Inteligencia Artificial para la detección de phishing y prevención del robo de identidad en adultos mayores en Colombia

Kevin Julian Neisa González - 20222020224
Facultad de Ingeniería, Universidad Distrital
Email: kjneisag@udistrital.edu.co

Resumen

El phishing se ha consolidado como una de las principales amenazas de ciberseguridad en Colombia, afectando de manera significativa a los adultos mayores debido a su limitada experiencia en el uso de tecnologías digitales. Esta vulnerabilidad ha provocado un aumento de casos de robo de identidad, con consecuencias que incluyen pérdidas económicas, suplantaciones legales y desconfianza hacia los medios digitales. Frente a esta problemática, se propone una investigación orientada al diseño de una solución preventiva basada en inteligencia artificial (IA), enfocada en la protección de adultos mayores en el entorno del hogar. La metodología contempla la identificación de los patrones más frecuentes en ataques de phishing, el diseño y entrenamiento de un modelo de IA capaz de detectar mensajes fraudulentos en correos electrónicos, mensajes de texto y enlaces web, y el desarrollo de un prototipo que genere alertas en tiempo real. Asimismo, se plantea el diseño de una interfaz accesible y amigable, adaptada a las necesidades cognitivas de los adultos mayores, que facilite la interacción con la herramienta. Se espera que esta propuesta contribuya a reducir los riesgos de robo de identidad, fortalecer la confianza digital de los adultos mayores y aportar un enfoque innovador a la ciberseguridad en el ámbito doméstico.

Palabras Clave - Phishing, robo de identidad, adultos mayores, ciberseguridad, inteligencia artificial, hogares digitales.

Abstract

Phishing has established itself as one of the main cybersecurity threats in Colombia, significantly affecting older adults due to their limited experience using digital technologies. This vulnerability has led to an increase in identity theft cases, with consequences that include financial losses, impersonation, and distrust of digital media. In response to this problem, we propose a research project aimed at designing a preventive solution based on artificial intelligence (AI), focused on protecting older adults in the home environment. The methodology includes identifying the most common patterns in phishing attacks, designing and training an AI model capable of detecting fraudulent messages in emails, text messages, and web links, and developing a prototype that generates real-time alerts. We also propose designing an accessible and user-friendly interface, adapted to the cognitive needs of older adults, to facilitate interaction with the tool. This proposal is expected to contribute to reducing the risks of identity theft, strengthen older adults' digital trust, and provide an innovative approach to cybersecurity in the home environment.

Keywords - Phishing, identity theft, older adults, cybersecurity, artificial intelligence, digital homes.

Introducción

En Colombia, el phishing se ha consolidado como una de las amenazas más relevantes en ciberseguridad, porque los atacantes usan correos, SMS, llamadas y enlaces fraudulentos para suplantar identidades y obtener datos sensibles (contraseñas, números de documento, datos bancarios). Los estudios sobre fraude muestran que este tipo de estafa ha crecido y diversificado sus vectores en los últimos años, afectando con especial gravedad a poblaciones con menor alfabetización digital. [3],[1]

En el ámbito doméstico, los hogares se convierten en el escenario principal donde se produce el robo de identidad por phishing: los atacantes apuntan a cuentas de correo, servicios bancarios en línea y llamadas de "soporte técnico" que inducen a transferencias o a revelar información personal. Estas tácticas explotan la urgencia y la confianza del usuario, y se manifiestan por canales múltiples (email, SMS, robocalls, redes sociales), lo cual evidencia la necesidad de soluciones que operen en tiempo real y de forma integrada. [3],[1]

Los adultos mayores constituyen un grupo especialmente vulnerable en los hogares colombianos por varias razones: brechas en habilidades digitales, impactos del ageismo en el diseño de tecnologías, dependencia de "expertos cálidos" (familiares) para resolver problemas técnicos, y dificultades para interpretar señales lingüísticas o formales de un mensaje fraudulento. Estudios sobre envejecimiento y tecnología muestran la heterogeneidad de este grupo y cómo la falta de interfaces inclusivas y de apoyo accesible incrementa su exposición al fraude. Además, investigaciones centradas en la autenticidad del contenido han identificado indicadores lingüísticos (gramática, sintaxis, tono) que los atacantes descuidan y que, si se enseñan o se automatizan, pueden ayudar a detectar phishing dirigido a personas mayores. [2],[5]

Las consecuencias en los hogares son múltiples: pérdidas económicas directas, suplantación de identidad en trámites familiares, impacto emocional y pérdida de confianza en servicios digitales. Aunque existen medidas tradicionales de higiene digital (contraseñas fuertes, MFA, antivirus), el dinamismo del phishing exige soluciones adaptativas. En este contexto, la Inteligencia Artificial (IA) presenta una alternativa prometedora: modelos de procesamiento de lenguaje natural y clasificación que identifiquen patrones de phishing en mensajes y enlaces, y que, al mismo tiempo, requieran un diseño inclusivo para no dejar fuera a poblaciones vulnerables. Sin embargo, el desarrollo de sistemas basados en IA debe atender riesgos de sesgo en los datos y a la necesidad de interfaces accesibles para adultos mayores. [1],[4],[5]

Por todo lo anterior, surge la necesidad de investigar y diseñar herramientas basadas en IA que estén orientadas al entorno doméstico y pensadas específicamente para adultos mayores en Colombia, con el objetivo de prevenir el robo de identidad por phishing y mejorar la confianza de este grupo en el uso de servicios digitales desde sus hogares. Esta investigación debe integrar hallazgos sobre diseño inclusivo, rasgos lingüísticos de mensajes fraudulentos y buenas prácticas de implementación de IA para poblaciones vulnerables. [1],[2],[3],[4],[5]

Metodología

La presente investigación adopta el modelo **CRISP–DM (Cross-Industry Standard Process for Data Mining)** para estructurar el proceso de análisis y diseño de una solución basada en inteligencia artificial orientada a la detección de phishing y la prevención del robo de identidad en adultos mayores en Colombia. Este enfoque permite organizar el estudio en seis fases integradas, garantizando reproducibilidad, claridad y rigor metodológico.

1. Comprensión del Negocio

El objetivo principal es analizar cómo la inteligencia artificial puede emplearse para identificar ataques de phishing y proteger a adultos mayores en entornos domésticos en Colombia. Se parte del reconocimiento de que esta población presenta mayor vulnerabilidad debido a brechas en alfabetización digital, dificultades cognitivas y exposición a múltiples vectores de ataque (correos, SMS, llamadas fraudulentas, enlaces web).

En esta fase se definió la pregunta de investigación: **"¿Cómo se aplican técnicas de inteligencia artificial para identificar ataques de phishing y proteger a los adultos mayores en el entorno del hogar?"**

También se formularon metas específicas:

- Identificar los patrones lingüísticos, estructurales y de comportamiento presentes en ataques de phishing.
- Evaluar técnicas de IA aplicadas en la detección de phishing revisadas en literatura reciente.
- Proponer lineamientos para un prototipo accesible, inclusivo y adaptado a adultos mayores.

2. Comprensión de los Datos

Se recopiló información secundaria proveniente de **cinco artículos recientes** sobre ciberfraude, IA para detección de phishing, plataformas inteligentes y diseño inclusivo para adultos mayores [6],[7],[1],[4],[3].

En esta fase se identificaron:

- **Tipos de datos utilizados por los estudios:**
  - Datasets públicos de phishing (correos, SMS, URLs).
  - Reportes institucionales y estadísticas de fraude.
  - Datos textuales con anotaciones para entrenamiento de NLP.
  - Interacciones de usuarios en plataformas educativas o asistivas.
- **Características relevantes:**
  - Indicadores lingüísticos (errores, tono, urgencia).
  - Indicadores técnicos (dominios falsos, acortadores, encabezados).
  - Señales de ingeniería social dirigidas a personas mayores.

Se realizó una revisión cualitativa inicial que permitió comprender cómo se estructuran los datasets y los enfoques más comunes en modelos de detección de phishing.

3. Preparación de los Datos

A partir del análisis de los artículos, se sintetizaron los procesos de preparación utilizados en las investigaciones revisadas:

- Limpieza de texto y normalización lingüística.
- Tokenización, stemming o lematización.
- Identificación de features relevantes: palabras clave, patrones de urgencia, análisis de encabezados, reputación de URLs.
- Construcción de matrices de características (TF-IDF, embeddings, n-grams).
- División de los datasets en entrenamiento, validación y prueba.

Aunque este estudio no construye un dataset propio, esta fase sistematiza las prácticas observadas para fundamentar las recomendaciones del prototipo a diseñar.

4. Modelado

En esta fase se analizan las técnicas de IA propuestas en los estudios comparados, considerando su aplicabilidad al contexto colombiano:

- **Machine Learning tradicional:** SVM, Random Forest, Naive Bayes para clasificación de correos y URLs [6], [4].
- **Deep Learning:** redes neuronales, LSTM, transformers para detección en tiempo real [7].
- **Procesamiento de Lenguaje Natural (NLP):** análisis de tono, sintaxis, embeddings semánticos para identificar señales de phishing dirigidas a adultos mayores [1], [4].
- **Agentes inteligentes y chatbots:** modelos orientados a asistencia, educación y alerta en tiempo real [6].

Se evaluaron también las métricas reportadas (accuracy, recall, F1-score), comparando desempeño y limitaciones según el tipo de dato y modelo.

5. Evaluación

La evaluación se desarrolló comparando los resultados, métricas y contribuciones de cada estudio con relación a la detección de phishing:

- Se priorizaron modelos con **alto recall**, dado que la población objetivo no debe exponerse a falsos negativos (mensajes peligrosos sin detectar).
- Se analizaron las limitaciones detectadas en los estudios:
    - Sesgo en datasets no representativos de poblaciones latinoamericanas.
    - Exceso de complejidad en herramientas no accesibles para adultos mayores.
    - Falta de integración entre notificación y explicación clara del riesgo.

Esta evaluación permitió generar criterios para una solución inclusiva y adaptada al contexto colombiano.

6. Despliegue

Aunque esta investigación no implementa un sistema completo, sí propone lineamientos para un prototipo de despliegue orientado al hogar:

- Utilizar modelos livianos o de inferencia rápida capaces de ejecutarse en dispositivos domésticos.
- Integrar una interfaz accesible con elementos como tipografías grandes, lenguaje claro y alertas visuales/sonoras.
- Considerar herramientas de despliegue mencionadas en los artículos:
    - **LLaMA u otros modelos locales**,
    - **Docker u Ollama** para contenedores seguros,
    - API o agente inteligente para análisis en tiempo real.
- Incorporar un módulo educativo que explique por qué un mensaje fue detectado como sospechoso, reforzando la alfabetización digital.

El resultado final de esta fase es una propuesta metodológica replicable que combina buenas prácticas de IA, diseño inclusivo y ciberseguridad centrada en adultos mayores.

Diseño, Desarrollo e Implementación

Esta sección describe el proceso técnico mediante el cual se construyó el prototipo funcional para la detección automática de mensajes de phishing orientados a la protección de adultos mayores en Colombia. El sistema se desarrolló siguiendo el enfoque CRISP-DM y empleando Python como lenguaje principal para la implementación.

4.1 Diseño del Sistema

El diseño del sistema se estructuró como un pipeline de procesamiento textual compuesto por tres módulos principales:

a) Preprocesamiento y representación del texto

Se diseñó un mecanismo de transformación lingüística basado en TF-IDF, utilizando n-gramas de 1 y 2 palabras para capturar patrones típicos del lenguaje fraudulento (por ejemplo, “urgente”, “verifique su identidad”, “actualice aquí”).

Este vectorizado se seleccionó por ser:

- interpretable,
- liviano,
- eficiente para textos cortos como correos o mensajes.

b) Clasificador

El sistema emplea un modelo Random Forest, elegido por su buen desempeño en problemas binarios con variables textuales, su estabilidad ante ruido y su facilidad para obtener explicaciones basadas en importancia de características.

c) Módulo de explicabilidad

Se incorporó un submódulo capaz de extraer:

- las palabras que más influyen en la decisión del modelo,
- mensajes explicativos redactados para adultos mayores, siguiendo principios de diseño accesible y lenguaje sencillo.

4.2 Desarrollo del Prototipo

El desarrollo se efectuó mediante un enfoque iterativo que incluyó:

a) Generación o carga del dataset

El sistema puede cargar un dataset real (CSV), pero también incluye un mecanismo de generación sintética orientada a adultos mayores, útil cuando se requieren pruebas controladas o cuando no se dispone de datos etiquetados.
Los mensajes generados incluyen temas habituales para la población adulta mayor:

- pago de pensión,
- bancos,
- urgencias falsas,
- soporte técnico,
- beneficios del gobierno.

Esto permite una validación inicial del modelo.

b) Entrenamiento del pipeline

El desarrollo incluyó:

- división estratificada 80/20,
- entrenamiento de TF-IDF + RandomForest en un único pipeline,
- generación automática de métricas: F1, precisión, recall, matriz de confusión.

c) Exportación de artefactos

Se implementó la persistencia mediante joblib:

- pipeline_tfidf_rf.pkl → modelo entrenado
- results_summary.json → métricas, features importantes, ejemplos explicados
- confusion_matrix.png → figura apta

Esto garantiza reproducibilidad y facilita el despliegue del modelo en entornos reales.

4.3 Implementación

La implementación se realizó íntegramente en Python, integrando bibliotecas como scikit-learn, pandas, matplotlib y fastapi.

El programa final consta de los siguientes componentes:

a) Pipeline de Entrenamiento

- Construye la vectorización TF-IDF.
- Entrena el clasificador RandomForest.
- Genera métricas y gráficos de evaluación.
- Extrae características relevantes para la explicabilidad.
- Guarda todos los artefactos en disco.

El código implementado garantiza:

- reproducibilidad,
- interpretabilidad,
- modularidad,
- claridad para auditoría y validación académica.

b) Módulo de Explicación (Interpretabilidad)

Para cada mensaje evaluado, el sistema devuelve:

- probabilidad de riesgo,
- clase predicha (phishing vs legítimo),
- lista de palabras sospechosas detectadas,
- una explicación en lenguaje sencillo adaptada a adultos mayores.

Esto fortalece la utilidad social del prototipo.

c) API de Inferencia (Implementación final)

Se desarrolló un endpoint de ejemplo en FastAPI:

POST /detect

```
{
   "text": "mensaje a analizar"
}
```

El servidor:

1. carga el pipeline entrenado,
2. calcula el riesgo de phishing,
3. devuelve una respuesta JSON con explicación.

Esto demuestra la capacidad del sistema para ser integrado en:

- aplicaciones móviles,
- sistemas bancarios,
- herramientas de alfabetización digital para adultos mayores,
- plataformas de ciberseguridad del sector público o privado.

DISCUSIÓN DE RESULTADOS

Los resultados obtenidos en el desarrollo del prototipo de detección de phishing mediante un modelo de aprendizaje automático muestran una tendencia coherente con lo reportado en la literatura reciente. En primer lugar, el rendimiento del clasificador Random Forest aplicado sobre representaciones TF-IDF evidencia una capacidad adecuada para identificar patrones lingüísticos asociados a intentos de fraude digital, lo cual coincide con los hallazgos presentados por Sannd y Cook [5], quienes concluyen que los correos fraudulentos comparten indicadores consistentes a nivel sintáctico y composicional. El modelo implementado en este trabajo replica esta observación al detectar características de lenguaje típicas de estafas, incluso en textos cortos.

Asimismo, los resultados permiten reforzar la postura de Sugunaraj et al. [3], quienes argumentan que la creciente sofisticación de los fraudes requiere herramientas automatizadas capaces de analizar señales textuales que muchas veces pasan desapercibidas entre adultos mayores. El desempeño del sistema construido en Python demuestra que el procesamiento automatizado puede, efectivamente, ofrecer un soporte real para esta población, alineándose con las

recomendaciones de investigación sobre prevención de fraudes a personas vulnerables.

Por otro lado, al comparar los resultados de este estudio con los reportados por Tummala et al. [1], se observa que los desafíos de diseño de sistemas de protección para adultos mayores y neurodiversos también emergen en este prototipo. Aunque el modelo logra clasificar textos con precisión aceptable, la utilidad práctica dependerá en gran medida de la integración de interfaces accesibles y mecanismos de explicación del riesgo. Esto sugiere que futuros desarrollos deben avanzar hacia modelos interpretables y sistemas interactivos que faciliten el entendimiento del usuario final, como recomiendan los autores.

El desempeño del prototipo también se relaciona con la perspectiva de Nguyen [2], quien examina cómo el edadismo impacta en el uso de tecnologías. Los resultados empíricos muestran que, aunque el modelo es técnicamente eficiente, su adopción real dependerá de que el diseño considere las limitaciones tecnológicas percibidas por los adultos mayores. Esto implica que la implementación debe complementarse con estrategias de alfabetización digital y una experiencia de usuario simplificada.

De manera similar, el enfoque propuesto se alinea con la plataforma ElderConnect descrita por Sayeed et al. [4], la cual integra IA para ofrecer asistencia preventiva a personas mayores. Aunque el sistema aquí desarrollado es más limitado y se concentra solo en la clasificación de textos, los resultados obtenidos demuestran que la detección automática puede ser un módulo clave en plataformas de mayor escala orientadas a la protección cibernética de poblaciones vulnerables.

Con respecto a los estudios revisados sobre la efectividad estadística de la IA frente al cibercrimen [7], los resultados del presente trabajo muestran que incluso modelos tradicionales como Random Forest, aplicados sobre texto procesado con TF-IDF, pueden alcanzar niveles significativos de precisión. Esto respalda la hipótesis de que técnicas de IA no necesariamente profundas pueden ser altamente útiles cuando se entrenan y se optimizan correctamente, lo cual concuerda con las tendencias estadísticas señaladas por dicho análisis.

Finalmente, los resultados del prototipo apoyan los hallazgos generales de Murugun et al. [6], quienes destacan que la inteligencia artificial es una herramienta fundamental para combatir amenazas de phishing debido a su capacidad de aprendizaje adaptativo. En este caso, el modelo entrenado muestra una mejora gradual conforme aumenta el volumen de datos procesados, evidenciando la escalabilidad del enfoque empleado.

En conjunto, la evidencia empírica obtenida demuestra que el sistema desarrollado no solo cumple con los objetivos planteados en el estudio, sino que también se posiciona de manera consistente respecto a los avances recientes presentados en la literatura. No obstante, las diferencias entre este prototipo y los sistemas más complejos analizados en los trabajos relacionados indican la necesidad de continuar la evolución del modelo hacia ambientes más accesibles, explicables y orientados a usuarios vulnerables.

Referencias

[1] P. Tummala, H. Choi, A. Gupta, T. A. Lapnas, Y. S. Chung, M. Peterson, G. Walther y H. Purohit, “Design Challenges for Scam Prevention Tools to Protect Neurodiverse and Older Adult Populations,” en *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, 2024. DOI: 10.1109/TPS-ISA62245.2024.00058.

[2] K. T. Nguyen, “Ageism and Its Impact on Information and Communications Technology Usage and Design,” en *2022 IEEE Global Humanitarian Technology Conference (GHTC)*, 2022. DOI: 10.1109/GHTC55712.2022.9911051.

[3] N. Sugunaraj, A. R. Ramchandra y P. Ranganathan, “Cyber Fraud Economics, Scam Types, and Potential Measures to Protect U.S. Seniors: A Short Review,” en *2022 IEEE International Conference on Electro Information Technology (eIT)*, 2022. DOI: 10.1109/eIT53891.2022.9813960.

[4] M. S. Sayeed, H. Tamut e I. K. Dutta, “ElderConnect: An AI-Powered Platform to Empower Seniors Against Cyberthreats,” en *2025 IEEE World AI IoT Congress (AIIoT)*, 2025. DOI: 10.1109/AIIoT65859.2025.11105307.

[5] P. Sannd y D. M. Cook, “Older Adults and the Authenticity of Emails: Grammar, Syntax, and Compositional Indicators of Social Engineering in Ransomware and Phishing Attacks,” en *2018 Fourteenth International Conference on Information Processing (ICINPRO)*, déc. 2018. DOI: 10.1109/ICINPRO43533.2018.9096878.

[6] Murugun S, Sheikh Haniah, Shambhavi M Koti, et al., “A Review on Phishing Threats and Data Security in Online Trading Systems using Artificial Intelligence Techniques,” *2024 Second International Conference on Advances in Information Technology (ICAIT-2024)*, IEEE, 2024. DOI: 10.1109/ICAIT61638.2024.10690690.

[7]Statistical Prospects of AI in Tackling Cyber Crimes, Revisión bibliográfica + análisis estadístico, 2022.