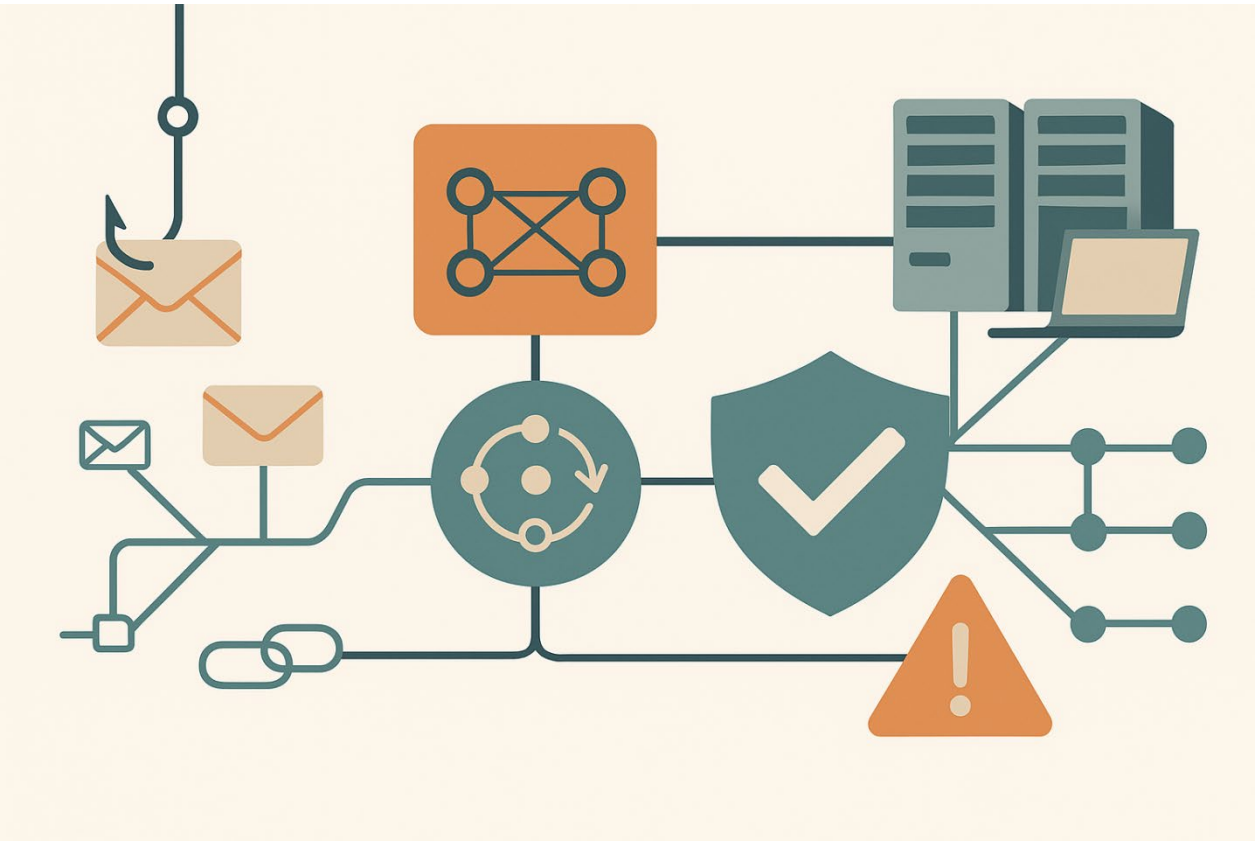
RIMI
Revista de Investigación Multidisciplinaria, Iberoamericana

AÑO 3. No. 3

ISSN:2992-7978



MODELO DE APRENDIZAJE PROFUNDO HIBRIDO PARA LA DETECCIÓN DE AMENAZAS DE PHISHING EN REDES ORGANIZACIONALES

Enrique Torres Romero[1]

[1] Doctorando Universidad Americana de Europa (UNADE)

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

AÑO 3. No. 3

ISSN:2992-7978

RESUMEN

Ante la creciente amenaza de ataques de suplantación de identidad (phishing) en las redes informáticas organizacionales, se desarrolló un modelo híbrido de aprendizaje supervisado para la detección automatizada y en tiempo real de correos electrónicos y sitios web maliciosos. Este enfoque buscó superar las limitaciones de los métodos tradicionales frente a ataques sofisticados, incluyendo aquellos impulsados por inteligencia artificial generativa. La investigación se centró en el diseño y evaluación de este modelo, abarcando la selección de datos, algoritmos y herramientas informáticas. La metodología incluyó el preprocesamiento del conjunto de datos de Mendeley Data, la estructuración del modelo, y las fases de entrenamiento, validación y pruebas. Para su desarrollo, se utilizaron Python, sus librerías, y Google Colab. El modelo alcanzó una exactitud de 0.9690, un ROC-AUC de 0.9951 y MCC de 0.9381. Se concluyó que la arquitectura seleccionada representó una opción eficiente y factible para su uso práctico en ambientes empresariales, fortaleciendo la resiliencia ante amenazas de phishing.

Palabras clave: Aprendizaje Profundo, Ciberseguridad, Detección de Phishing.

ABSTRACT

Given the increasing threat of identity spoofing (phishing) attacks on organizational computer networks, a hybrid supervised learning model was developed for the automated, real-time detection of malicious email and websites. This approach sought to overcome the limitations of traditional methods against sophisticated attacks, including those driven by generative artificial intelligence. The research focused on the design and evaluation of this model, covering data, algorithm, and software tool selection. The methodology included preprocessing the Mendeley Data dataset, structuring the model, and the training, validation, and testing phases. Python, its libraries, and Google Colab were used for its development. The model achieved an accuracy of 0.9690, an ROC-AUC of 0.9951, and an MMC of 0.9381. It was concluded that the selected architecture represented an efficient and feasible option for practical use in enterprise environments, strengthening resilience against phishing threats.

Keywords: Deep Learning, Cibersecurity, Phishing Detection.

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

INTRODUCCIÓN

El aumento exponencial del comercio electrónico a través de tiendas en línea ha traído un sinfín de beneficios al sector empresarial; pero también la exposición a diversas formas de amenazas, en específico la suplantación de identidad (del inglés "phishing") de correos electrónicos y sitios web corporativos indispensables en una organización.

La ciberseguridad se ha priorizado en las organizaciones, ante la digitalización de los entornos comerciales y las amenazas del phishing manifestado a través de mensajes de texto, llamadas telefónicas, correos electrónicos, sitios web, entre otros.

La evolución del phishing desde sus comienzos, que se caracterizaba por un lenguaje persuasivo y errores ortográficos, ha avanzado hasta incluir métodos más sofisticados, como la integración de la inteligencia artificial generativa. Esta tecnología permite el envío masivo y automatizado de ataques, lo que limita aún más la eficacia de los sistemas tradicionales de defensa, que se basan en reglas y listas, sobre todo en la detección de ataques de día cero y sus variantes. Esto evidencia la fragilidad de la seguridad en las redes informáticas, poniendo en riesgo tanto los activos como la reputación de las empresas.

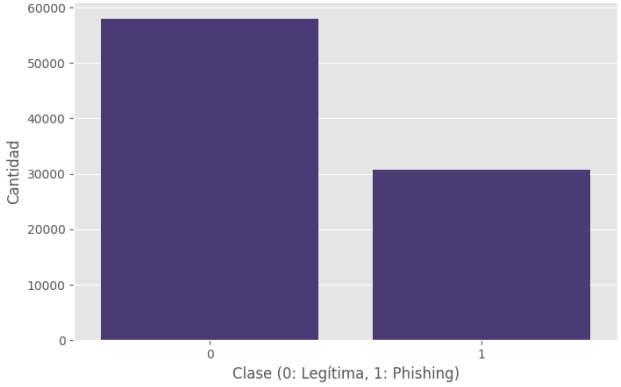
Si bien los sistemas de detección basados en aprendizaje automático muestran una efectividad aceptable en ciertos casos, a menudo requieren intervención manual constante. En contraste, los modelos de aprendizaje profundo se distinguen por su capacidad de extraer información de conjuntos de datos de forma automática, detectando e identificando patrones maliciosos con mínima intervención humana.

La propuesta de investigación se centra en el diseño de un modelo hibrido, constituido por dos algoritmos principales desarrollando procesos en independiente y fusionado salidas para una predicción final: redes neuronales convolucionales (CNN) y redes de memoria a corto y largo plazo bidireccional (BiLSTM), para aprovechar las ventajas que los caracterizan y una mayor capacidad de generalización en comparación con las tecnologías tradicionales que les preceden.

La experimentación con el conjunto de datos reales de Mendeley Data, integrado por 88,647 instancias de localizador de recursos uniforme (URLs.) y páginas HTML (legitimas "0": 58,000 y phishing "1": 30647) como se observa en la **Figura 1**, permite garantizar la validez del proceso de aprendizaje del modelo durante su entrenamiento.

Figura 1

Distribución de Clases en el Conjunto de Datos



Nota: elaboración propia con datos de Mendeley Data

La evaluación del rendimiento del modelo en la detección del phishing analiza métricas de arquitecturas similares que han tenido éxito y fallos en la identificación de patrones de falsos positivos y

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under
CC BY-NC-SA 4.0

ISSN:2992-7978

negativos, lo que permite conocer limitaciones y posibles mejoras de optimización, reflejando mayor fiabilidad del modelo al momento de evaluarse en un entorno simulado.

El riesgo exponencial de los activos empresariales ante las amenazas cada vez más sofisticadas del phishing, hace urgente y prioritaria la implementación de un sistema de detección inteligente, automatizado y escalable que opere en tiempo real como el propuesto, que fortalezca las medidas de seguridad de las redes informáticas en el entorno organizacional.

El objetivo general de esta investigación es proponer y evaluar experimentalmente un modelo híbrido de aprendizaje profundo automatizado que permita gestionar la detección de amenazas de phishing en un entorno simulado. Para alcanzarlo, se han identificado cuatro objetivos especifico: primero, identificar el conjunto de datos; segundo, seleccionar los algoritmos adecuados que integren el modelo hibrido; tercero, seleccionar las herramientas informáticas óptimas para el entrenamiento y evaluación; y cuarto, evaluar el modelo en un entorno de simulación para probar su efectividad.

La propuesta del proyecto de investigación plantea que la creación de un modelo híbrido para la administración e identificación de amenazas de phishing fortalece considerablemente las medidas de seguridad en las redes informáticas de una entidad conectada a internet.

MARCO TEÓRICO

La eficacia del phishing, según (Jakobsson & Myers, 2007, p. 28), se basa en el nivel de manipulación psicológica de la víctima y en las vulnerabilidades técnicas. Para contrarrestarlo o eliminarlo, las medidas a tomar requieren la combinación de diversos elementos, que incluyen tecnologías, procesos, educación del usuario y una investigación constante sobre su evolución técnica.

La ciberseguridad empresarial es determinante en la protección de los activos digitales, así lo expresa (Donaldson et al., 2015), quien la define como una estrategia integral que involucra individuos, procesos y herramientas, fundamentales en la construcción de una férrea defensa y solidez en una organización.

El aprendizaje automático se ha hecho una herramienta indispensable en la ciberseguridad al interior de las organizaciones como lo señala (Thomas et al., 2019), dado el constante aumento y complejidad de los datos que se manejan. Permite la detección automatizada de amenazas potenciales de malware, anomalías y patrones maliciosos en forma rápida y eficiente, transformando y reforzando la ciberseguridad.

Es fundamental el seguimiento y cumplimiento de las mejores prácticas en el desarrollo de los modelos de aprendizaje profundo, así lo expresa (Ketkar et al., 2021) para obtener mejores resultados en su implementación y a futuro.

Las CNN y las LSTM son fundamentales en la estructura de los modelos de aprendizaje profundo en la detección de patrones maliciosos en grandes conjuntos de datos, así lo indica (Goodfellow et al., 2016) al permitir a los modelos el aprendizaje de representaciones complejas y abstractas de los datos, esencial para la extracción de información significativa y descubrimiento de patrones ocultos.

ESTADO DEL ARTE

Zhang et al. (2021) propone un modelo híbrido, donde segmenta la URL en palabras clave y las transforma en una matriz de vectores para alimentar la CNN, lo que permite extraer características locales y presentarlas como entrada al BiLSTM. Así se obtienen características dependientes de memoria a largo plazo de manera bidireccional, alcanzando una precisión de 0.9884.

La propuesta del modelo hibrido CNN-LSTM de (Ujah-Ogbuagu et al., 2024) busca combinar y mejorar capacidades en la detección de URLs. maliciosas; la CNN empeñada en la extracción de

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under
CC BY-NC-SA 4.0

ISSN:2992-7978

características textuales y por otra parte la LSTM aprender y memorizar relaciones contextuales secuenciales, con una exactitud del 0.9890 y 0.9680 al entrenarse por separado en dos conjuntos de datos.

En sus trabajos de investigación (Uzoaru et al., 2024) entrena las CNN y LSTM en forma independiente y combinada con un conjunto de datos de 20,000 URLs., (phishing y legitimas) y, hace comparaciones en rendimiento con una exactitud del 0.9920, 0.9680 y 0.976 respectivamente.

El modelo hibrido propuesto por (Zonyfar et al., 2023) involucra la combinación de una CNN y una LSTM sumando ventajas en la distinción de URLs. legitimas y phishing. También entrena en forma individual ambos algoritmos involucrados en la combinación, logrando métricas de rendimiento superiores en el entrenamiento del modelo combinado.

Las CNN y LSTM, funcionan de forma independiente, sus salidas se concatenan en la capa sigmoide, para la clasificación final de la extracción de características de HTML y URLs de modo que se obtiene un rendimiento de 0.9834 (Ariyadasa et al., 2020).

Se presenta un enfoque que utiliza un modelo LSTM con GCM analizando independientemente las URLs y el contenido HTML para finalmente combinarlo y obtener una predicción final de 0,9642 (Ariyadasa et al., 2022).

Una estrategia multimodal que se apoya en la configuración CNN-LSTM, FastText-BiLSTM y BERT-MLP para la detección de las características de las URLs, del código JavaScript y el contenido de las páginas web maliciosas, combinándolo finalmente a través de fusión tardía con un acierto del 0,9790 (Belfedhal, 2023).

El enfoque multimodal aplicado en la detección del phishing, que combina un modelo de integración de datos y una arquitectura híbrida de red neuronal convolucional en conjunción a capas densidad de análisis, se combina con fusiones en una capa compartida con un acierto de 0,9800 (Ali et al., 2024).

METODOLOGÍA

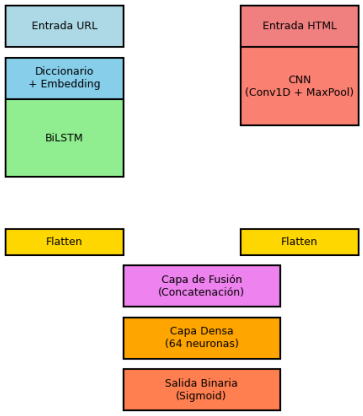
Método y técnicas

La metodología sugerida se enfocó en la creación de un modelo de identificación de phishing que utilizó aprendizaje profundo supervisado, destinado a clasificar de forma binaria correos electrónicos y sitios web phishing o legítimos.

En la **Figura 2**, se muestra el núcleo de la solución que consistió en una arquitectura mixta que combina una CNN y una BiLSTM, utilizando sus capacidades complementarias para realizar un análisis multimodal de datos.

Figura 2

Arquitectura del Modelo Híbrido CNN-BiLSTM para Detección de Phishing



Nota: elaboración propia

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

ISSN:2992-7978

La CNN se ocupó de extraer características espaciales del contenido HTML de las páginas web, mientras que la BiLSTM manejó las secuencias de caracteres para captar dependencias temporales y contextuales.

El procedimiento de preprocesamiento del conjunto de datos se dividió en dos flujos principales:

Las técnicas de las que se sirvió el contenido HTML para aplicarles limpieza incluyeron todas las técnicas de eliminación de ruido, etiquetas, comentarios, scripts y estilos, al igual que la extracción y normalización del texto.

En el caso de URLs se preprocesó asegurando que las cadenas fueran tanto completas como normalizadas, prestando atención muy explícita en el manejo de protocolos y decodificación. Se tokenizaron a nivel carácter y se transformaron en vectores numéricos de tamaño fijo a través de una capa de incrustación (embedding), imprescindible para la entrada de BiLSTM.

Para el problema del desequilibrado del conjunto de datos se aplicó el método SMOTE-Tomek Link. El resultado de esto es la combinación de sobremuestreo de la clase menor (phishing) y el submuestreo de la clase mayor (legítima), que demostró incrementar significativamente las capacidades del modelo de la tarea de identificar con exactitud las URLs de phishing y, en consecuencia, mejorar el rendimiento global.

A la vez, se aplicó el Método de Importancia de Permutación (PIM) para la selección de las características más relevantes de las URLs con la intención de disminuir la dimensionalidad de los datos y mejorar la eficiencia y precisión del modelo, desechando atributos redundantes o irrelevantes.

La arquitectura de la CNN se formó por varias capas convolucionales, capas de normalización por lotes y capas de max-pooling, para dar paso a una capa Flatten destinada a preparar la salida.

La arquitectura BiLSTM se constituyó por capas apiladas, que incluyeron celdas de memoria con puertas de entrada y de olvido para lograr capturar las dependencias a largo plazo dentro de las secuencias.

Las salidas flatten de la CNN y la LSTM se unieron en una capa de fusión que alimentó a una capa densa, que aprendió las características combinadas y que acabó en una capa de salida con una única neurona y activación sigmoide para la predicción binaria definitiva.

Población y muestra

La muestra de datos del estudio se conformó por un conjunto de datos de sitios web de phishing de Mendeley, que abarca un total de 88,647 instancias, de las cuales 58,000 correspondieron a sitios web legítimos y 30,647 a sitios web de phishing, incluyendo la URL y el contenido HTML de cada instancia que formó parte del conjunto.

Para la implementación del modelo la muestra de datos se fraccionó convenientemente en las fases del conjunto entrenamiento, validación y prueba, donde el 70% fue destinado al entrenamiento del modelo, un 15% para la validación, calibración de hiperparámetros y evitar el sobreajuste, y el 15% restante para la evaluación del rendimiento final del modelo con datos no conocidos.

Experimentación

El procedimiento seguido para desarrollar dicho modelo de gestión de amenazas de suplantación de identidad se ejecutó en un entorno de experimentación controlado, el cual combinó los recursos computacionales locales de alto rendimiento con la flexibilidad brindada por las plataformas en nube para realizar un entrenamiento intensivo del modelo de aprendizaje profundo.

El proceso de desarrollo se llevó a cabo en un ciclo iterativo de fases de diseño arquitectónico, implementación del código, entrenamiento del modelo, pruebas y evaluación del rendimiento.

La fase de diseño se enfocó en la conceptualización de la arquitectura híbrida CNN-BiLSTM y la definición de los pasos de preprocesamiento de datos y la fase de implementación en el desarrollo de los algoritmos y de la infraestructura.

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

ISSN:2992-7978

Las pruebas se hicieron de forma constante utilizando el conjunto de validación para el ajuste de hiperparámetros asimismo el conjunto de prueba para la evaluación imparcial de la capacidad de generalización del modelo.

Herramientas informáticas

En el desarrollo e implementación del modelo, se utilizó Python y librerías PyTorch y Keras; además de Google Colab en la nube.

Para el preprocesamiento y manipulación del conjunto de datos, se utilizó Scikit-learn, NumPy, Matplotlib y Pandas.

Se aprovechó de unidades de procesamiento gráfico (GPU) para el aceleramiento de los cálculos involucrados.

La optimización del modelo fue la minimización de la función de pérdida, que era la entropía cruzada binaria, esencial en problemas de clasificación binaria. Se seleccionó el algoritmo optimizador Adam, eficaz en el ajuste adecuado y progresivo de la tasa de aprendizaje.

El entrenamiento se realizó con un periodo de 50 épocas, un tamaño de lote de 64 instancias y una tasa de aprendizaje inicial de 0.001. Se monitoreó la función de pérdida y precisión en los subconjuntos de entrenamiento y validación, en la detección de situaciones de sobreajuste o de subajuste.

Para evitar el sobreajuste se insertó capas de Dropout estratégicamente después de las capas de max-pooling y de las primeras capas densas con tasas de Dropout del 0.25 y del 0.5.

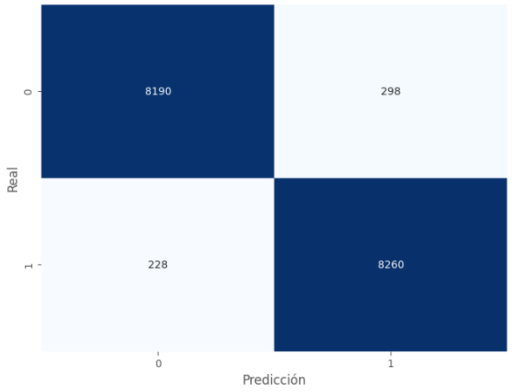
El rendimiento del modelo se evaluó con métricas derivadas de la matriz de confusión, adicionalmente, se creó la curva ROC-AUC para conocer la capacidad discriminativa del modelo en distintos puntos de corte, y se calculó el Coeficiente de Correlación de Matthews (MCC) como métrica robusta y necesaria, sobre todo por el desbalanceo de los datos.

RESULTADOS

La evaluación del desempeño del modelo híbrido CNN-BiLSTM se presenta por medio del estudio de la matriz de confusión y de métricas derivadas, de forma que se obtiene una idea exhaustiva de su capacidad de clasificación, como se visualiza en las **Figura 3**, **Figura 4**, **Figura 5** y **Figura 6**.

Figura 3

Matriz de Confusión del Modelo CNN-BiLSTM



Nota: elaboración propia con datos de Mendeley Data

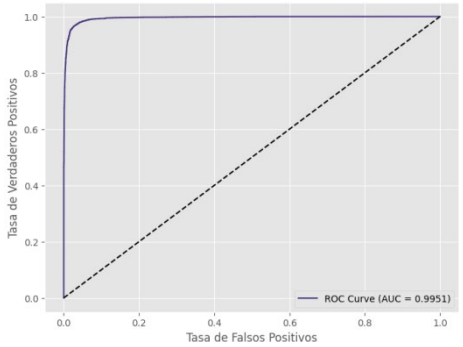
Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

ISSN:2992-7978

Figura 4

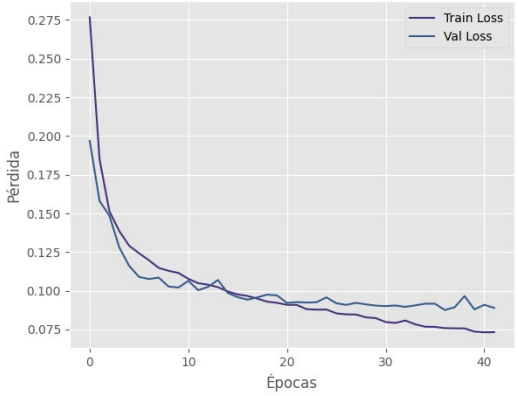
Curva ROC del Modelo CNN-BiLSTM



Nota: elaboración propia con datos de Mendeley Data

Figura 5

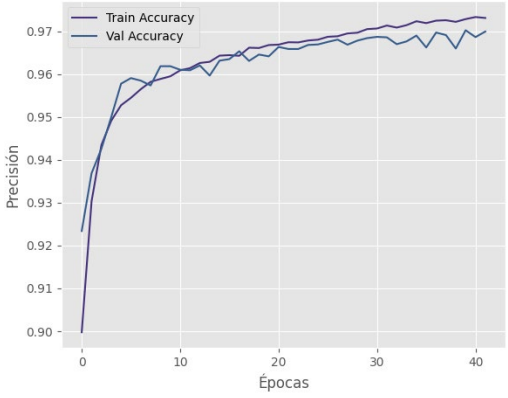
Evaluación de la Función de Perdida Durante el Entrenamiento y Validación



Nota: elaboración propia con datos de Mendeley Data

Figura 6

Evaluación de la Exactitud Durante el Entrenamiento y Validación



Nota: elaboración propia con datos de Mendeley Data

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

ISSN:2992-7978

Con el conjunto de datos de Mendeley, el modelo CNN-BiLSTM consigue los siguientes resultados de rendimiento: accuracy: 0.9690, precision: 0.9652, recall: 0.9731, F1-score: 0.9691, ROC-AUC: 0.9951 y MCC: 0.9381

DISCUSIÓN

Análisis de rendimiento

Aunque la evaluación general del modelo indica un buen desempeño, es esencial analizar cómo se comporta en tipos específicos de ataques de phishing, como aquellos dirigidos a la banca, a las redes sociales o a las instituciones gubernamentales. Sin embargo, el Conjunto de Datos de Phishing de Mendeley, en su actual formato, no categoriza las URL de phishing según sectores o tipos de objetivos (banca, redes sociales, gobiernos, etc.). Por lo tanto, el enfoque del rendimiento del modelo de la presente investigación no se realiza bajo ese marco de categorización. El análisis por tipo de phishing sería una línea sugerente para futuras investigaciones que demandaría la recogida o la anotación de conjuntos de datos que incluyeran dicha información sectorial.

Manejo del desequilibrio de datos

Uno de los problemas que pueden aparecer en los conjuntos de datos de detección de phishing es la desproporción de clases; en el caso de los conjuntos de datos de phishing, las instancias legítimas superan a las de phishing. Esto quiere decir también que puede inducir un sesgo en el modelo hasta hacerlo decantarse por la clase mayoritaria y, a la larga, alcanzar un rendimiento que no sea el óptimo en la detección de la clase minoritaria (phishing) que es sobre la que se centra el interés. Para eludir dicho sesgo se aplica estrictamente el método SMOTE-Tomek Link en la fase de preprocesado. Aunque se realiza en el periodo de preprocesado, SMOTE es un buen método para equilibrar la distribución de clases en los conjuntos de datos. El uso de este método de sobremuestreo (SMOTE) y submuestreo (Tomek Link) es necesario para evitar el sesgo hacia la clase legítima y así mejorar la capacidad de detección de URL de phishing.

De ese modo, la submuestra de las características obtenida a través de PIM permite que el modelo se concentre en los atributos más discriminativos, minimizando el efecto de las características irrelevantes o ruidosas que pueden inducir sesgos, tal y como fue indicado anteriormente. Las métricas de evaluación Recall y F1-score, junto con MCC, son fundamentales para monitorear y garantizar que el modelo no presente un sesgo significativo hacia la clase mayoritaria, ya que son mucho menos sensibles a las diferencias en el número de clases en comparación con la precisión simple.

Comparación con la literatura previa

Se presentan los resultados del modelo propuesto en comparación con investigaciones anteriores que incluyen tanto arquitecturas de aprendizaje profundo convencionales como híbridas. Al comparar modelos que utilizan CNN y LSTM, el modelo desarrollado muestra ser, por lo menos, competitivo e incluso superior en algunos casos.

Por ejemplo, un Random Forest que implementa un enfoque basado en firmas, aunque limitado para identificar nuevos sitios web maliciosos y que requiere un tiempo considerable para la ingeniería de características manual, logra una precisión de 0.9300 (Praba et al., 2024). En contraste, (Yerima & Alzaylaee, 2020) sugieren utilizar una CNN unidimensional para la extracción de características tanto de URLs como del contenido de sitios maliciosos, la cual supera a varios clasificadores tradicionales con una precisión del 0.9820 en la detección de phishing.

El estudio de (Zaimi et al., 2024) se centra en el problema de datos desbalanceados mediante la aplicación de técnicas SMOTE-Tomek Link y PIM en el preprocesamiento de datos, evaluando cuatro

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under
CC BY-NC-SA 4.0

ISSN:2992-7978

clasificadores de aprendizaje profundo, entre los que destaca un modelo CNN-LSTM con una precisión de 0.9688

(Alshdadi, 2024) propone un modelo con arquitectura LSTM-PSO que integra en su entrenamiento páginas web consideradas sospechosas debido a su contenido textual, logrando resultados prometedores con una precisión de 0.9830

La investigación (Sultana et al., 2023) abarca una CNN diseñada para identificar características correlativas locales de URLs, y LSTM para aprender dependencias semánticas, configurando una arquitectura CNN-Attention-LSTM que alcanza una precisión de 0.9700.

La consolidación de métricas de rendimiento altas, como las obtenidas por el modelo propuesto (accuracy: 0.9690, precisión: 0.9652, recall: 0.9731, F1-score: 09691, ROC-AUC: 0.9951 y MCC: 0.9381), le otorga ventajas esenciales en la detección de phishing en empresas donde los falsos negativos son muy costosos. Si bien algunos modelos previos pueden mostrar métricas ligeramente superiores en ciertos contextos, la robustez y la capacidad de generalización del modelo, potenciadas por las técnicas de preprocesamiento y la arquitectura híbrida, lo posicionan como una solución altamente competitiva y viable para entornos organizacionales.

Implicaciones de los hallazgos

En la matriz de confusión de la **Figura 3** se aprecia un buen rendimiento del modelo con el conjunto de datos de prueba, al mostrarse un número bajo de falsos positivos y falsos negativos. Esto significa que el modelo no está bloqueando erróneamente sitios legítimos, así como tampoco permite el paso de muchos sitios de phishing, lo cual es crucial para la operatividad y seguridad de una organización

La curva ROC indicada en la **Figura 4**, permite observar la alta capacidad discriminativa del modelo para distinguir entre una clase y otra, pudiendo predecir tasas altas de verdaderos positivos y tasas bajas de falsos positivos.

En la valoración de la pérdida mostrada en la **Figura 5**, se aprecia la reducción de la función de pérdida (tanto en entrenamiento como en validación), donde ambas gráficas, a medida que avanzan las épocas, revelan un adecuado aprendizaje, optimización de los pesos y una buena generalización del modelo.

La evaluación de la exactitud indicada en la **Figura 6** evidencia una mejora continua a lo largo de las épocas de las curvas de entrenamiento y validación, lo que se traduce en una mayor capacidad predictiva y en una alta precisión con respecto a datos no vistos.

Los resultados de la investigación corroboran la mayor eficacia de los modelos híbridos CNN-BiLSTM para la detección de phishing, en particular cuando se incorporan técnicas avanzadas de preprocesamiento. En este sentido, se demuestra la robustez del modelo, sobre todo en cuanto a la aplicación de técnicas de equilibrado del dato y selección de variables. La eficacia de la curva ROC-AUC en discernir al cotejar diferentes umbrales con el propósito de verificar la capacidad del modelo en clasificar, confirma su poderío para clasificar correctamente entre sitios “legítimos” y de “Phishing”. Asimismo, el cálculo del MCC, con el fin de proporcionar una métrica equilibrada y fiable del rendimiento, también se revela crucial en los casos de distribución de datos desequilibrados, similar a la que se encuentra en Mendeley, donde la métrica de precisión simple podría resultar engañosa.

Se pone de manifiesto que el balanceo de datos mediante SMOTE-Tomek Link y la selección de características mediante PIM constituyen pasos imprescindibles que aumentan considerablemente la precisión, exactitud, recall y F1 del modelo, superando así el rendimiento de este último sin optimizaciones.

La arquitectura multimodal que aborda de manera indistinta las URL y el contenido HTML permite al modelo captar un rango más amplio de patrones maliciosos, abarcando desde las irregularidades en la estructura de las URL hasta los elementos engañosos en el diseño de las páginas web.

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

ISSN:2992-7978

El modelo propuesto tiene una capacidad de generalización muy elevada, lo que le hace un candidato prometedor para su uso en entornos empresariales en la detección de amenazas en entornos simulados.

Limitaciones del estudio

A pesar del alto nivel de rendimiento general, se evidencia necesaria la optimización continua para reducir la tasa de falsos negativos (baja recuperación para URL de phishing), fundamentalmente en la detección de sitios de phishing que se asemejan mucho a URL simples y legítimas.

Un reto a raíz de la investigación es la demanda de recursos computacionales para el entrenamiento de modelos de aprendizaje profundo, que puede suponer una dificultad para su uso en entornos reales donde hay escasos recursos.

La ausencia de una categorización detallada de los ataques de phishing en el conjunto de datos de Mendeley complica el análisis del desempeño del modelo en tipos específicos de ataques, como los bancarios o en redes sociales, lo que representa una limitación para esta investigación y una futura área de trabajo.

La constante evolución de las tácticas de phishing subraya la importancia de reentrenar el modelo de manera periódica con datos actualizados para asegurar su efectividad a largo plazo.

Futuras líneas de investigación

Nuevas líneas de exploración contemplan, por ejemplo, importar el corpus de datos en la medida necesaria para poder introducir esta categorización sectorial, la búsqueda de técnicas de reentrenamiento en el sentido de contener el modelo a las tácticas que comparten el phishing que están en constante avance.

Se investiga la posibilidad de reducir los recursos computacionales para permitir un mejor rendimiento en equipos con restricciones de hardware, así como la incorporación de mecanismos de explicabilidad (XAI) para mejorar la confianza y la comprensión sobre las decisiones del modelo.

CONCLUSIONES

La presente investigación aborda el creciente y complejo peligro del phishing en las redes informáticas organizacionales, una amenaza que las soluciones de seguridad tradicionales no pueden contener por sí solas. Como consecuencia de esta problemática, se propone el desarrollo de un modelo de gestión de detección de phishing, basado en una arquitectura híbrida de Deep Learning CNN-BiLSTM, evaluado en un entorno de simulación.

Los resultados preliminares de su evaluación experimental respaldan la hipótesis principal de la investigación: el desarrollo de un modelo de gestión para la detección de phishing en un entorno de aprendizaje supervisado, preciso y confiable, basado en una arquitectura híbrida de Deep Learning CNN-BiLSTM, potencia las medidas de seguridad de las redes informáticas de una organización conectada a internet en un sentido amplio. Por lo tanto, se logra el objetivo de diseñar y evaluar la estructura de un modelo de detección de phishing en un entorno simulado.

La exhaustiva evaluación llevada a cabo en el Conjunto de Datos evidencia un rendimiento superior, con cifras como una precision de 0.9652, una accuracy del 0.9690, un recall de 0.9731 y un F1-score de 0.9691. Estos resultados se logran gracias a un minucioso proceso de implementación de técnicas avanzadas de preprocesamiento, como la técnica SMOTE-Tomek Link para equilibrar los datos y el PIM para la selección de características. Se demuestra que el equilibrio de los datos y la elección de características son procesos fundamentales que aumentan notablemente la efectividad del modelo para reconocer

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under
CC BY-NC-SA 4.0

adecuadamente las URL de phishing, contrarrestando el sesgo que se presenta en conjuntos de datos desiguales.

La arquitectura híbrida CNN-BiLSTM resulta ser muy efectiva, dado que combina las capacidades propias de las CNN para extraer patrones espaciales del contenido HTML y las URL con las capacidades de las BiLSTM para capturar dependencias temporales y contextuales en las secuencias de URL, lo que permite un análisis multimodal de gran nivel. Las métricas derivadas de la matriz de confusión, la curva ROC-AUC y el MCC, confirman la elevada capacidad discriminativa del modelo, incluso en contextos con desequilibrio de clases.

Esta investigación aporta notablemente al ámbito de la ciberseguridad al proponer una solución de detección de phishing que está preparada para superar los inconvenientes de los métodos tradicionales y la creciente sofisticación de los ataques, incluso aquellos pertinentes a inteligencia artificial generativa.

El uso de este modelo en un entorno de organización empresarial simulado representa una defensa de alto nivel que ayuda a reducir considerablemente las pérdidas económicas y de reputación provocadas por ataques de phishing.

La implementación multimodal (URL y HTML) y la supervisión avanzada del preprocesamiento hacen de este modelo un candidato prometedor para proteger información crítica y usuarios finales en entornos simulados.

REFERENCIAS

Ali, A. K., Ghaib, A. A., Al-atbee, M., & Abduljabbar, Z. A. (2024). *AMPDF: A Hybrid Deep Learning Framework for Multi-Modal Phishing Detection in Cybersecurity*. DOI: 10.52783/jisem.v10i27s.4416

Alshdadi, A. A. (2024, July). LSTM-PSO: NLP-based model for detecting phishing attacks. *In Proceedings of the First International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security* (pp. 70-79). https://aclanthology.org/2024.nlpaics-1.9/

Ariyadasa, S., Fernando, S., & Fernando, S. (2020). Detecting phishing attacks using a combined model of LSTM and CNN. *Int. J. Adv. Appl. Sci*, 7(7), 56-67. https://doi.org/10.21833/ijaas.2020.07.007

Ariyadasa, S., Fernando, S., & Fernando, S. (2022). Combining long-term recurrent convolutional and graph convolutional networks to detect phishing sites using URL and HTML. *IEEE Access, 10*, 82355-82375. https://doi.org/10.1109/ACCESS.2022.3196018

Belfedhal, A. E. (2023). Multi-modal deep learning for effective malicious webpage detection. *Revue d'Intelligence Artificielle*, 37(4), 1005. https://doi.org/10.18280/ria.370422

Donaldson, S., Siegel, S., Williams, C. K., & Aslam, A. (2015). *Enterprise cybersecurity: how to build a successful cyberdefense program against advanced threats*. Apress.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press. https://doi.org/10.4258/hir.2016.22.4.351

Jakobsson, M., & Myers, S. (Eds.). (2007). *Phishing and countermeasures: understanding the increasing problem of electronic identity theft*. Published by John Wiley & Sons, Inc., Hoboken, New Jersey, First published: 18 May 2006.

Ketkar, N., Moolayil, J., Ketkar, N., & Moolayil, J. (2021). *Deep learning with Python: learn best practices of deep learning models with PyTorch* (pp. 243-285). New York, NY, USA:: Apress. https://doi.org/10.1007/978-1-4842-5364-9

Praba, M. B., Duddukunta, K. R., Bezawada, V. S., & Addanki, S. V. (2024, August). Enhancing Web Security through Machine Learning: A Random Forest Approach to Malicious URL Detection. *In 2024 4th Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1-6). IEEE. DOI: 10.1109/ASIANCON62057.2024.10837992

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0

Sultana, R., Rahman, M. A., & Khan, M. I. (2023, December). Hybrid Model Based Phishing Websites Detection Using Deep Learning Technique. *In 2023 26th International Conference on Computer and Information Technology (ICCIT)* (pp. 1-6). IEEE. DOI: 10.1109/ICCIT60459.2023.10441639

Thomas, T., Vijayaraghavan, A. P., & Emmanuel, S. (2019). *Machine learning approaches in cyber security analytics*. Singapore: Springer. https://doi.org/10.1007/978-981-15-1706-8

Ujah-Ogbuagu, B. C., Akande, O. N., & Ogbuju, E. (2024). A hybrid deep learning technique for spoofing website URL detection in real-time applications. *Journal of Electrical Systems and Information Technology*, 11(1), 7. https://doi.org/10.1186/s43067-023-00128-8

Uzoaru, G. C., Odikwa, N. H., & Agbugba, O. A. (2024). Intelligent Phishing Website Detection Model Powered by Deep Learning Techniques. *Asian J. Res. Com. Sci, 17*(1), 71-85. DOI: 10.9734/AJRCOS/2024/v17i1414

Yerima, S. Y., & Alzaylaee, M. K. (2020, March). High accuracy phishing detection based on convolutional neural networks. *In 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS)* (pp. 1-6). IEEE. DOI: 10.1109/ICCAIS48893.2020.9096869

Zaimi, R., Hafidi, M., & Mahnane, L. (2024). *A Permutaion Importance Based feature selection method and Deep Learning Model to Detect Phishing Websites*. DOI: https://doi.org/10.21203/rs.3.rs-3943049/v1

Zhang, Q., Bu, Y., Chen, B., Zhang, S., & Lu, X. (2021). Research on phishing webpage detection technology based on CNN-BiLSTM algorithm. *In Journal of Physics: Conference Series* (Vol. 1738, No. 1, p. 012131). IOP Publishing. doi:10.1088/1742-6596/1738/1/012131

Zonyfar, C., Lee, J. B., & Kim, J. D. (2023). HCNN-LSTM: hybrid convolutional neural network with long short-term memory integrated for legitimate web prediction. *Journal of Web Engineering, 22*(5), 757-782. https://doi.org/10.13052/jwe1540-9589.2251

Revista de Investigación Multidisciplinaria Iberoamericana, RIMI © 2023 by Elizabeth Sánchez Vázquez is licensed under

CC BY-NC-SA 4.0