de localhost.

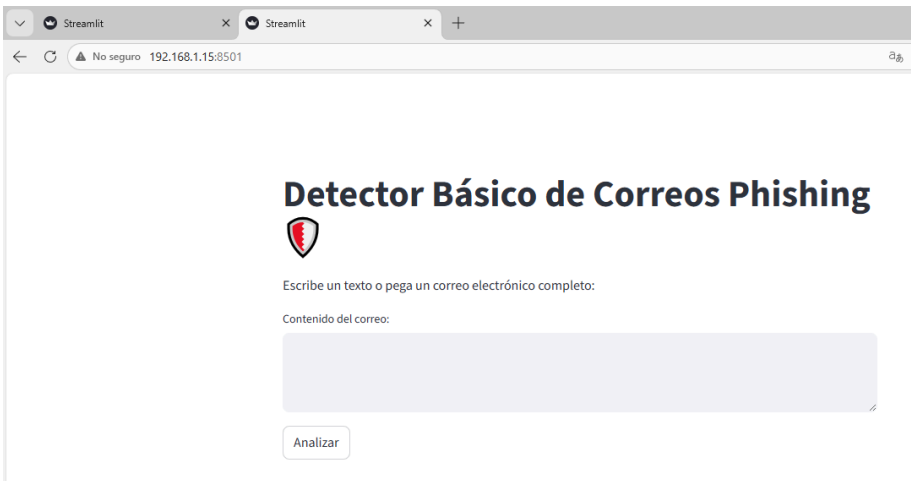

Figura 8: Imagen aplicación detector de correos electrónicos Phishing por medio de dirección IP

Fuente: Elaboración propia (2025)

La figura 8, corresponde a la visualización de la aplicación desde el navegador por medio de la IP: 192.168.1.15:8501

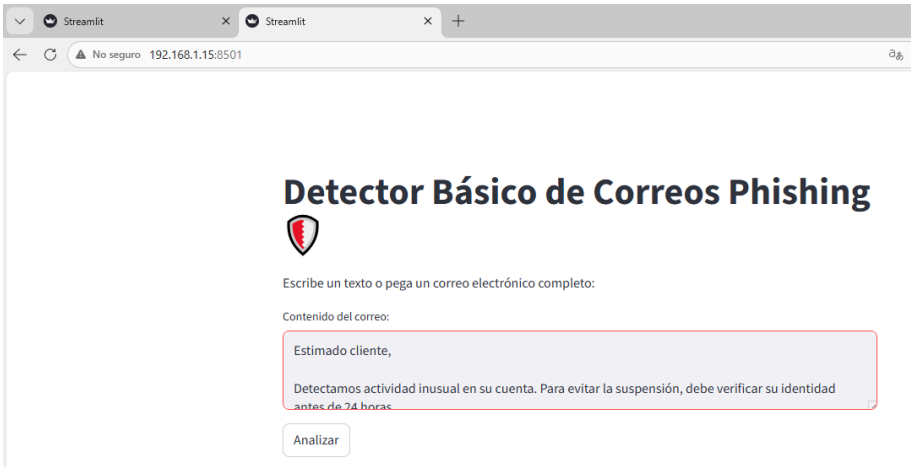

Figura 9: Imagen aplicación detector de correos electrónicos Phishing prueba

Fuente: Elaboración propia (2025)

La figura 9, corresponde a la visualización de la aplicación desde el navegador realizando una prueba con un texto de correo electrónico aparentemente engañoso.

Mensaje utilizado en la aplicación:

Estimado cliente,

Detectamos actividad inusual en su cuenta. Para evitar la suspensión, debe verificar su identidad antes de 24 horas.

Verifique aquí: hxxps://banco-nacional.secure-login.example/test?session=748192

(Nota: enlace desactivado para pruebas)

Si no completa el proceso, su acceso será restringido.

Equipo de Seguridad — Banco Nacional

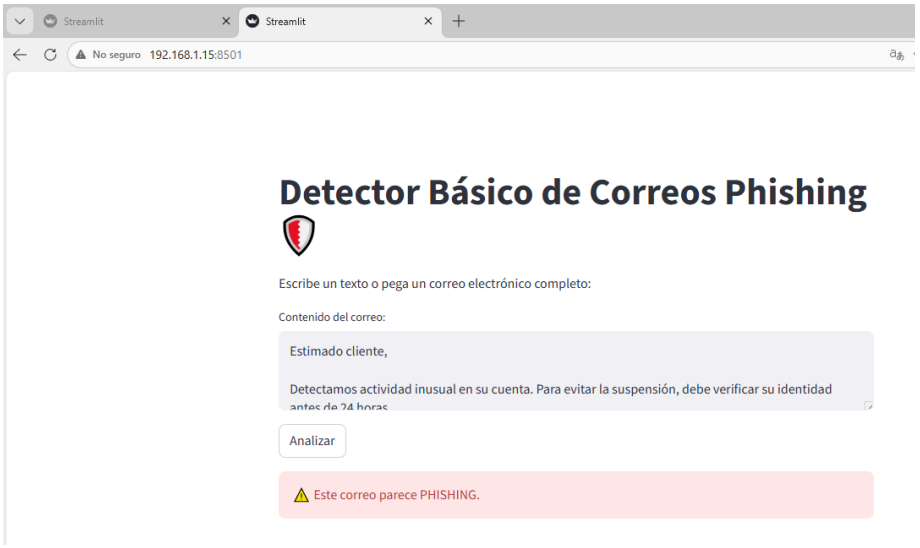

Figura 10: Imagen aplicación detector de correos electrónicos Phishing resultado

Fuente: Elaboración propia (2025)

La figura 10, , corresponde a la visualización de la aplicación después de dar clic en el botón *analizar*.

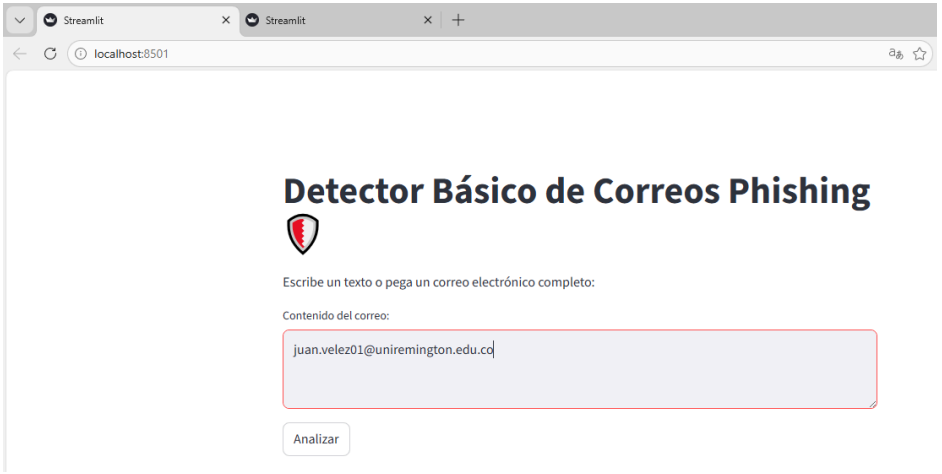

Figura 11: Imagen aplicación detector de correos electrónicos Phishing prueba

Fuente: Elaboración propia (2025)

La figura 11, corresponde a la visualización de la aplicación desde el navegador realizando una prueba con un correo electrónico para saber si es seguro o Phishing.

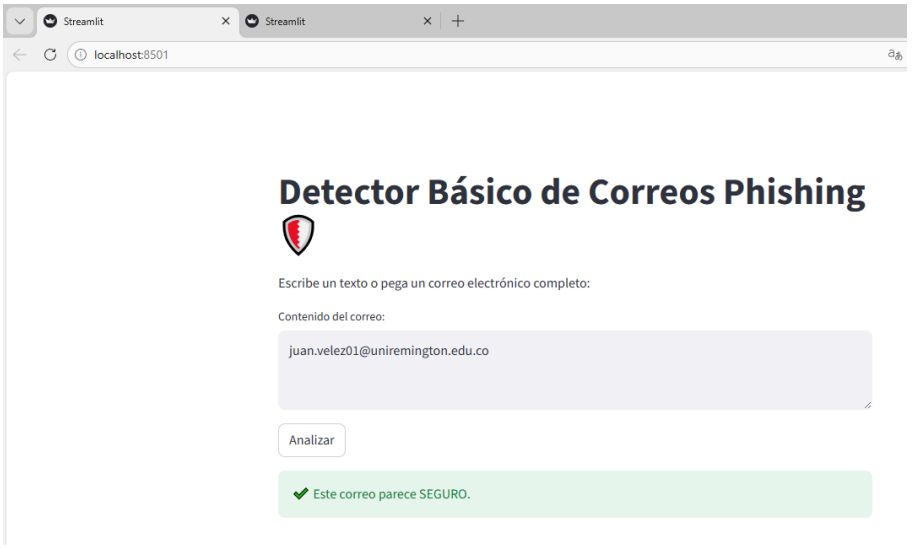

Figura 12: Imagen aplicación detector de correos electrónicos Phishing resultado.

Fuente: Elaboración propia (2025)

La figura 12, corresponde a la visualización de la aplicación después de dar clic en el botón *analizar*, en el cual se puede observar que es un correo electrónico seguro.

Arquitectura Big Data utilizada

Aunque el prototipo funciona en local, la estructura del proyecto está pensada para escalar a una arquitectura tipo Big Data. La propuesta incluye:

- Ingesta masiva de correos desde servidores corporativos o proveedores de correo.
- Procesamiento distribuido utilizando herramientas como Data Warehouse en caso de manejar volúmenes muy grandes.
- Almacenamiento distribuido, por ejemplo HDFS o bases NoSQL como MongoDB.

- Entrenamiento y actualización periódica del modelo usando pipelines automáticos.
- Interfaz de análisis en tiempo real desplegada con Streamlit o un servidor web.

Esta arquitectura permite que el sistema pase de analizar unos pocos correos a analizar miles por minuto, alineándose con el enfoque Big Data del proyecto.

Conclusiones

El identificador de correos tipo Phishing brinda un entendimiento mucho mayor de cómo Big Data, el aprendizaje automático y la ciberseguridad pueden trabajar en total sincronía para enfrentar este problema que está en crecimiento no solo en Colombia, sino también en el mundo. Este sistema demuestra que los métodos tradicionales no son suficientes para garantizar la seguridad, sino que cada vez se necesitan herramientas con técnicas que permitan procesar un gran volumen de datos en poco tiempo.

El desarrollo de esta aplicación demostró que incluso un modelo sencillo puede garantizar resultados efectivos para la detección de este tipo de fraudes y que no es necesario el uso de modelos complejos, sino de una correcta representación del lenguaje y datasets bien estructurados y continuamente actualizados. Así mismo, esta herramienta se puede escalar con el tiempo a entornos que permitan analizar miles de correos simultáneamente, almacenarlos de manera distribuida y entrenar modelos con volúmenes mucho más amplios.

Para finalizar, este trabajo justifica y refuerza la idea de lo importante que es que las empresas colombianas empiecen a hacer uso de este tipo de herramientas de Machine Learning y Big Data para garantizar la seguridad de la información para así, reducir riesgos,

prevenir fraudes y optimizar la capacidad de respuesta, abriendo la puerta a sistemas más inteligentes, adaptativos y alineados con los desafíos actuales del entorno digital.

Referencias

- Hadnagy, C., & Fincher, M. (2015). *Phishing Dark Waters: The Offensive and Defensive Sides of Malicious Emails*. Wiley.
- Hadnagy, C. (2018). *Social engineering: The science of human hacking* (2nd ed.). Wiley.
- ESET. (2025). *¿Qué es phishing? Guía completa 2025*. https://www.eset.com/latam/blog/cultura-y-seguridad-digital/que-es-phishing-guia-completa-2025/
- Mitnick, K., & Simon, W. (2001). *El arte del engaño: Controlando el factor humano en la seguridad*. Wiley.
- Verizon. (2023). *Data breach investigations report (DBIR)*. Verizon Enterprise. https://www.verizon.com/business/resources/Ta5a/reports/2023-dbir-public-sector-snapshot.pdf
- Erl, T., Bühler, P., & Khattak, W. (2016). *Big Data Fundamentals: Concepts, Drivers & Techniques*. Prentice Hall.
- López Fandiño, V. M. (2023). *Sistemas de Big Data*. Ra-Ma Editorial.
- Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt.
- Guttman, B. and Roback, E. (1995), An Introduction to Computer Security: the NIST Handbook, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, [online], https://doi.org/10.6028/NIST.SP.800-12r1

- National Institute of Standards and Technology. (2018). *Framework for improving critical infrastructure cybersecurity*. NIST.
- Stallings, W. (2018). *Computer security: Principles and practice* (4th ed.). Pearson.
- Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Brooks, C. J., Grow, C., & Craig, P. (2018). *Cybersecurity Essentials*. Wiley / Sybex.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Centro Cibernético Policial. (2024). *Balance anual CAI Virtual 2024* [Infografía]. Policía Nacional de Colombia.
- El Colombiano. (2024). *Aumentan los casos de fraudes digitales y phishing en Colombia*. https://www.elcolombiano.com
- Infobae. (2024). *Colombia entre los países más afectados por ataques de phishing según expertos*. https://www.infobae.com
- Kaspersky. (2023). *Reporte de amenazas en América Latina 2023*. https://latam.kaspersky.com
- ESET. (2024). *Panorama del phishing en Latinoamérica 2024*. https://www.eset.com/latam