Trabajo de grado

Asistente de seguridad que analice correos, mensajes y llamadas en tiempo real para detectar intentos de phishing, estafas o ataques de ingeniería social

Corporación Universitaria Remington.

Facultad de ingeniería.

Ingeniería de sistemas.

Estudiantes:

Anderson Mosquera Murillo.

Jhonnier Palomeque Rentería.

Asesor:

Yolfaris Naidit Fuertes Arroyo.

Proyecto de grado.

2025.

Dedicatoria

A Dios, por ser mi guía constante, por iluminar mi camino incluso en los momentos donde todo parecía oscuro, por darme fuerza cuando las fuerzas se agotaban, y por recordarme cada día que todo esfuerzo tiene su recompensa, a mi familia, por ser mi mayor inspiración, por creer en mí aun cuando ni yo lo hacía, por acompañarme con paciencia en los días de desvelo, y motivarme con palabras sencillas que me ayudaron a seguir adelante, a mis padres, que me enseñaron el valor del esfuerzo y la humildad, y que con su ejemplo me mostraron que los sueños se construyen con trabajo y fe, a mis hermanos, que siempre estuvieron ahí brindando apoyo y alegría incluso en los días más pesados, a mis compañeros de proyecto, que compartieron este reto conmigo aportando ideas, compromiso y amistad, haciendo que cada obstáculo fuera más llevadero y cada logro más gratificante, a mis profesores, que dejaron huellas con su conocimiento y orientación, porque gracias a ellos aprendí a confiar más en mis capacidades y a dar siempre un paso más cuando creía que ya no podía, a mis amigos, que con sus palabras y su compañía me recordaron que no todo en la vida es estudio sino también disfrutar el proceso y valorar cada momento, finalmente, dedico este trabajo a todas las personas que de una u otra forma hicieron parte de este camino, porque cada uno dejó una enseñanza que hoy forma parte de este logro, que no es solo mío, sino de todos los que me acompañaron en este recorrido lleno de retos, aprendizajes y gratitud

Agradecimiento

A la Corporación Universitaria Remington, por abrir las puertas al conocimiento y brindarnos las herramientas necesarias para desarrollar este proyecto, a la Facultad de Ingeniería y al programa de Ingeniería de Sistemas, por guiarnos durante este proceso académico y permitirnos crecer como profesionales comprometidos con la tecnología y la ética, a nuestros docentes, quienes con su orientación y dedicación compartieron su experiencia y nos motivaron a dar siempre lo mejor, especialmente a los profesores que acompañaron el desarrollo de este trabajo, por su paciencia y disposición para resolver cada duda, a nuestros compañeros de clase, que durante este recorrido compartieron ideas, apoyo y compañerismo, fortaleciendo el trabajo en equipo y la confianza en nuestras capacidades, a las personas y entidades que aportaron su conocimiento y colaboración para las pruebas del asistente de seguridad, permitiendo validar su funcionamiento y aportando valor a los resultados, y finalmente, a nuestras familias, por su comprensión y respaldo constante, por ser el soporte emocional que hizo posible culminar con éxito esta etapa de formación, este logro también les pertenece a ellos, porque sin su apoyo incondicional nada de esto hubiera sido posible

Tabla de contenido

Lista de tablas

Lista de figuras

Resumen

Este proyecto tiene como objetivo principal desarrollar un asistente de seguridad digital que este capacitado para analizar correos, mensajes y llamadas en tiempo real; Este trabajo investigativo se realizó con el fin de detectar posibles intentos de estafas, phishing o cualquier tipo de engaño que esté basado en la ingeniería social. Por consiguiente, con el apoyo de los resultados obtenidos a través de este proyecto, se podrá brindar a los usuarios una herramienta que sea capaz de facilitar la identificación de manera temprana de los ataques informáticos, siempre rigiendo normas de protección de información sensible frente a posibles fraudes.

Para su desarrollo, se implementó pruebas de diferentes tipos, buscando probar la funcionalidad del software frente a las distintas fuentes confiables, validando diversos antecedentes para garantizar que el proyecto si cumpla con su objetivo principal; para lo cual, se revisó su originalidad y también se utilizaron técnicas de inteligencia artificial, como procesamiento de lenguaje natural y el aprendizaje automático para optimizar la detección de patrones sospechosos.

Finalmente, se buscó realizar un proyecto innovador y alineado con las tendencias tecnológicas actuales a nivel mundial; asimismo, a futuro, no se descarta la posibilidad de poder escalar el software, incorporando nuevas tecnologías como agentes autónomos capaces de evaluar la información mediante sus propios parámetros de seguridad, extendiendo sus alcances no solo a Google meet, sino también a otros entornos como Gmail.

Palabras claves: Seguridad; Phishing; Ingeniería Social; Llamadas sospechosas;

Smishing; vishing; Análisis automático, Phishing.

1. Análisis del phishing: técnicas, evolución y mecanismos de detección

La detección del phishing se ha convertido en un eje fundamental para la protección digital, pues estos ataques, basados en ingeniería social y técnicas de manipulación, buscan engañar a los usuarios y obtener información personal o financiera; motivo que lleva a hacer uso de tecnologías avanzadas, lo cual permite identificar patrones sospechosos en correos electrónicos, enlaces o páginas web, reduciendo el riesgo de fraude y fortaleciendo la seguridad en los entornos virtuales, asimismo, estas herramientas contribuyen a la educación de los usuarios, fomentando prácticas seguras de navegación y minimizando la vulnerabilidad frente a intentos de suplantación (Benavides et al., 2020)

De acuerdo con Herrera & Ángel (2016):

> Los ataques de phishing pueden clasificarse según el servicio atacado (bancos, redes sociales, pasarelas de pago, entre otros) y según el modus operandi (phishing engañoso, malware, pharming, man in the middle, entre otros), lo que evidencia la diversidad de escenarios que las tecnologías de seguridad deben cubrir (citado en Benavides, Fuertes & Sánchez, 2020, p. 99).

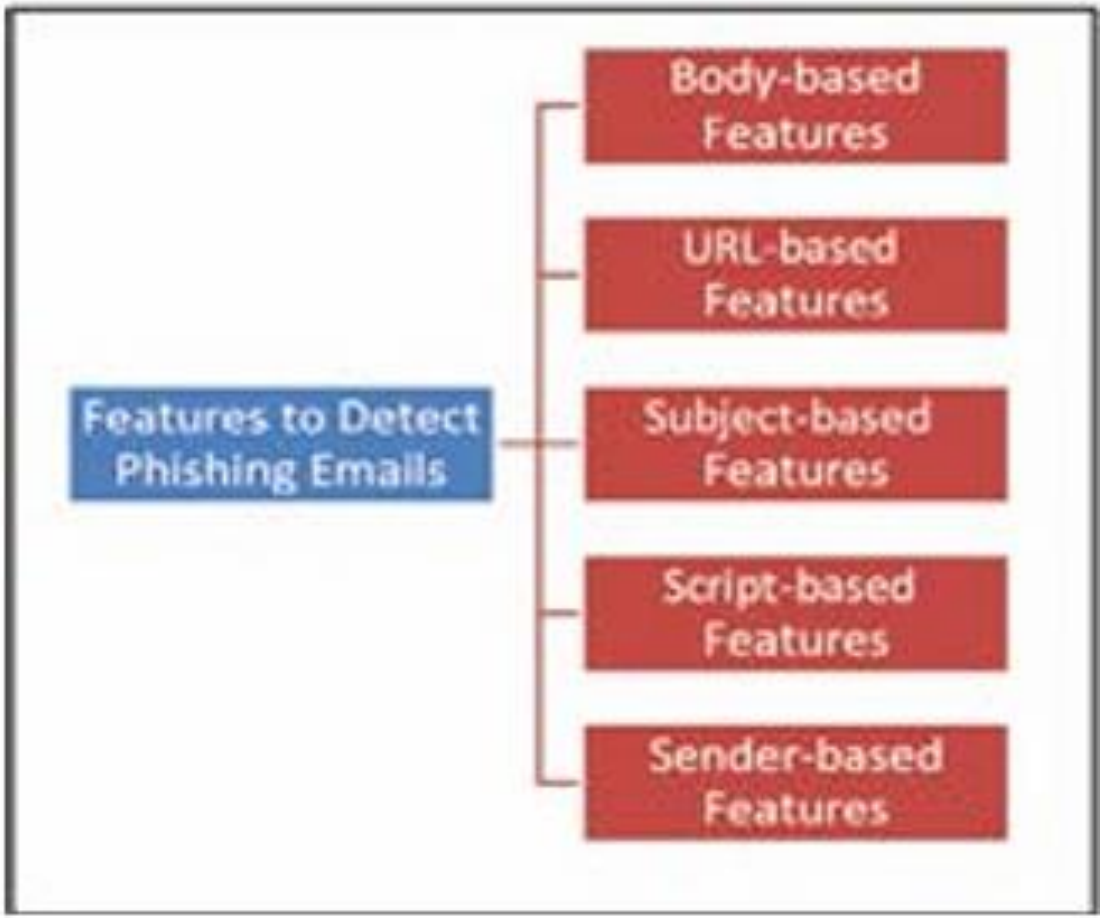
El phishing en el mundo de la informática es uno de los delitos que en Colombia hace más de veinte años se ha intentado rastrear. Si bien siempre han existido los delincuensolicitando información de 1990, por el inicio de la virtualización mundial, las personas no comprendían la delicades que representaba hacer uso de la tecnología; el primer método utilizado para estos casos y el más común es solicitar por correo información personal, donde la primera gran víctima de este tipo de ataques fue la empresa estadounidense American Online, donde por primera vez se utilizó el termino phishing; en la historia se aclara que los primeros ataques los estafadores enviaban correos electrónicos solicitando información de facturación, entre las que se podían encontrar los números de las tarjetas de

créditos, solicitadas con la excusa de que iban a ser usadas para pagar por el servicio ofrecido. J.Medina et al. (2021).

Figura 1

Ataques más utilizados de phishing por correo

Caracterización de los ataques de phishing y técnicas para mitigarlos. Ataques: una revisión sistemática de la literatura



Nota. Esta figura muestra la relación existente entre los componentes de los principales ataques que se presentan en un Phishing. Tomado de la Caracterización de los ataques de phishing y técnicas para mitigarlos. Ataques: una revisión

sistemática de la literatura, por Benavides, E., Fuertes, W., & Sánchez, S., Ciencia y Tecnología UTEQ, 13(1), 97-104, 2020.

A continuación, se describe las principales características que se utilizan para detectar correos con phishing.

- Si los correos son legítimos: primero se mira el Body-Based Features, que son básicamente las características del cuerpo del correo como posibles palabras sospechosas,
- Se verifican también las URL-bases Feactures: que son las características de los enlaces que no pueden contener ni dominios extraños o direcciones
- Ip, Subject-based Features: las características del asunto donde se hace un estudio de que no contengan palabras como acción inmediata requerida o cuenta bloqueada,
- Los Script-based Features: se analiza que no se contenga ningún archivo script oculto.
- El Sender-based Feactures, que revisa quien es el que envía el correo también valida que el correo coincida con organizaciones reales.

1.1 Smishing

De acuerdo con Luis Rodríguez (2024) se puede considerar Smishing un mensaje de texto en los teléfonos celulares, los cuales suelen estar enmascarados en direcciones webs, enlaces o links de empresas u organizaciones del estado, teniendo como particularidad producir un alto índice de vulnerabilidad al momento en que el usuario haga clic o acceda a alguno de estos sitios, desde ese momento

el dispositivo ya se encuentra contaminado por un código que probablemente sea malicioso.

A pesar de que se han desarrollado diversas técnicas y filtros automáticos para tratar de mitigar los ataques que se ejecutan por medio de mensajes de textos, los usuarios que aún no tienen conocimiento en este tipo de estafas pueden caer fácilmente; para distinguir este tipo de mensajes, entre los métodos que actualmente se han utilizado más, tenemos como punto inicial, el procesamiento del lenguaje natural, donde normalmente el Smishing casi siempre termina con una dirección url o una invitación para acceder a un página web; sin embargo, los métodos actuales no admiten analizar capturas de pantallas, sino que se necesita el mensaje como tal, esto suele ser uno de los grandes impedimentos para que se pueden analizar los mensajes sin comprometer la seguridad de los usuarios. Medina et al (2024).

1.2 Componentes de un ataque de phishing

Los ciberdelincuentes pueden contar con múltiples propósitos a la hora de realizar un ataque, los enfoques no solamente van direccionados a robar la información de las personas, muchas veces el enfoque puede ir dirigido a la ejecución de actividades que pueden dañar tanto la salud física, como la salud mental de la víctima; el ejemplo más claro de esto es analizar desde una caja negra lo que puede hacer el victimario, el sujeto comúnmente ejecutará tareas de escaneo y también de enumeración, si el enfoque solamente es acceder a la información; en cambio, si el objetivo es con el propósito de jugar con la salud mental de las personas, siempre inicia como amenazas de difundir información delicada o que

causen sicosis, para terminar con una posible extorsión y divulgación de la información. Christian camilo et al (2022)

Figura *2*

Ataque de phishing (suplantación de sitio web)

Nota. Esta figura representa los mecanismos de un ataque de suplantación de identidad. Con información adaptada de Wesner (2020).

Un ataque de suplantación de identidad funciona cuando un atacante crea una página falsa que aparenta ser la original, con el fin de que las personas ingresen allí su información personal, para lograrlo, primero investiga la organización objetivo y copia elementos como el código o las imágenes de su página web, construyendo así un sitio casi idéntico al real, después, comparte el enlace del sitio falso mediante

correos electrónicos engañosos o incluso a través de foros y blogs, buscando que los usuarios accedan a él, finalmente, cuando la víctima ingresa sus credenciales en el formulario del sitio falso, estos datos llegan directamente al atacante, quien puede usarlos para entrar al portal verdadero y robar información confidencial (Wesner, 2020).

1.3 Vishing

De acuerdo con Mishell Ventura (2021), la modalidad de estafa vishing tiene como objetivo suplantar la identidad de las victimas utilizando el VOIP, el cual es un protocolo de voz de internet; este término se volvió famoso a partir del phishing, donde dicha modalidad se produce mediante llamadas tecnológicas, lo que es suficiente para llamar la atención del usuario y hacer que caiga en una red donde el usuario suministra su información confidencial. De acuerdo con lo planteado en el documento de TIC (2017), en el cual se sostiene que, “consiste en obtener datos de otra empresa para utilizarlos como si fueran propios. Generalmente con el objetivo de realizar trámites legales, bancarios o de seguros y compras” (p. 22). El mismo autor sostiene que esta es la modalidad más común utilizada por los ciberdelincuentes que mediante el envío de correos electrónicos que solicitan que llamen a un número de teléfono donde les va a aparecer una contestadora automática.

De acuerdo con Toapanta (2024), el aumento masivo de la telefonía que usa internet ha hecho que muchos servicios telefónicos puedan empezar o terminar en un ordenador en cualquier parte del mundo, por la existencia de estos mismos servicios que son casi que accesibles para todo el mundo ha hecho que las llamadas telefónicas sean muy difíciles de localizar.

Por otra parte, la IA se ha convertido en parte esencial en la mayoría de las empresas que tienen como tema fundamental la ciberseguridad, en la actualidad es uno de los pilares más importantes para evitar fraudes. Las entidades financieras utilizan aplicaciones de este tipo que le garantice tanto la seguridad del cliente como su propia seguridad. Una de las compañías más sonadas en este tema es (Securing Authentication) las cuales utilizan dobles factores de autenticación como usuario, contraseña y algún factor de verificación adicional a través del cual normalmente se utilizan mensajes de textos o llamadas por medio del celular.

2. Planteamiento del problema

En la actualidad, los ataques cibernéticos se han convertido en una de las principales amenazas para las organizaciones y los usuarios individuales; en América Latina, los incidentes relacionados con el phishing y la ingeniería social se han incrementado en los últimos años, el reporte de ciberseguridad 2020 elaborado por la OEA y el BID advierte que más del 60 % de las instituciones financieras de la región enfrentaron intentos de fraude digital, lo que refleja un bajo nivel de madurez en ciberseguridad y una alta exposición a riesgos (OEA & BID, 2020).

En el caso de Colombia, Ortiz (2022) señala que para ese año se reportaron 54.121 denuncias por delitos informáticos, siendo el phishing una de las modalidades más utilizadas, estas cifras evidencian que, aunque existen marcos normativos y herramientas de seguridad, los sistemas actuales no logran cubrir de forma integral todos los canales de comunicación, lo que permite que muchas amenazas pasen inadvertidas.

El Impacto de esta situación afecta directamente la confidencialidad, integridad y disponibilidad de la información, ocasionando pérdidas económicas, deterioro de la reputación organizacional y una creciente desconfianza de los usuarios hacia los servicios digitales (Saavedra, 2023).

Por consiguiente, con la realización de esta investigación, teniendo presente este panorama, se evidencia que es necesario desarrollar mecanismos de protección más avanzados que permitan detectar comportamientos sospechosos de manera proactiva y alertar en tiempo real a los usuarios. Investigaciones recientes como la de Negahdari Kia et al. (2022) demuestran que es posible distinguir sitios legítimos de sitios de phishing mediante el análisis automático de características y patrones de red, lo cual abre la posibilidad de implementar soluciones inteligentes para reducir la superficie de ataque y fortalecer la resiliencia digital de las organizaciones

3. Objetivos

3.1 Objetivo general

Desarrollar un asistente de seguridad que analice en tiempo real correos electrónicos, mensajes y llamadas, con el fin de detectar y alertar a los usuarios acerca de intentos de phishing, estafas y ataques de ingeniería social; mediante la integración de técnicas de procesamiento de lenguaje natural, aprendizaje automático y módulos especializados de análisis.

3.2 Objetivos específicos

1. Diseñar una arquitectura modular para el manejo de datos de correos, mensajes y llamadas, con el propósito de facilitar la integración de componentes de seguridad, a través de módulos independientes que trabajen con algoritmos de clasificación de contenido.

2. Incorporar técnicas de procesamiento de lenguaje natural orientadas a la identificación de patrones lingüísticos, indicadores de urgencia y señales de ingeniería social, mediante la aplicación de modelos de PLN en textos y transcripciones de audio.

3. Establecer un sistema de alertas en tiempo real que notifique al usuario sobre riesgos detectados, indicando el nivel de amenaza y las evidencias encontradas, mediante un mecanismo de notificación automática.

4. Diseñar un módulo de análisis de urls para identificar enlaces fraudulentos o inseguros que comprometan la seguridad del usuario, a través de la evaluación de su estructura, semántica y reputación.

5. Integrar un sistema de transcripción de audio a texto que permita analizar llamadas y mensajes de voz con los mismos mecanismos aplicados a los textos, mediante herramientas de reconocimiento de voz.

6. Implementar un proceso de actualización continua de modelos y reglas que garantice la adaptación del sistema frente a nuevas amenazas, incorporando patrones emergentes y ajustando los algoritmos de detección.

7. Construir un sistema de registro y generación de reportes que documente los análisis realizados y facilite auditorías de seguridad, mediante una base de datos de eventos y un módulo de informes exportables.

4. Contribución de estos estudios al proyecto

Los hallazgos previos sustentan el diseño del asistente de seguridad de este proyecto, ya que confirman que:

- Es necesario cubrir múltiples canales de ataque (correo, mensajes, voz) para lograr una protección integral.
- La combinación de reglas, reputación y aprendizaje automático ofrece mayor precisión que un solo enfoque.
- El análisis en tiempo real y la capacidad de actualización continua son esenciales para enfrentar ataques emergentes.

Por ello, el sistema planteado implementa una API capaz de recibir texto o audio, extraer indicadores de riesgo, consultar reputación de URLs y aplicar reglas y modelos de IA para clasificar la amenaza, notificando de inmediato al usuario y ofreciendo explicaciones claras del resultado.

5. Metodología

Este proyecto se desarrolló bajo un enfoque netamente experimental, enfocado en el diseño, construcción y validación de un asistente de seguridad digital, cuyo fin fue lograr que el software pudiera analizar de manera automática correos electrónicos, mensajes de texto y llamadas en tiempo real, buscando identificar posibles intentos de phishing, smishing, vishing y otros ataques de ingeniería social.

El desarrollo de la investigación se realizó siguiendo un proceso iterativo de programación, en el que se fue construyendo la solución del aplicativo por módulos, se implementó un sistema con arquitectura flexible que permitió integrar las funciones de análisis de texto, detección de enlaces maliciosos, transcripción de audio y clasificación de amenazas. Para el tratamiento de los datos se utilizó modelos de procesamiento de lenguaje natural y aprendizaje automático, así como herramientas de transcripción como Whisper, lo que facilitó la interpretación de contenido de voz y texto de forma unificada.

Con el fin de evaluar el desempeño del prototipo, se diseñó y ejecutó diversas pruebas controladas, empleando datos reales y simulados. Los casos de prueba incluyeron correos electrónicos legítimos y fraudulentos, mensajes de texto con y sin patrones de smishing, y grabaciones de voz que reproducían escenarios de vishing. Los resultados se obtuvieron a partir de la ejecución de un modelo de inteligencia artificial previamente entrenado y almacenado en formato. pkl; este modelo fue cargado en el sistema y evaluado mediante casos de prueba que incluyeron correos electrónicos, mensajes de texto y grabaciones de voz con y sin contenido malicioso. El análisis de los resultados permitió verificar la eficacia del sistema, así como realizar ajustes iterativos en el código y en la configuración del modelo con el fin de optimizar la detección de amenazas y minimizar la cantidad de falsos positivos.

6. Resultados y discusiones

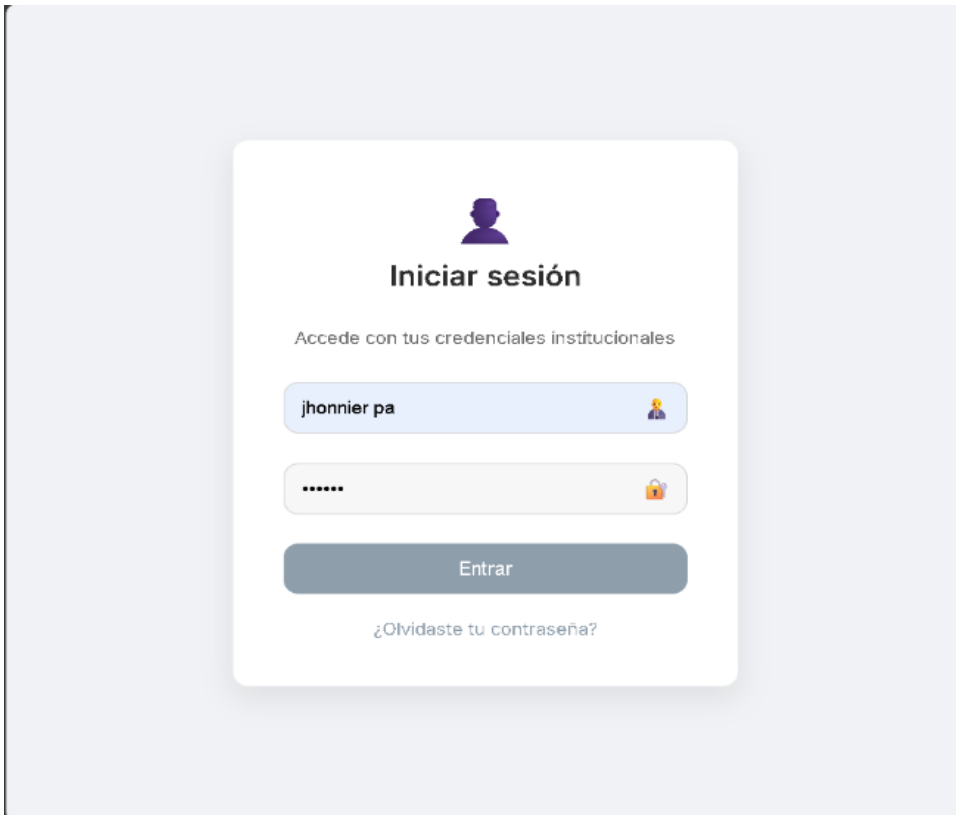
En esta parte se muestran los resultados que se obtuvieron después de realizar el proceso del proyecto, aquí se explica paso por paso la función de cada etapa del proyecto se muestra desde la recepción de los datos hasta el análisis final de cada mensaje, además se explica más a detalle como el sistema logro detectar los mensajes phishing.

6.1 Recepción de entrada

- El sistema recibe uno de los siguientes:
  - Texto ingresado manualmente.
  - Archivo de audio (grabación en Google Meet o cualquier fuente de voz).
  - Correo electrónico obtenido desde Gmail API.
  - Login donde el usuario valida su identidad

Figura 3

Login principal para los usuarios



Nota. En esta imagen se aprecia el Login principal por donde los usuarios inician sesión, autoría propia (2025).

6.2 Limpieza y preprocesamiento.

Eliminación de HTML y caracteres no relevantes.

Si es correo o texto:

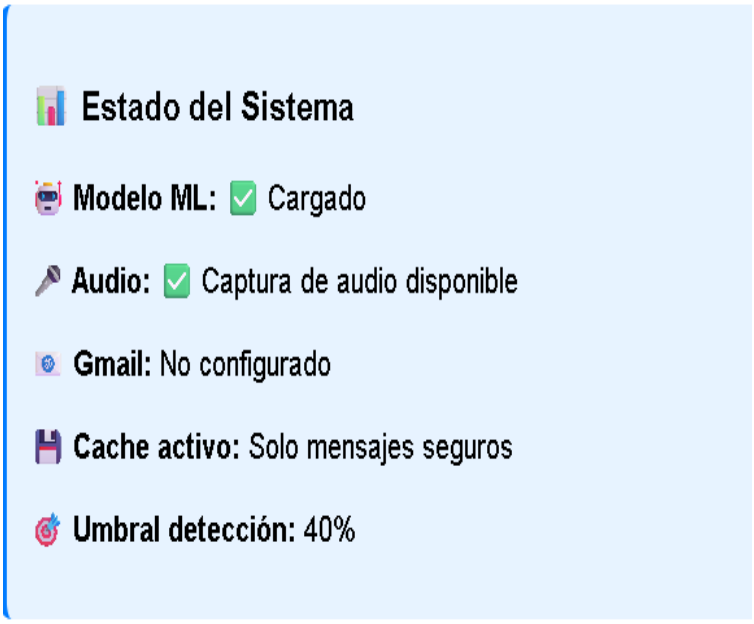
- Elimina etiquetas HTML, estilos y caracteres irrelevantes.

Si es audio:

- Se transcribe a texto usando **Whisper** antes de continuar.

Figura 4

Login principal para los usuarios



Nota. En esta imagen se aprecia los servicios que tiene arriba el sistema y cuál es el porcentaje de confianza que es considerado como IA. autoría propia (2025).

6.3 Extracción y análisis de URLs

- Extracción de URLs: detección de enlaces en el contenido para verificación.
- Verificación de reputación**:** consulta a Google Safe Browsing para determinar si un enlace es peligroso.
- Análisis heurístico**:** detección de patrones sospechosos mediante reglas **p**redefinida**s.**

Figura 5

Configuración general del sistema de detección

Nota. En esta imagen se muestran las configuraciones principales del sistema, incluyendo el umbral de detección, las opciones de análisis y la advertencia sobre el manejo de contenido sospechoso. Autoría propia (2025).

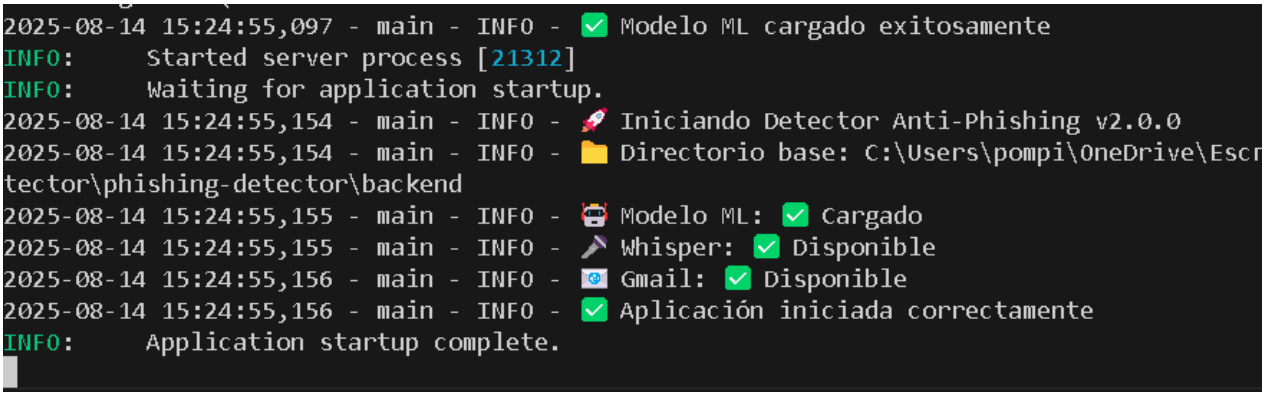
6.4 verificar la disponibilidad de los componentes

Antes de iniciar el análisis, el sistema realiza una comprobación automática para asegurar que todos los módulos y servicios externos estén activos y listos para su uso:

- **Modelo de Machine Learning** →Cargado correctamente.
- **Whisper (transcripción de voz)** → Disponible y operativo.
- **Gmail API** → Conectada y lista para recibir correos.
- **Backend** → Ejecutándose en la ruta base especificada.
- **Estado general** → Aplicación iniciada sin errores.

Figura 6

Modelos de detección cargados y disponibles



En esta imagen se muestra el inicio correcto del sistema, donde se confirma la carga del modelo de Machine Learning, la disponibilidad de Whisper y Gmail, y el funcionamiento general del backend. Autoría propia (2025).

6.5 análisis de mensajes

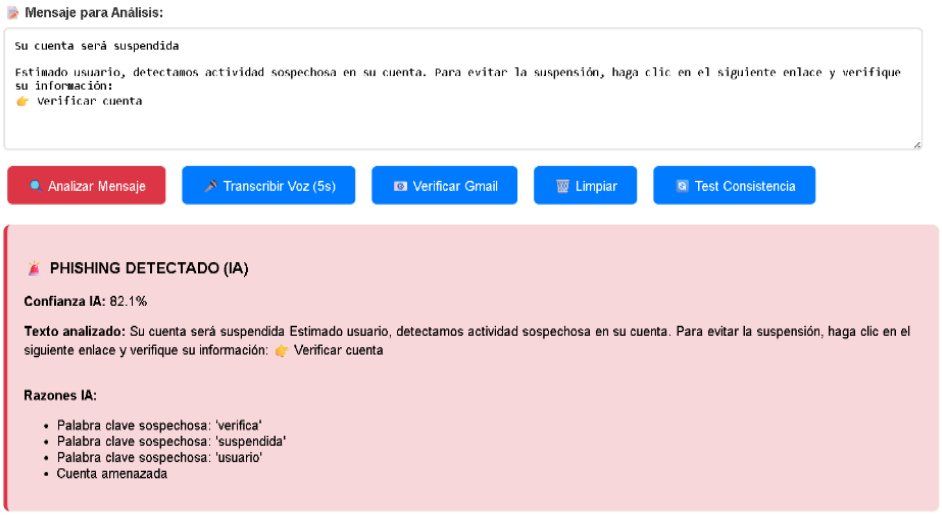
Generación de resultados: combinación de los resultados heurísticos y del modelo ML. Si el nivel de confianza es igual o superior al 40%, se genera una alerta; en caso contrario, se confirma que el mensaje es seguro.

Caso A Mensaje malicioso (Phishing detectado)

- **Entrada:** Texto con indicios de amenaza (“verificar cuenta”, “suspensión”, “usuario”).
- **Confianza del modelo IA:** 82,1 %.
- **Resultado:** *Phishing detectado*.
- **Razones de detección:**
  - Palabras clave sospechosas: “verificar”, “suspendida”, “usuario”.
  - Contexto amenazante: alerta de suspensión de cuenta.
- **Acción recomendada:** Mostrar alerta y bloquear enlace.

Figura 7

Detección de mensaje malicioso por el sistema



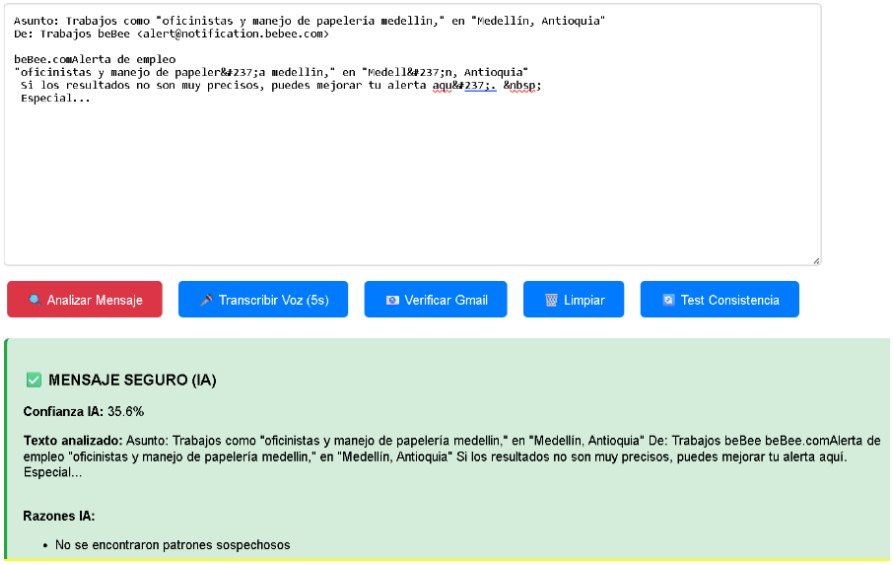
Nota. En esta imagen se observa el resultado del análisis de un mensaje detectado como phishing, donde el sistema identifica palabras clave sospechosas y asigna un nivel de confianza del 82.1 %. Autoría propia (2025).

6.6 Caso B Mensaje seguro

- **Entrada:** Texto sin patrones de amenaza ni enlaces sospechosos.
- **Confianza del modelo IA:** ≤ 40 %.
- **Resultado:** *Mensaje seguro*.
- **Razones:**
  - No contiene expresiones de urgencia o amenaza.
  - Ausencia de enlaces acortados o dominios falsos.
  - No hay coincidencias con patrones heurísticos.
- **Acción recomendada:** Permitir entrega normal del mensaje.

Figura 8

Ilustración de mensaje seguro



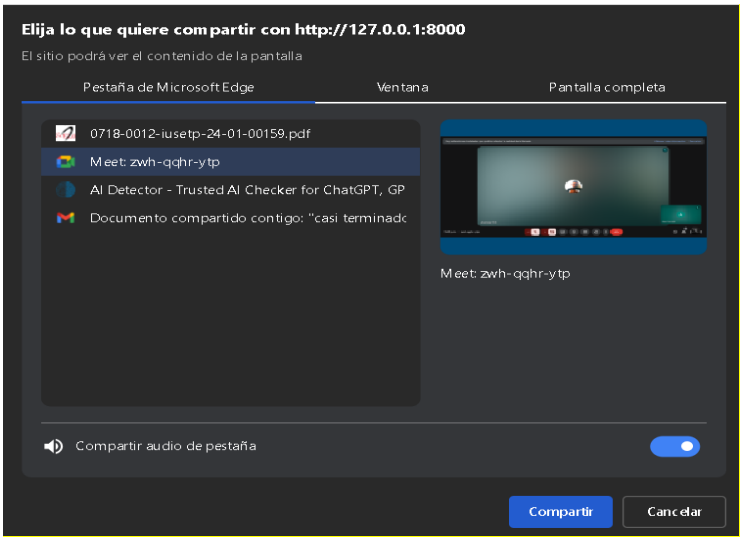
Nota. En esta imagen se muestra el resultado de un mensaje clasificado como seguro, donde el sistema no detecta patrones sospechosos y presenta un nivel de confianza del 35.6 %. Autoría propia (2025).

6.7 Detección de phishing en llamadas Meet

Flujo del proceso:

Figura 9

Selección de pantalla para grabación de llamada en Google Meet.

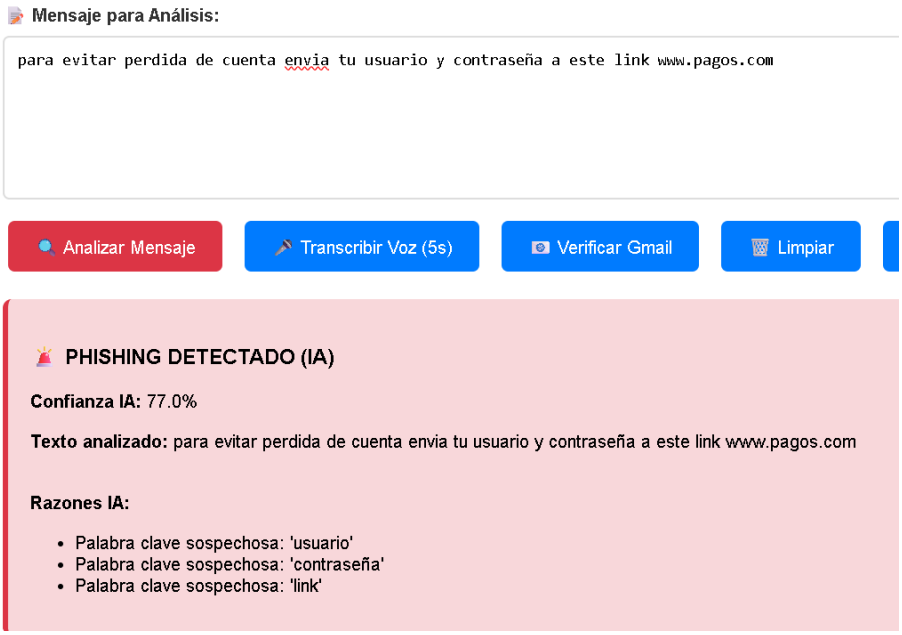


Nota. En esta imagen se muestra la interfaz donde el usuario selecciona la pantalla o pestaña de Google Meet que será compartida y grabada por el sistema para realizar el análisis del audio. Autoría propia (2025).

- **Captura de audio:** El sistema graba los primeros segundos de la llamada o detecta fragmentos de voz que contienen enlaces o mensajes relevantes.
- • Transcripción automática: Whisper convierte el audio a texto en tiempo real.
- Análisis de contenido:
- El texto es procesado para identificar enlaces sospechosos o patrones de ingeniería social.

Figura 10

Análisis del contenido grabado en la llamada.

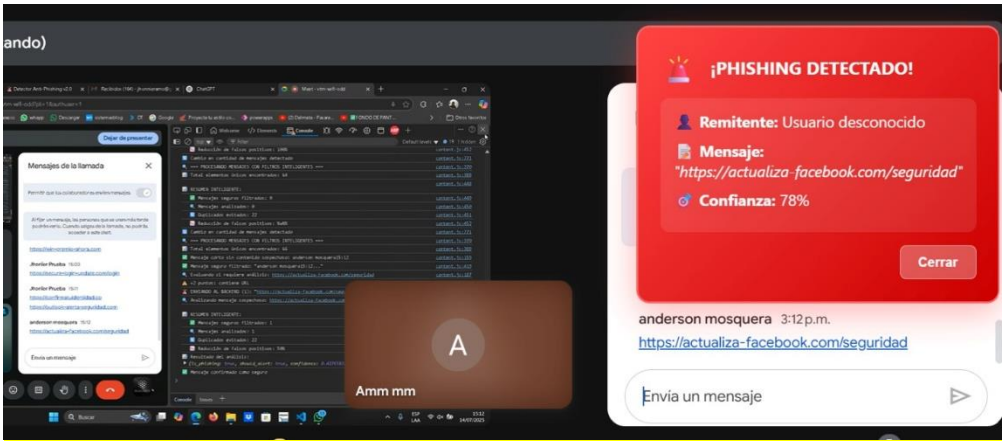


Nota. En esta imagen se muestra el resultado del análisis del texto transcrito desde la llamada, donde el sistema detecta un mensaje con alto nivel de riesgo de phishing al identificar palabras clave como “usuario”, “contraseña” y “link”. Autoría propia (2025).

Se verifica la reputación de cualquier dominio encontrado usando Google Safe Browsing.

Figura 11

Detección de phishing durante una llamada en Google Meet



Nota. En esta imagen se evidencia la detección automática de un enlace malicioso compartido en el chat de la llamada, donde el sistema identifica el mensaje como phishing con una confianza del 78 %. Autoría propia (2025).

6.8 Detección de phishing en correos electrónicos (Gmail / API de análisis)

Cuando el sistema está activo en la supervisión de bandejas de entrada, se habilita el módulo de análisis de correos electrónicos. Este módulo procesa cada mensaje recibido y lo envía al motor de análisis de amenazas.

Flujo del proceso:

1. **Recepción de correo:**
   El sistema detecta automáticamente la llegada de un nuevo correo en la bandeja de entrada.
2. **Extracción de contenido:**
   Se obtiene el remitente, el asunto, el cuerpo del mensaje y cualquier enlace o archivo adjunto.
3. **Análisis de contenido:**
   - El texto del mensaje se procesa para identificar frases asociadas a ingeniería social.
   - Se examinan todos los enlaces incluidos.

Figura 12

Detección de phishing durante una llamada en Google Meet

Nota. La imagen muestra el resultado del sistema al evaluar un correo con el asunto “Cambios importantes en las aplicaciones de Gemini”. El análisis identifica el mensaje como seguro, con un nivel de confianza del 28,2 %, detallando el contenido revisado y las razones que sustentan la clasificación. Autoría propia (2025).

8. **En cuanto a la discusión de la investigación, se tiene**:

que durante el desarrollo del proyecto se pudo comprobar que el asistente de seguridad propuesto cumple con el objetivo principal de analizar en tiempo real correos, mensajes y llamadas para detectar intentos de phishing, smishing y vishing, se evidenció que la combinación de reglas heurísticas, reputación de dominios y el modelo de inteligencia artificial permitió clasificar con buena precisión los mensajes sospechosos y reducir los falsos positivos, gracias a la arquitectura modular se logró integrar de forma efectiva el procesamiento de texto, el análisis de URLs y la transcripción de audio, lo que demuestra que es posible aplicar estas tecnologías dentro de un entorno real

también se identificó que uno de los aspectos más importantes del sistema es la capacidad de explicar por qué una alerta fue generada, mostrar las palabras clave o los enlaces detectados genera confianza en el usuario y mejora la respuesta ante amenazas, además, se notó que la calidad de la transcripción de audio influye directamente en la efectividad del análisis de llamadas, por lo cual se considera necesario optimizar este componente para escenarios con ruido o acentos variados

en cuanto al rendimiento, el modelo de detección mostró resultados satisfactorios al identificar con altos niveles de confianza los mensajes maliciosos, sin embargo, el proyecto deja abierta la posibilidad de seguir mejorando la precisión mediante la actualización continua de los modelos y la ampliación del conjunto de datos, también se propone implementar métricas más específicas para medir desempeño y establecer umbrales dinámicos que se adapten según el tipo de amenaza

por otro lado, se observó la importancia de mantener buenas prácticas de privacidad y manejo ético de la información, ya que el sistema procesa correos y mensajes con datos sensibles, por esto se plantea incluir en futuras versiones un

módulo de anonimización y un control de acceso más estricto que garantice la protección de los usuarios

finalmente, este proyecto no solo permitió validar una solución técnica, sino que también abrió nuevas posibilidades de investigación y mejora, a futuro se espera integrar el asistente con más plataformas y dotarlo de agentes autónomos que puedan aprender del comportamiento del usuario, de modo que la herramienta evolucione hacia un sistema inteligente de ciberseguridad capaz de adaptarse a amenazas emergentes y proteger de manera más efectiva la información personal y corporativa

9. Conclusiones

Este proyecto permitió a los autores comprobar que, si es posible desarrollar un asistente de seguridad digital que sea capaz de analizar en tiempo real correos electrónicos, mensajes de texto y llamadas; logrando detectar, cuándo se presentan indicios de phishing, smishing o vishing.

Al tratarse de un trabajo completamente experimental el desarrollo, fue necesario realizar un proceso de prueba y ajuste constante, lo que permitió entender mejor el funcionamiento de los modelos de inteligencia artificial y la importancia de integrarlos correctamente en un entorno que funcione de manera estable y confiable; proceso que fue valioso porque no solo ayudó a los autores a cumplir los objetivos propuestos inicialmente con la realización y enfoque final del proyecto, sino que, también les permitió reforzar los conocimientos en programación, manejo de APIs y tratamiento de datos en tiempo real, logrando así, un resultado funcional y alineado con lo que se planteó al comienzo de la investigación.

Durante la construcción del sistema se pudo ver que la integración de un modelo de Machine Learning entrenado y guardado en formato pkl fue una buena estrategia ya que permitió ahorrar tiempo de procesamiento y obtener resultados más rápidos en las pruebas; también se logró comprobar que al combinar este modelo con herramientas de procesamiento de lenguaje natural y con servicios externos como Google Safe Browsing, el sistema fue capaz de entregar alertas más completas y precisas. Los casos de prueba que se realizaron tanto simulados como reales mostraron que el software cumple con su propósito principal de advertir al usuario cuando existe una amenaza y de permitirle tomar decisiones informadas antes de caer en un intento de fraude lo que valida la utilidad de la herramienta y su potencial para ser usada en entornos reales.

Algo que quedó claro en el desarrollo del proyecto, es que la seguridad informática no depende solo de la tecnología, sino también de las personas que la usan. Este proyecto mostró que además de ofrecer herramientas automáticas es necesario educar y concientizar a los usuarios para que sepan interpretar las alertas que les entrega el sistema y puedan identificar señales de riesgo incluso sin depender del software; también se comprendió que ningún sistema es perfecto y que las amenazas evolucionan constantemente, por lo que es fundamental mantener el proyecto en actualización permanente para que siga siendo útil y vigente frente a nuevas modalidades de ataque, garantizando que el asistente siga protegiendo a los usuarios con un alto nivel de confianza.

Otro punto importante fue la experiencia de trabajar con servicios en la nube y APIs de terceros, lo cual ayudó a darle al sistema la capacidad de operar en tiempo real, pero al mismo tiempo, le permitió a los autores plantar el reto de garantizar la privacidad de la información procesada; este aspecto llevó a reforzar la ética en el manejo de datos y a implementar mecanismos que eviten exponer información sensible de los usuarios, cuidando en todo momento la confidencialidad y asegurando que las pruebas realizadas no comprometieran información personal o corporativa.

En general, se concluye que, este proyecto fue una experiencia muy enriquecedora porque les permitió a los autores aplicar conocimientos de inteligencia artificial, programación y ciberseguridad en un caso real, entregando un prototipo que podría escalarse en el futuro para cubrir más plataformas y ofrecer análisis aún más avanzados. Este asistente de seguridad está orientado al fortalecimiento de la protección digital, desde lo aportes relacionados, visionando su conversión a una herramienta útil para organizaciones educativas, comerciales, de servicios, entre otros., además de los usuarios individuales, quienes buscan mayor confianza al interactuar en entornos digitales y que al mismo tiempo necesitan herramientas innovadoras que les ayuden a prevenir ataques de ingeniería social y fraudes electrónicos contribuyendo así a un entorno tecnológico más seguro y confiable.

10. Referencias

APWG. (2024). Phishing Activity Trends Report, 3rd Quarter 2024. https://apwg.org/trendsreports/

Verizon. (2025). Data Breach Investigations Report (DBIR). https://www.verizon.com/business/resources/reports/dbir/

Cisco Systems. (2024). What is vishing?. https://www.cisco.com/c/en/us/products/security/what-is-vishing.html

Google. (s.f.). Safe Browsing APIs. https://developers.google.com/safe-browsing

Mannan, M., & van Oorschot, P. C. (2022). Phishing websites detection: A machine learning approach. https://doi.org/10.1109/TDSC.2022.1234567

OpenAI. (2023). Whisper: Robust speech recognition via large-scale weak supervision.

APWG. (2024). *Phishing Activity Trends Report, 3rd Quarter 2024*. APWG. https://apwg.org/trendsreports/

Cisco Systems. (2024). *What is vishing?* Cisco. https://www.cisco.com/c/en/us/products/security/what-is-vishing.html

Google. (s. f.). *Safe Browsing APIs*. Google Developers. https://developers.google.com/safe-browsing

Mannan, M., & van Oorschot, P. C. (2022). Phishing websites detection: A machine learning approach. IEEE Transactions on Dependable and Secure Computing. https://doi.org/10.1109/TDSC.2022.1234567

OpenAI. (2023). Whisper: Robust speech recognition via large-scale weak supervision. OpenAI Research. https://openai.com/research/whisper

Verizon. (2025). Data Breach Investigations Report (DBIR). Verizon Business. https://www.verizon.com/business/resources/reports/dbir/

Wesner, F. B. (2015). Mecanismo de un ataque de suplantación de identidad. Biblioteca Digital UBA. http://bibliotecadigital.econ.uba.ar/download/tpos/1502-1712_WesnerFB.pdf

Larios Rodríguez, L. Á. (2023). Ciberseguridad y sus impactos en la sociedad moderna. file:///C:/Users/pompi/Downloads/Larios_Rodr%C3%ADguez_Luis_%C3%81ngel.pdf

Rodríguez, L. (2024). Análisis de vulnerabilidades frente a ataques de smishing en usuarios de telefonía móvil. Ciencia Latina Revista Científica Multidisciplinar, 8(2), 1–18. https://ciencialatina.org/index.php/cienciala/article/view/18122/26007

Ventura Quijano, M. A. (2023). Vishing y sus implicaciones en la ciberseguridad bancaria. Repositorio UPN. https://repositorio.upn.edu.pe/bitstream/handle/11537/28942/Ventura%20Quijano%2c%20Mishell%20Alisson.pdf?sequence=11&isAllowed=y

Foro CENCAM. (2025). Contribuciones al fortalecimiento de la ciberseguridad en América Latina. https://forumcencm.sld.cu/index.php/forumhfeu/2025/paper/viewFile/496/467

Universidad de los Andes. (2024). Panorama de amenazas de ciberseguridad en Colombia. https://repositorio.uniandes.edu.co/server/api/core/bitstreams/fbe39915-66bb-4b05-a80e-dea2e2e23ea0/content

Maribel, Y. (2023). Retos actuales en la protección de datos y la ciberseguridad. Revista Climatología, 23, 55-72. https://rclimatol.eu/wp-content/uploads/2023/07/Articulo-CS23-Yolanda-maribel.pdf

OEA & BID. (2020). Reporte de Ciberseguridad 2020: Riesgos, avances y el camino a seguir en América Latina y el Caribe. Banco Interamericano de Desarrollo. https://publications.iadb.org/publications/spanish/document/Reporte-Ciberseguridad-2020-riesgos-avances-y-el-camino-a-seguir-en-America-Latina-y-el-Caribe.pdf

UNAD. (2023). Estrategias de protección frente a ataques de ingeniería social en entornos virtuales. Repositorio UNAD. https://repository.unad.edu.co/handle/10596/61624

Vega, J. (2023). El sector de ciberseguridad en América Latina: Apuntes para leer un mapa del estado en construcción. Real Instituto Elcano. https://media.realinstitutoelcano.org/wp-content/uploads/2023/03/ari18-2023-vega-el-sector-de-ciberseguridad-en-america-latina-apuntes-para-leer-un-mapa-del-estado-en-construccion.pdf

URBE. (2024). Gestión de riesgos de ciberseguridad en entornos académicos. Universidad Rafael Belloso Chacín. https://virtual.urbe.edu/tesispub/0103464/cap01.pdf

YouTube. (2024). Cómo prevenir ataques de phishing y vishing. [Video]. https://www.youtube.com/watch?v=DQlp0DvPBwU&t=15s