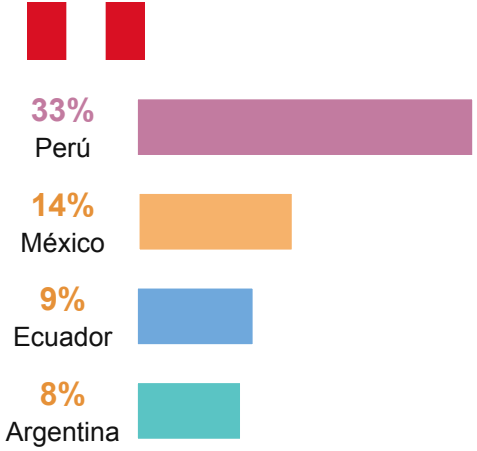
Sistema de decisión para prevenir fraude de ingeniería social

Problema

Perú es el país de América Latina que más ataques de phishing recibe con la finalidad de consolidar algún tipo de estafa (Hernandez.M,2022).

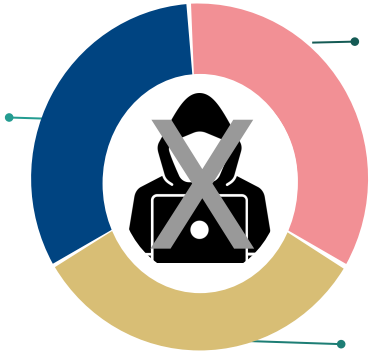


La FTC reportó que las pérdidas por fraude de phishing en 2023 superaron los 10,000 millones de dólares (FTC, 2024)

Perú en el 2023 acumuló un total de 53.8 millones de soles (Wang et al., 2024).

Motivación

-Experiencias de víctimas de phishing y smishing.
-Sofisticación de ataques y la alta demanda de soluciones efectivas.
-Educar a los usuarios para identificar y prevenir estos riesgos, promoviendo un entorno digital más seguro.



Contribución a la sociedad

-Aumenta conciencia sobre seguridad digital.
-Reducir fraudes y estafas.
-Promover un entorno digital más seguro.

Contribución a la IA

-Conjunto de datos robusto.
-Publicaciones y colaboración.
-Integración con otras tecnologías.
-Adaptación de arquitectura a entornos similares.

Contexto

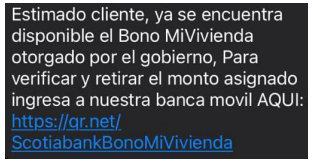
CAUSAS Y FACTORES DE PHISHING

Principales motivaciones

- Beneficios financieros
- Robo de identidad.
- Espionaje industrial.
- Distribución del malware.

Causas

- Vulnerabilidad en nuestros sistemas informáticos.
- Desinformación
- Evolución tecnológica.
- Falta de conciencia



Factores

- Las mujeres son más susceptibles al phishing.
- Individuos con competencias básicas.
- Individuos que usan Internet con más frecuencia.
- Percepción de la autoeficacia.

MÉTODOS DE PHISHING

- Vishing
- Smishing
- Phishing

(Ribeiro et al., 2024; Ticona et al., 2024)

Dataset

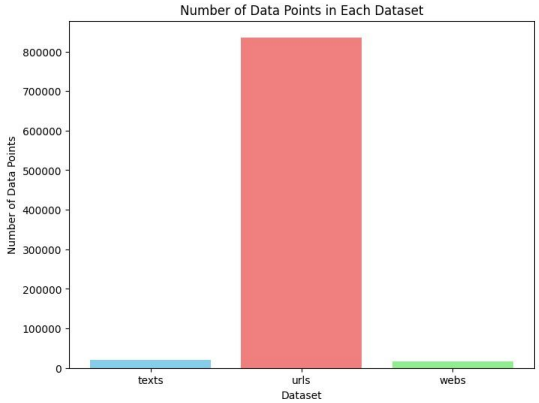
Obtención del conjunto de datos inicial X0,y0

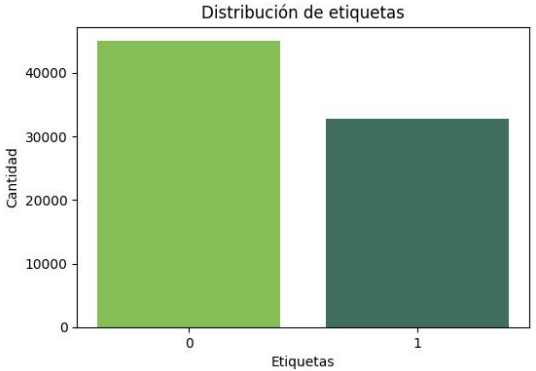
Datos PhishingTank

Datos Mendeley

Datos UCI

Datos HuggyngFace





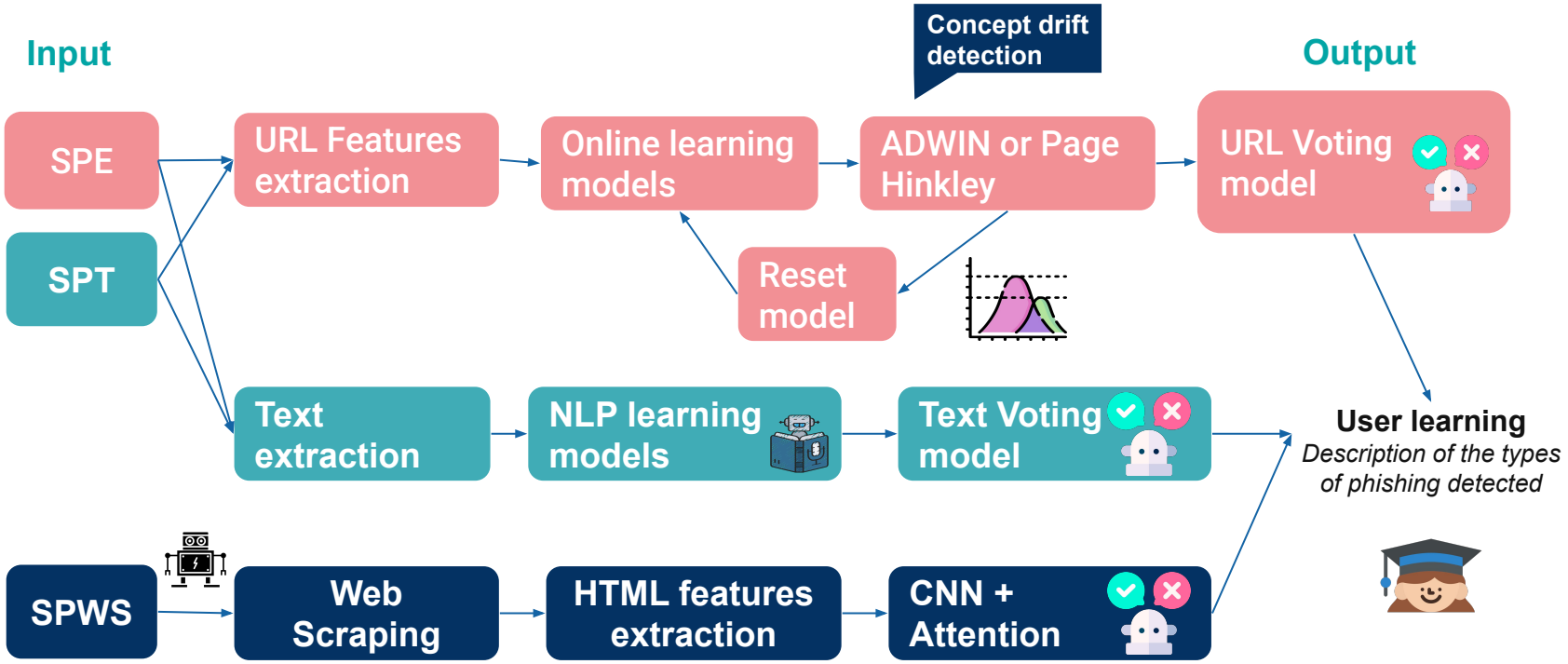
Abordar desequilibrio de datos y escasez de datos multilingüe

Google Neural Machine Translation → **Paraphrase-Generator T5 Model** → **Class weighting**

SPE: Suspicious Phishing Email + Text (18,754 samples) x5 → 93,770 samples
SPT: Suspicious Phishing url (235795 samples)
SPWS: Suspicious Phishing Website (83,251 samples) [Parafrasear código]

Existen muchos datos sobre phishing. Universidades, empresas y estados los almacenan en su área de TI o ciberseguridad.

Arquitectura de la propuesta



SPE: Suspicious Phishing Email
SPT: Suspicious Phishing Text
SPWS: Suspicious Phishing Website

Se utiliza adaptive windowing y Page-Hinkley debido que la distribución de features cambian a la par de nuevos esquemas de fraude.

PhawAI

Técnicas de inteligencia artificial

Algoritmo de aprendizaje en línea para URL

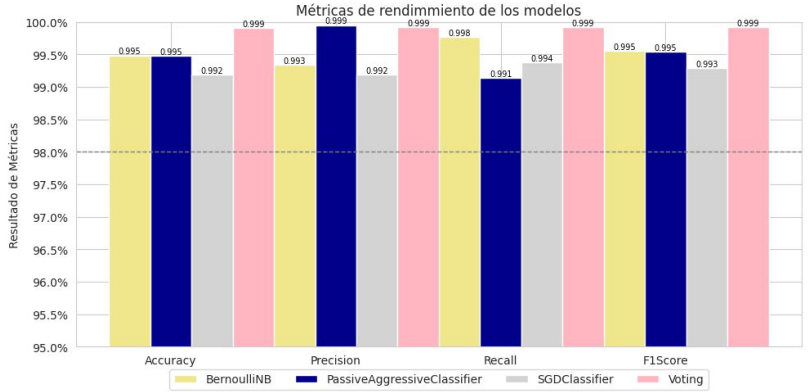
$$Voting = Ensemble([PA, BNB, SGD])$$
$$Data = Stream.iter("Phishing\ risk")$$

For X, y in Data:

$$\hat{y} = Voting(X)$$
$$Voting.Learn(X, y)$$
$$Update\ F1(y, \hat{y})$$
$$Adwin.Update(\hat{y} = y)$$
$$PageHinkley.Update(\hat{y} = y)$$

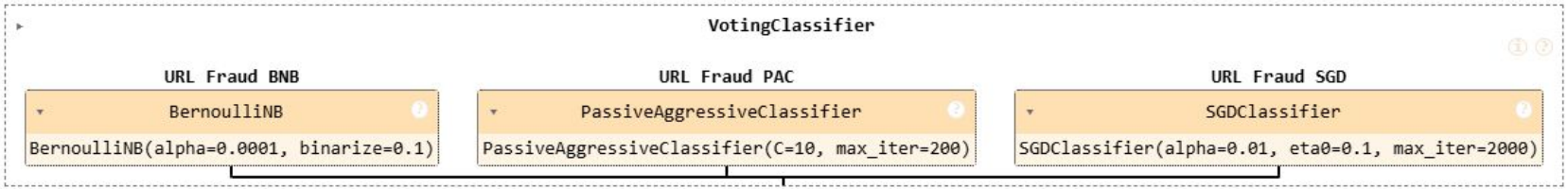
IF $Drift\ Detected\ (Adwin \vee PageHinkley)$ then

$$Voting = Ensemble([PA, BNB, SGD])$$



	Model	Accuracy	Precision	Recall	F1Score	Training Time
0	BernoulliNB	0.994826	0.993370	0.997633	0.995497	0.458212
1	PassiveAggressiveClassifier	0.990776	0.985827	0.998262	0.992005	2.800088
2	SGDClassifier	0.991348	0.990314	0.994637	0.992471	3.701061
3	Voting model	0.999067	0.999186	0.999186	0.999186	2.383109

Clasificador ensemble para toma de decisión



Algoritmo de aprendizaje en línea ante cambio de distribución de features por atacantes.

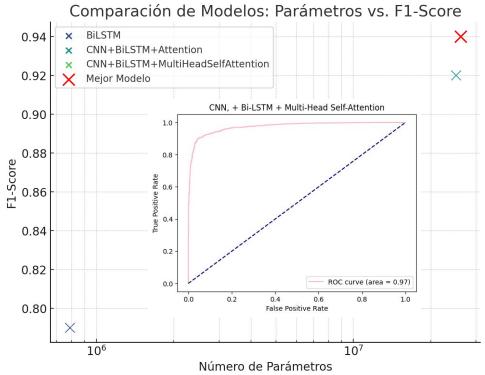
PhawAI

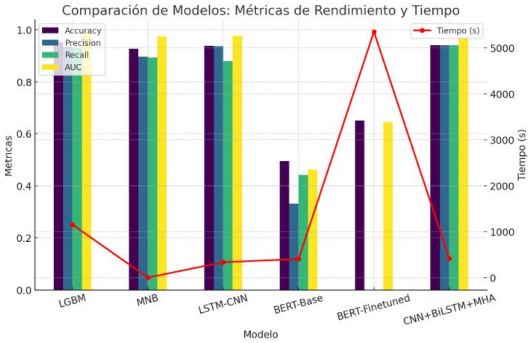
Técnicas de inteligencia artificial

Modelos para clasificar contenido web de phishing

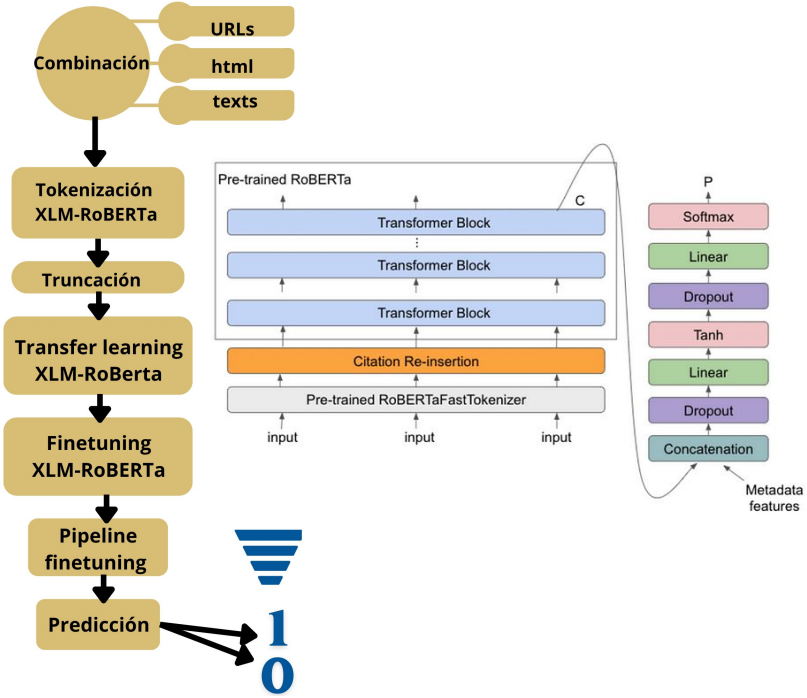
Layer (type)
input_layer (InputLayer)
embedding (Embedding)
conv1d (Conv1D)
max_pooling1d (MaxPooling1D)
bidirectional (Bidirectional)
multi_head_attention (MultiHeadAttention)
flatten (Flatten)
dropout_1 (Dropout)
dense (Dense)
dropout_2 (Dropout)
dense_1 (Dense)

Mejor modelo probado





Modelo para clasificar contenido url, web y textual multilingüe



Se necesita arquitectura robustas y rápidas para detectar un ataque de phishing de distintas modalidades.

Evaluación de propuesta

Desempeño de features de HTML y URLS

Correlación de Xy	Selección	Explicación de features	Concept drift	Selección de modelos con CV
Selección por Feature importance.	Selección basada en información mutua (q85). **I(x;y) = Hx + Hy - Hx,y**	Impacto global y local mediante SHAP y LIME.	¿Cambio de distribución? Re-entrenar o incrementar. **ADWIN = |μWo - μW1| > e (concept drift)** **Page-Hinkley = Σ∂x - min ∂x > e**	

Desempeño de modelos

$$False\ Positive\ Rate = \frac{FP}{TN + FP}$$

Minimizar interrupciones y molestias.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

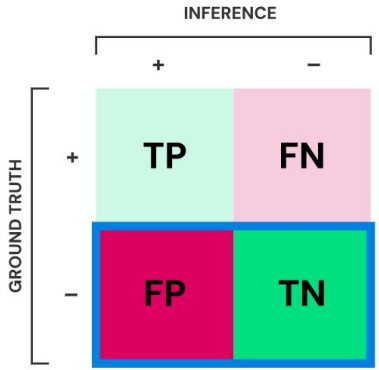
$$Recall = \frac{TP}{TP + FN}$$

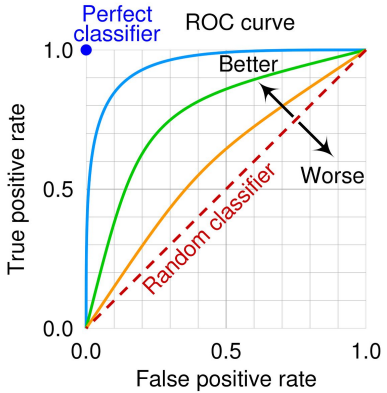
$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Balance de precision-recall

$$AUC - ROC$$

Umbral óptimo para minimizar FP y FN.





Se debe seleccionar las mejores features sin depender del modelo y predecir con una baja tasa de falsos positivos.

Impactos

Detección temprana de ataques de ingeniería social.

Mejora de la confianza del usuario en la seguridad digital y reducir brecha tecnológica en adopción de medidas de seguridad.

Análisis en tiempo real de patrones de comportamiento sospechosos.

Disminución de falsos positivos, mejorando la precisión.

Identificación de nuevas tácticas de fraude a través del aprendizaje automático.

Reducción de costos asociados a la gestión de incidentes de seguridad.

Beneficios

Reducción de pérdidas de datos y económicas.

Aumento de la eficiencia operativa al reducir el tiempo de respuesta.

Capacitación continua y adaptabilidad automática del modelo a través de nuevos datos.

Protección de grupos vulnerables contra ataques de ingeniería social y mejorar en la confianza en la seguridad digital.

Integración con otros sistemas de seguridad para una defensa más robusta.

Disminución de costos operativos en resolución de incidentes.

Oportunidades

- Complemento para la red de comunicación estatal como privada.
- Establecimiento de canales de comunicación efectivos para compartir información sobre amenazas detectadas y mejores prácticas.
- Desarrollo de Protocolos de Respuesta

Bibliografía

Al Tawil, A., Almazaydeh, L., Qawasmeh, D., Qawasmeh, B., Alshinwan, M., & Elleithy, K. (2024). Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT. *Computers, Materials & Continua*, *81*(2), 3395–3412. https://doi.org/10.32604/CMC.2024.057279

CMC | Comparative Analysis of Machine Learning Algorithms for Email Phishing Detection Using TF-IDF, Word2Vec, and BERT. (n.d.). Retrieved February 13, 2025, from https://www.techscience.com/cmc/v81n2/58675

Innab, N., Osman, A. A. F., Ataelfadiel, M. A. M., Abu-Zanona, M., Elzaghmouri, B. M., Zawaideh, F. H., & Alawneh, M. F. (2024). Phishing Attacks Detection Using Ensemble Machine Learning Algorithms. *Computers, Materials & Continua*, *80*(1), 1325–1345. https://doi.org/10.32604/CMC.2024.051778

Mauricio Hernández Armenta. (2022). *ESET: Perú es el país con más ataques de phishing en América Latina*. Pagina Web. https://forbes.pe/tecnologia/2022-10-26/eset-peru-es-el-pais-con-mas-ataques-de-phishing-en-america-latina

Ribeiro, L., Guedes, I. S., & Cardoso, C. S. (2024). Which factors predict susceptibility to phishing? An empirical study. *Computers & Security*, *136*, 103558. https://doi.org/https://doi.org/10.1016/j.cose.2023.103558

Shafin, S. S. (2024). An Explainable Feature Selection Framework for Web Phishing Detection with Machine Learning. *Data Science and Management*. https://doi.org/10.1016/J.DSM.2024.08.004

Ticona, J. C. A., Calcina, K. M. C., Lipe, J. J. L., Valero, M. L., Cabrera, R. M. M., García, H. L. T., & Parí, N. C. (2024). CAUSAS Y CONSECUENCIAS DEL INCREMENTO DE LOS DELITOS INFORMÁTICOS EN LA CIUDAD DE PUNO 2023. *REVISTA DE DERECHO*, *9*(1), 2024. https://doi.org/10.47712/RD.2024.V9I1.262

Wang, M., Zang, X., Cao, J., Zhang, B., & Li, S. (2024). PhishHunter: Detecting camouflaged IDN-based phishing attacks via Siamese neural network. *Computers & Security*, *138*, 103668. https://doi.org/10.1016/J.COSE.2023.103668

Con pérdidas por fraude en todo el país que en 2023 superaron los $10,000 millones de dólares, la FTC intensifica sus esfuerzos para proteger al público | Comisión Federal de Comercio. (n.d.). Retrieved February 15, 2025, from https://www.ftc.gov/es/noticias/con-perdidas-por-fraude-en-todo-el-pais-que-en-2023-superaron-los-10000-millones-de-dolar