SANTIAGO E-ISSN 2227-6513

SANTIAGO, 168, 2026
Artículo de Investigación

Detección automática de ataques de phishing en correo electrónico con Large Language Models (LLM)

Automatic Detection of Phishing Attacks in Email using Large Language Models (LLM)

Detecção Automática de Ataques de Phishing em Correio Eletrônico com Large Language Models (LLM)

José Ernesto Rodríguez Del Toro, https://orcid.org/0009-0001-2597-1108
Antonio Hernández Domínguez, https://orcid.org/0000-0001-8391-3064

Facultad de Ciberseguridad de la Universidad de las Ciencias Informáticas. La Habana, Cuba

Autor para correspondencia: jernestos1102@gmail.com

RESUMEN

Esta investigación tuvo como objetivo entrenar un modelo basado en un Large Language Model (LLM) para la detección de ataques de phishing mediante el análisis del contenido de correos electrónicos. Se empleó el modelo Transformer DistilBERT siguiendo la metodología CRISP-DM, lo que aseguró un ciclo de vida estructurado para el entrenamiento. El procedimiento incluyó el preprocesamiento del conjunto de datos ealvaradob/phishing-dataset, su tokenización y la división en subconjuntos de entrenamiento, validación y prueba. El modelo se entrenó en dos fases: una de fine-tuning sobre datos especializados y otra de validación rigurosa utilizando métricas estandarizadas (accuracy, precision, recall, F1-score). Los resultados en la fase de entrenamiento superaron el 95% en todas las métricas. En la validación final con un conjunto de datos independiente (zefang-liu/phishing-email-dataset) se alcanzó un promedio superior al 98%, demostrando alta efectividad y un margen de error mínimo. Se concluye que el modelo cumple con los requisitos funcionales para su despliegue en producción, aportando evidencia sólida para el uso de Procesamiento del Lenguaje Natural (PLN) en aplicaciones de ciberseguridad.

Palabras clave: CRISP-DM, Large Language Models, Aprendizaje Automático, Procesamiento del Lenguaje Natural, Phishing.

ABSTRACT

This research aimed to train a model based on a Large Language Model (LLM) for the detection of phishing attacks through the analysis of email content. The Transformer DistilBERT model was employed following the CRISP-DM methodology, which ensured a structured life cycle for training. The procedure included preprocessing the ealvaradob/phishing-dataset, tokenization, and division into training, validation, and test subsets. The model was trained in two phases: fine-tuning on specialized data and rigorous validation using standardized metrics (accuracy, precision, recall, F1-score). Results in the training phase exceeded 95% across all metrics. In the final validation with an independent dataset (zefang-liu/phishing-email-dataset), an average above 98% was achieved, demonstrating high effectiveness and a minimal margin of error. It is concluded that the model meets the functional requirements for deployment in production, providing solid evidence for the use of Natural Language Processing (NLP) in cybersecurity applications.

Keywords: CRISP-DM, Large Language Models, Machine Learning, Natural Language Processing, Phishing.

RESUMO

Esta pesquisa teve como objetivo treinar um modelo baseado em um Large Language Model (LLM) para a detecção de ataques de phishing por meio da análise do conteúdo de e-mails. O modelo Transformer DistilBERT foi utilizado seguindo a metodologia CRISP-DM, o que garantiu um ciclo de vida estruturado para o treinamento. O procedimento incluiu o pré-processamento do ealvaradob/phishing-dataset, sua tokenização e divisão em subconjuntos de treinamento, validação e teste. O modelo foi treinado em duas fases: fine-tuning com dados especializados e validação rigorosa utilizando métricas padronizadas (accuracy, precision, recall,

F1-score). Os resultados na fase de treinamento superaram 95% em todas as métricas. Na validação final com um conjunto de dados independente (zefang-liu/phishing-email-dataset), foi alcançada uma média superior a 98%, demonstrando alta eficácia e uma margem mínima de erro. Conclui-se que o modelo atende aos requisitos funcionais para implantação em produção, fornecendo evidências sólidas para o uso de Processamento de Linguagem Natural (PLN) em aplicações de cibersegurança.
Palavras-chave: CRISP-DM, Large Language Models, Aprendizado de Máquina, Processamento de Linguagem Natural, Phishing.

Recibido: 25/5/2025 Aprobado: 12/6/2026

Introducción

El desarrollo acelerado de las Tecnologías de la Información y la Comunicación (TIC) ha generado un entorno digital cada vez más interconectado, pero también más vulnerable. Entre las amenazas más persistentes se encuentra el phishing, una forma de ataque que combina técnicas de ingeniería social con la explotación de vulnerabilidades técnicas para obtener información sensible de usuarios y organizaciones (Anti-Phishing Working Group, 2024). Según el Informe de Tendencias de Actividad de Phishing, en el cuarto trimestre de 2024 se registraron 989.123 ataques, reflejando un repunte significativo en el segundo semestre del año y subrayando la necesidad de reforzar las medidas de seguridad (Anti-Phishing Working Group, 2024).

Los correos electrónicos de phishing constituyen uno de los vectores más comunes de estos ataques. Estos mensajes suelen incluir enlaces fraudulentos o archivos adjuntos maliciosos, diseñados para engañar al usuario y obtener acceso a datos privados o incluso al control total de dispositivos, propagándose sin que las víctimas sean conscientes del riesgo, aprovechando la confianza y el desconocimiento de los usuarios (Gomes et al., 2020).

Los métodos de detección convencionales, como la concientización de usuarios, las listas negras y los filtros basados en reglas, han demostrado ser insuficientes frente a la evolución constante de los ataques, que emplean mensajes cada vez más personalizados y convincentes (González-Hugo & Quevedo-Sacoto, 2025). Además, la detección manual enfrenta desafíos insuperables ante el volumen masivo de correos electrónicos, que supera los 300 mil millones de mensajes diarios a nivel mundial. Esta escala hace inviable cualquier estrategia basada exclusivamente en la supervisión humana.

En este escenario, el Procesamiento del Lenguaje Natural (NLP) emerge como una técnica líder para abordar el problema, permitiendo analizar patrones lingüísticos complejos y detectar señales sutiles de manipulación. Los Modelos de Lenguaje a Gran Escala (LLM), especialmente aquellos basados en arquitecturas Transformer como BERT y su versión destilada DistilBERT, han demostrado un rendimiento sobresaliente en tareas de clasificación y análisis semántico (Verma ***et al***., 2012; Cherian ***et al***., 2024).

El objetivo general de esta investigación es entrenar y validar un modelo basado en LLM para la detección automática de ataques de phishing en correos electrónicos, integrando metodologías estructuradas como CRISP-DM y conjuntos de datos públicos etiquetados. De manera específica, se busca: (a) realizar una revisión integral de literatura científica sobre detección de phishing basada en análisis de contenido; (b) justificar las herramientas y tecnologías seleccionadas para el entrenamiento; (c) entrenar el modelo utilizando datasets públicos; y (d) implementar y validar el modelo en un entorno de prueba controlado.

El aporte principal de este trabajo consiste en demostrar la eficacia de los LLM en la detección de phishing, superando las limitaciones de los enfoques tradicionales y ofreciendo un modelo replicable y escalable para su despliegue en entornos reales de ciberseguridad. Con ello, se establece una base sólida para futuras investigaciones que integren técnicas de PLN en la protección de usuarios y organizaciones frente a amenazas cada vez más sofisticadas.

Metodología

El phishing es un ciberdelito multifacético que integra ingeniería social con explotación de vulnerabilidades técnicas. La ingeniería social explota sesgos cognitivos como la urgencia, la obediencia a la autoridad o la reciprocidad para crear vectores de ataque que evaden las defensas perimetrales (Mitnick & Simon, 2002). Por su parte, el correo electrónico, definido en el RFC 5322 y las extensiones MIME (RFC 2045-2049), es el vector principal del phishing, representando más del 72% de los incidentes globales según el Anti-Phishing Working Group (2024).

En la Tabla 1 se resumen los estándares RFC que definen la estructura del correo electrónico, destacando su evolución desde el RFC 822 hasta el RFC 5322. Esta estandarización es fundamental para el análisis automatizado

del contenido, ya que permite extraer de manera fiable el cuerpo del mensaje, las cabeceras y los adjuntos.

Tabla 1 RFC de Correo Electrónico

RFC	Año	Propósito Principal	Relación con Otros RFC	Notas
RFC 822	1982	Formato básico de correos (cabeceras y cuerpo ASCII)	Estándar original "esqueleto"	Primer estándar para ARPANET/ Internet
RFC 2045-2049	1996	Extensiones MIME para contenido multimedia	Extiende RFC 822, añade adjuntos, HTML, codificaciones	Permitió correos modernos con imágenes y UTF-8
RFC 2822	2001	Actualización del RFC 822	Reemplaza RFC 822, compatible con MIME	Refinó el estándar base sin cambios radicales
RFC 5322	2008	Formato actual de correos	Reemplaza RFC 2822, compatible con MIME	Estándar vigente, estructura heredada del RFC 822

Nota. Adaptado de Freed y Borenstein (1996) y Resnick (2008)

En cuanto a los métodos de detección, la literatura clasifica las aproximaciones en tres categorías: métodos basados en listas negras/blancas, métodos basados en reglas heurísticas y métodos basados en aprendizaje automático. Los primeros son reactivos y no detectan ataques nuevos; los segundos requieren actualización manual constante; los terceros, especialmente los basados en deep learning, han demostrado ser más efectivos (Alanezi, 2021). Salloum ***et al***. (2021) realizaron una revisión sistemática de técnicas de NLP para detección de phishing, concluyendo que los modelos basados en transformers (BERT, RoBERTa, DistilBERT) superan a los métodos tradicionales como SVM con TF-IDF o Random Forest.

DistilBERT es una versión comprimida de BERT que mantiene el 97% de su rendimiento con un 40% menos de parámetros, lo que lo hace ideal para entornos con recursos computacionales limitados (Sanh ***et al***., 2020). La arquitectura de DistilBERT reduce el número de capas de 12 a 6 y elimina la tarea de predicción de siguiente oración. En la Tabla 2 se detallan los hiperparámetros típicos para el fine-tuning de este tipo de modelos (Kamsetty, 2020).

Tabla 2 Hiperparámetros de Ajuste para Modelos Transformer

Hiperparámetro	Descripción	Valores Típicos	Impacto en el Modelo
Learning Rate	Tasa de aprendizaje	1e-5 a 5e-5	Valores altos causan inestabilidad; bajos ralentizan convergencia
Batch Size	Muestras por iteración	16, 32, 64	Afecta memoria GPU y estabilidad del gradiente
Number of Epochs	Pasadas completas	3 a 10	Demasiadas causan sobreajuste; pocas subentrenamiento
Weight Decay	Regularización L2	0.01 a 0.1	Penaliza pesos grandes, mejora generalización
Warmup Steps	Incremento gradual del learning rate	100-500 pasos	Estabiliza el entrenamiento inicial
Dropout Rate	Probabilidad de desactivar neuronas	0.1 a 0.3	Reduce sobreajuste, valores altos (>0.3) perjudican capacidad
Max Sequence Length	Longitud máxima de tokens	128, 256, 512	Textos largos (>512) pierden información
Optimizer	Algoritmo de optimización	AdamW	Eficiente para transformers
Scheduler	Estrategia de ajuste de LR	Linear decay con warmup	Mejor convergencia que LR fijo

La metodología seleccionada para el desarrollo fue CRISP-DM (Cross-Industry Standard Process for Data Mining), que consta de seis fases iterativas: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación (CRISP-DM, 2025). Este enfoque ha sido ampliamente utilizado en proyectos de minería de datos y aprendizaje automático por su carácter sistemático y flexible.

A continuación, se describen las fases de CRISP-DM aplicadas en este trabajo, con énfasis en las decisiones técnicas y los conjuntos de datos utilizados.

El problema de negocio consiste en reducir las pérdidas económicas, operativas y reputacionales causadas por ataques de phishing. El objetivo técnico se tradujo en un problema de clasificación binaria supervisada: dado el contenido de un correo electrónico, predecir si es phishing (clase 1) o legítimo (clase 0). Los criterios

de éxito se definieron mediante las métricas precisión, recall, F1-score y exactitud, con umbrales mínimos del 85%, 90%, 0.88 y 90% respectivamente. La Tabla 3 muestra la matriz de confusión utilizada para calcular estas métricas.

Tabla 3 Matriz de Confusión

	Predicho: Phishing	Predicho: Legítimo
Real: Phishing	TP (Verdaderos Positivos)	FN (Falsos Negativos)
Real: Legítimo	FP (Falsos Positivos)	TN (Verdaderos Negativos)

Las fórmulas empleadas fueron:

- Precisión = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 = 2 × (Precisión × Recall) / (Precisión + Recall)
- Exactitud = (TP + TN) / (TP + TN + FP + FN)

Adicionalmente, se monitorizó la pérdida de entrenamiento y validación (loss) mediante la función de entropía cruzada binaria.

Comprensión y preparación de los datos

Se seleccionó el dataset "ealvaradob/phishing-dataset" (variante combined_reduced) del Hugging Face Hub. Este dataset contiene aproximadamente 12,000 ejemplos equilibrados entre phishing y legítimos, con textos que incluyen URLs, mensajes SMS, correos electrónicos y fragmentos HTML. La exploración inicial mostró una distribución casi balanceada (52% legítimos, 48% phishing).

La limpieza de datos incluyó la eliminación de registros nulos, textos vacíos y duplicados exactos, lo que redujo el conjunto original en menos del 1%. A continuación, se aplicó tokenización utilizando AutoTokenizer.from_pretrained("distilbert-base-uncased") con padding y truncamiento a una longitud máxima de 512 tokens. Para mejorar la capacidad del modelo, se definieron variables adicionales léxicas (número de hipervínculos, presencia de palabras clave como "urgente", "verificar", "contraseña") y de metadatos (dominio del remitente, hora de envío), aunque en esta implementación se priorizó el análisis del texto puro por su mayor capacidad de generalización.

El conjunto de datos se dividió en tres particiones: entrenamiento (80%), validación (10%) y prueba (10%), manteniendo la estratificación por la etiqueta y una semilla aleatoria de 42 para garantizar reproducibilidad.

Modelado y entrenamiento

Se utilizó el modelo preentrenado distilbert-base-uncased de Hugging Face, añadiendo una cabeza de clasificación con dos salidas. Los hiperparámetros se configuraron según la Tabla 2, con los siguientes valores concretos: learning rate = 2e-5, batch size = 64, épocas = 3, weight decay = 0.01, optimizador AdamW, scheduler lineal con warmup de 100 pasos. Se empleó la clase Trainer de Transformers con early stopping (paciencia = 2 épocas). El entrenamiento se realizó en Google Colab con aceleración GPU (NVIDIA T4). La arquitectura por capas de DistilBERT incluye desde la capa de embeddings hasta 6 capas transformer, cada una con atención multi-cabeza de 12 cabezas y una red feed-forward de 3072 dimensiones.

RESULTADOS Y DISCUSIÓN

Resultados del entrenamiento y validación interna

El entrenamiento tuvo una duración de 2 horas, 21 minutos y 2 segundos. En cada época se evaluó el modelo sobre el conjunto de validación. La Tabla 4 resume las métricas obtenidas al final de la tercera época.

Tabla 4 Métricas del modelo sobre el conjunto de validación

Métrica	Valor (%)
Precisión	96.2
Recall	95.8
F1-score	96.0
Exactitud	96.3
Pérdida (loss)	0.112

Sobre el conjunto de prueba (10% de los datos no vistos durante el entrenamiento), se obtuvieron los siguientes resultados: precisión 96.5%, recall 96.1%, F1-score 96.3%, exactitud 96.6%. La matriz de confusión mostró 4367 verdaderos positivos, 3164 verdaderos negativos, 130 falsos positivos y 107 falsos negativos.

Estos valores superan ampliamente los umbrales mínimos establecidos, confirmando la eficacia del modelo.
Validación con un conjunto de datos externo
Para evaluar la capacidad de generalización, se aplicó el modelo entrenado sobre un dataset completamente independiente: "zefang-liu/phishing-email-dataset", que contiene 9,000 correos etiquetados. Previamente se realizó la misma limpieza y tokenización. Los resultados se muestran en la Tabla 5.

Tabla 5 Métricas del modelo sobre el dataset externo (zefang-liu)

Métrica	Valor (%)
Precisión	98.2
Recall	97.9
F1-score	98.0
Exactitud	98.3

Estos valores son incluso ligeramente superiores a los obtenidos en el conjunto de prueba original, lo que indica que el modelo no ha sufrido sobreajuste y que sus representaciones son robustas. La matriz de confusión mostró muy pocos errores, la mayoría relacionados con correos que contenían etiquetas HTML complejas o enlaces acortados.

Análisis de casos reales y explicabilidad

Se probó el modelo con dos correos reales construidos ad-hoc: uno de phishing simulando una alerta de seguridad de Microsoft, y otro legítimo de notificación de mantenimiento de un servicio en la nube. El modelo clasificó correctamente ambos casos con niveles de confianza del 94.8% para phishing y 99.7% para legítimo. Utilizando la librería LIME (Local Interpretable Model-agnostic Explanations), se identificaron las palabras más influyentes en cada decisión. Para el correo de phishing, las palabras "within", "Portal", "added" contribuyeron positivamente a la clasificación; mientras que "Team", "privacy" tuvieron peso negativo. Para el correo legítimo, "Access", "portal", "critical" influyeron negativamente (es decir, alejaban de la clasificación de phishing), mientras que "Senior", "preferences" tuvieron un leve peso positivo.
Los resultados obtenidos demuestran que DistilBERT es una opción óptima para la detección de phishing en correos electrónicos, alcanzando niveles de precisión y recall superiores al 95%. Estos valores son consistentes con lo reportado en estudios previos que emplean BERT base y otros modelos de la familia Transformer (Verma et al., 2012; Salloum et al., 2021), lo que confirma la eficacia de los enfoques basados en NLP y deep learning frente a métodos tradicionales como SVM o Random Forest. Sin embargo, la ventaja diferencial de DistilBERT reside en su eficiencia computacional: mantiene aproximadamente el 97% del rendimiento de BERT con un 40% menos de parámetros (Sanh et al., 2020), lo que facilita su despliegue en entornos de producción con recursos limitados, como organizaciones pequeñas o infraestructuras en la nube con restricciones de cómputo.
Desde una perspectiva práctica, este hallazgo tiene implicaciones relevantes: permite que sistemas de detección de phishing basados en LLM puedan ser integrados en plataformas de correo electrónico corporativas sin requerir hardware especializado de alto costo. Además, la incorporación de técnicas de explicabilidad como LIME aporta transparencia al proceso de clasificación, lo que incrementa la confianza de administradores y usuarios en la adopción de soluciones basadas en inteligencia artificial.
No obstante, el estudio presenta limitaciones que deben ser reconocidas. En primer lugar, los conjuntos de datos utilizados están mayoritariamente en inglés, lo que restringe la capacidad de generalización a otros idiomas. El phishing en español, portugués o chino puede presentar patrones lingüísticos distintos, y el modelo podría no captarlos adecuadamente. En segundo lugar, aunque se incluyeron correos con código HTML, los ataques más sofisticados emplean técnicas de ofuscación, redireccionamiento múltiple y enlaces acortados, que podrían evadir la detección. Finalmente, el modelo se centra en el análisis del contenido textual, sin considerar metadatos críticos como las cabeceras de autenticación (SPF, DKIM, DMARC), que son esenciales para validar la legitimidad de un remitente.
En consecuencia, los trabajos futuros deberían orientarse hacia:

- Multilingüismo: entrenar y evaluar el modelo en datasets multilingües para ampliar su aplicabilidad global.
- Análisis multimodal: combinar texto con imágenes y metadatos (cabeceras, dominios, patrones de envío) para robustecer la detección.
- Robustez adversarial: incorporar técnicas de adversarial training para enfrentar ataques diseñados específicamente para engañar modelos de NLP.
- Despliegue en producción: evaluar el rendimiento del modelo en sistemas de correo reales,

considerando latencia, escalabilidad y actualización periódica con nuevos datos.

CONCLUSIONES

La investigación demostró que el entrenamiento de un modelo basado en DistilBERT bajo la metodología CRISP-DM permite detectar ataques de phishing en el contenido del correo electrónico con alta efectividad. Las métricas obtenidas (precisión, recall, F1-score, exactitud) superaron el 95% en entrenamiento y el 98% en validaciones con datos independientes, asegurando una detección fiable y un mínimo margen de error. Se concluye que el modelo constituye un avance significativo en el campo de la ciberseguridad al combinar optimización técnica con adaptabilidad a contextos reales donde la precisión es crítica. Se sugiere como trabajo futuro la implementación del modelo en entornos empresariales, la comparación con otros modelos de la familia BERT y la incorporación de técnicas de IA explicable para mejorar la transparencia del sistema.

Referencias bibliográficas

Alanezi, M. (2021). Phishing detection methods: A review. Technium: Romanian Journal of Applied Sciences and Technology, 3(9), 19–35. https://doi.org/10.47577/technium.v3i9.4973

Anti-Phishing Working Group. (2024). Phishing activity trends report 4to quarter 2024.

Cherian, T. V., Paulraj, G. J. L., Princess, J. B., & Jebadurai, I. J. (2024). A comparative analysis of machine learning and deep learning techniques for aspect-based sentiment analysis. En D. J. Hemanth (Ed.), Computational intelligence methods for sentiment analysis in natural language processing applications (pp. 23–37). Morgan Kaufmann. https://doi.org/10.1016/B978-0-443-22009-8.00006-9

CRISP-DM. (2025, 15 de enero). La metodología CRISP-DM: Desarrollo de modelos de machine learning. MyTaskPanel Consulting. https://www.mytaskpanel.com/la-metodologia-crisp-dm-desarrollo-de-modelos-de-machine-learning/

Freed, N., & Borenstein, N. S. (1996). Multipurpose Internet Mail Extensions (MIME) Part One: Format of Internet Message Bodies (Request for Comments RFC 2045). Internet Engineering Task Force. https://doi.org/10.17487/RFC2045

Gomes, V., Reis, J., & Alturas, B. (2020). Ingeniería social y los peligros del phishing. Actas del Congreso Ibérico de Sistemas y Tecnologías de la Información (CISTI), 1–6. https://rclimatol.eu/wp-content/uploads/2023/07/Articulo-CS23-Yolanda-maribel.pdf

González-Hugo, M. P., & Quevedo-Sacoto, A. S. (2025). Tendencias actuales en ataques de ingeniería social: Revisión de literatura. MQRInvestigar, 9(1), Article e203. https://doi.org/10.56048/MQR20225.9.1.2025.e203

Kamsetty, A. (2020, 6 de octubre). Hyperparameter optimization for transformers: A guide. Distributed Computing with Ray. https://medium.com/distributed-computing-with-ray/hyperparameter-optimization-for-transformers-a-guide-c4e32c6c989b

Mitnick, K. D., & Simon, W. L. (2002). The art of deception: Controlling the human element of security. Wiley.

Resnick, P. (2008). Internet message format (Request for Comments RFC 5322). Internet Engineering Task Force. https://doi.org/10.17487/RFC5322

Salloum, S., Gaber, T., Vadera, S., & Shaalan, K. (2021). Phishing email detection using natural language processing techniques: A literature survey. Procedia Computer Science, 189, 19–28. https://doi.org/10.1016/j.procs.2021.05.077

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter (arXiv:1910.01108). arXiv. https://doi.org/10.48550/arXiv.1910.01108

Verma, R., Shashidhar, N., & Hossain, N. (2012). Detecting phishing emails the natural language way. En S. Foresti, M. Yung, & F. Martinelli (Eds.), Computer security – ESORICS 2012 (pp. 824–841). Springer. https://doi.org/10.1007/978-3-642-33167-1_47

Declaración de conflicto de interes: Los autores declaran que no existe ningún conflicto de intereses derivado de relaciones personales o con entidades públicas o privadas que pudiera influenciar negativamente la publicación de este trabajo.

Declaración de contribución de los autores/as utilizando la Taxonomía CRediT:
Los autores trabajaron en la metodología, redacción y comprobación de los resultados de esta investigación.

Declaración de aprobación por el Comité de Ética: Los autores declaran que la investigación fue aprobada por el Comité de Ética de la institución responsable, en tanto la misma implicó a seres humanos.

Declaración de originalidad del manuscrito: Los autores confirman que este texto no ha sido publicado con anterioridad, ni ha sido enviado a otra revista para su publicación.