

Detector de phishing en correos electrónicos

1st Sandro Carrillo

*Escuela de Ciencia de la Computación
Universidad Nacional de Ingeniería
Lima, Perú
elcorreo@uni.pe*

2nd Ariana Lopez

*Escuela de Ciencia de la Computación
Universidad Nacional de Ingeniería
Lima, Perú
elcorreo@uni.pe*

3rd Albert Argumedo

*Escuela de Ciencia de la Computación
Universidad Nacional de Ingeniería
Lima, Perú
elcorreo@uni.pe*

4th Juan de Dios Lerzundi

*Escuela de Ciencia de la Computación
Universidad Nacional de Ingeniería
Lima, Perú
juan.lerzundi.r@uni.pe*

Resumen—El phishing por correo electrónico es una de las amenazas más frecuentes en ciberseguridad. Este trabajo presenta un sistema de detección de correos de phishing basado en técnicas de aprendizaje supervisado y procesamiento de lenguaje natural. Se desarrolló un pipeline que incluye limpieza de texto, tokenización y vectorización con TF-IDF, y se evaluaron dos clasificadores: Multinomial Naive Bayes y Random Forest. Los resultados (placeholders) muestran que Random Forest obtiene mayor recall y F1-score, indicando una mejor capacidad para identificar correos maliciosos. Finalmente, se discuten las implicancias del preprocesamiento, limitaciones del modelo y líneas de mejora como el uso de embeddings contextualizados.

Palabras clave—phishing, clasificación, TF-IDF, Random Forest, Naive Bayes, aprendizaje supervisado.

I. INTRODUCCIÓN

A. Contexto y Motivación

El correo electrónico es una herramienta de comunicación masiva usada en ámbitos personales, académicos y empresariales. Su popularidad la convierte en un vector atractivo para ataques de ingeniería social y fraude digital. El phishing (y su forma dirigida spear-phishing) busca engañar al usuario para obtener credenciales o información sensible, provocando pérdidas económicas, robo de identidad y compromiso de sistemas (Fette, Sadeh, & Tomasic, 2007).

B. Importancia del Problema

Los filtros tradicionales basados en listas negras o reglas simples no son suficientemente robustos frente a mensajes que imitan lenguaje legítimo o utilizan pequeñas variaciones en dominios. Por ello, la aplicación de técnicas de IA y procesamiento de lenguaje natural (PLN) para analizar contenido y metadatos del correo es una línea prometedora para mejorar la detección (Abu-Nimeh et al., 2007).

C. Objetivos del Proyecto

Desarrollar y evaluar un prototipo de detector de phishing en correos electrónicos utilizando técnicas de Machine Learning y PLN, comparando Naive Bayes y Random Forest sobre un dataset preprocesado, con el objetivo de maximizar la detección (recall) y mantener una F1 alta.

D. Restricciones

- El proyecto se centra en la detección de phishing general (no spear-phishing dirigido).
- Se parte de un dataset ya preprocesado (el informe detalla el proceso de preprocesamiento realizado).

II. MARCO TEÓRICO

A. Definición y naturaleza del phishing

El phishing es una técnica de ingeniería social que emplea correos (u otros canales) simulando entidades legítimas para inducir a la víctima a revelar credenciales, ejecutar acciones o entregar información sensible. Sus variantes incluyen spear-phishing (ataques dirigidos), whaling (dirigido a ejecutivos) y ataques basados en URL/malware (ver Garera et al., 2007; Fette et al., 2007).

B. Representación del texto y fundamentos de PLN

- **Tokenización y normalización:** separación en tokens, conversión a minúsculas, normalización Unicode y tratamiento de caracteres especiales (URLs, correos, números).
- **Stopwords, lematización/stemming:** reducción de forma de palabras para disminuir sparsity.
- **Bag-of-Words / n-grams:** representación básica que captura presencia/frecuencia de términos y frases cortas; útil con Naive Bayes y modelos lineales.
- **TF-IDF:** ponderación por frecuencia inversa a la frecuencia en corpus, reduce peso de términos comunes.
- **Embeddings y representaciones densas:** word2vec/GloVe representan palabras en espacios vectoriales densos; capturan similitud semántica y mejoran cuando se requiere generalización semántica (Mikolov et al., 2013; Pennington et al., 2014).
- **Representaciones contextualizadas:** Transformers (BERT, RoBERTa) generan embeddings contextuales por token/frase y suelen mejorar el rendimiento en tareas de clasificación de texto si hay datos suficientes (Vaswani et al., 2017; Devlin et al., 2019).

C. Modelos de clasificación supervisada

- **Naive Bayes (Multinomial/Bernoulli):** modelo probabilístico asumiendo independencia condicional de características; rápido y efectivo en text mining con bag-of-words/TF-IDF (Manning et al., 2008).
- **Árboles y ensamblados (Random Forest, Gradient Boosting):** capturan interacciones no lineales entre features; Random Forest es robusto a ruido y poca sintonía de hiperparámetros.
- **SVM:** buena separación en espacios de alta dimensión, efectivo con kernels y cuando hay margen claro entre clases.
- **Redes neuronales y Transformers:** CNNs/RNNs (antes dominantes en PLN) y hoy Transformers. Requieren más datos y cómputo, pero capturan mejor semántica y contexto.
- **Evaluación y selección de modelos:** emplear cross-validation, curvas ROC/PR, y reportar precision/recall/F1 por clase (especialmente recall para detectar phishing).

D. Tipos de features relevantes en detección de phishing

- **Contenido textual:** términos sospechosos (“password”, “verify”), patrones de lenguaje urgente, uso de imperativos, errores ortográficos.
- **Características estructurales:** presencia de URLs, número de enlaces, longitud del correo, HTML vs texto plano, uso de imágenes.
- **Características del remitente/metadatos:** dominio del remitente, coincidencia entre From y Return-Path, IP de origen, registros SPF/DKIM/DMARC.
- **Características de la URL:** longitud, uso de subdominios, caracteres especiales, presencia de IP en lugar de dominio, similitud visual con dominios legítimos (homoglyphs).
- **Comportamiento y contexto:** tasa de envíos por remitente, patrones temporales, reputación del dominio.

E. Problemas técnicos centrales

- **Desbalance de clases:** los correos legítimos suelen superar en número a los phishing; requiere técnicas de muestreo (oversampling/SMOTE), o métricas robustas (PR-AUC).
- **Drift y evolución de ataques:** campañas nuevas cambian vocabulario y tácticas; necesario reentrenamiento y monitorización.
- **Adversarialidad:** atacantes pueden adaptar textos para evadir detectores; conviene estudiar adversarial ML y robustez (Biggio & Roli, 2018).

III. ESTADO DEL ARTE

A. Enfoques históricos y baselines

- **Reglas heurísticas y blacklists:** sistemas tempranos usaban listas negras de dominios/URLs y reglas de coincidencia de patrones; funcionan contra técnicas conocidas pero no contra evasiones (nazario corp).

- **Filtrado bayesiano y ML clásico:** Multinomial Naive Bayes y SVMs se convirtieron en baseline por su eficiencia y desempeño en datasets textuales (Abu-Nimeh et al., 2007; Fette et al., 2007).

B. Métodos basados en características (features)

- Investigaciones tempranas mostraron que combinar features de contenido, URL y metadatos mejora la detección. Garera et al. (2007) y Bergholz et al. (2010) analizaron heurísticas de URL y contenido.
- Modelos de ensamblado (Random Forest, Gradient Boosting) demostraron robustez frente a ruido y capacidad para priorizar features relevantes (importancia de variables).

C. Deep learning y representaciones modernas

- **CNN/RNN:** capturan patrones locales y secuencias (uso en clasificación de correo).
- **Transformers (BERT y variantes):** han mostrado mejoras marcadas en clasificación de texto y tareas de seguridad cuando se dispone de datos de calidad o se hace fine-tuning (Devlin et al., 2019).
- **Enfoques híbridos:** combinar embeddings BERT con features manuales (URL, encabezados) suele ser altamente efectivo, aprovecha semántica de texto y reglas estructurales.
-

D. Datasets y benchmarks comunes

- **Enron Email Dataset:** corpus grande de correos empresariales (usado para spam/filtrado y como fuente de ham).
- **SpamAssassin public corpus:** colección etiquetada de spam y ham.
- **Phishing corpora / repositorios:** conjuntos públicos con muestras de phishing (por ejemplo PhishTank/Nazario/PhishCorpus).
- Investigaciones recientes también construyen datasets a partir de correos reales anotados y de campañas de phishing actuales; calibrar modelos en datos recientes es crucial.

E. Evaluación práctica y métricas

- **Recall prioritario:** en detección de phishing, disminuir falsos negativos suele ser más importante que minimizar falsos positivos.
- **PR-AUC vs ROC-AUC:** cuando la clase positiva es rara, PR-AUC refleja mejor la capacidad del clasificador para identificar positivos relevantes.
- **Explicabilidad y análisis de errores:** LIME/SHAP son herramientas para entender decisiones y depurar falsos positivos/negativos.

F. Resumen de hallazgos empíricos

- Los modelos que combinan features manuales (URL, encabezados) con representaciones semánticas (embeddings o Transformers) tienden a obtener los mejores resultados.

- Random Forest y Gradient Boosting son fuertes competidores cuando los recursos son limitados; Transformers dominan cuando hay datos y cómputo suficientes.
- La robustez a nuevas campañas requiere pipelines de reentrenamiento y detección de deriva.

G.

IV. METODOLOGÍA

El sistema consta de los siguientes módulos:

A. Flujo General

- 1) **Preprocesamiento:** normalización y limpieza del texto del correo (remoción de HTML, URLs, tokens no alfábéticos), reemplazo de direcciones y números por tokens especiales, minúsculas, eliminación de stopwords, lematización/stemming.
- 2) **Extracción de características:** representación mediante TF-IDF (n-grams unígrafo y bigrama), inclusión opcional de features booleanos (presencia de URL, cantidad de enlaces, uso de palabras claves sospechosas) y metadatos (dominio remitente, encabezados).
- 3) **Partición de datos y balanceo:** división Train/Test (p. ej. 80/20) y validación cruzada (5-fold). En presencia de desbalance, aplicación de técnicas como SMOTE o submuestreo estratificado.
- 4) **Entrenamiento:** entrenamiento de Multinomial Naive Bayes y RandomForestClassifier (scikit-learn).
- 5) **Evaluación:** cálculo de accuracy, precision, recall, F1, ROC-AUC y PR-AUC; generación de matriz de confusión y análisis de importancia de variables (Random Forest).
- 6) **Análisis y despliegue:** análisis de casos erróneos, propuestas de mejora y diseño de prototipo para integración en pipeline de correo.

B. Preprocesamiento

- 1) **Limpieza:** eliminar etiquetas HTML, normalizar saltos de línea, eliminar caracteres no ASCII si procede.
- 2) **Normalización:** pasar a minúsculas, normalizar acentos.
- 3) **Tokens especiales:** reemplazar URLs por <URL>, direcciones de email por <EMAIL> y números por <NUM>.
- 4) **Tokenización:** separación en tokens — se consideraron n-grams (1,2).
- 5) **Stopwords & Lematización:** remover palabras funcionales y lematizar para reducir variabilidad morfológica.
- 6) **Vectorización:** TF-IDF con límite de vocabulario (p. ej. max_features=20,000), ngram_range=(1,2), y sublinear_tf=True.

C. Modelos

- **Multinomial Naive Bayes:** `sklearn.naive_bayes.MultinomialNB(alpha=1.0)` como baseline para datos textuales.
- **Random Forest:** `sklearn.ensemble.RandomForestClassifier(n_estimators=200, max_depth=None, n_jobs=-1,`

`random_state=42)` usado para capturar interacciones entre tokens y es robusto a features ruidosos.

- **Validación:** 5-fold cross-validation y evaluación sobre test holdout (80/20). Se reportan métricas promedio y desviación estándar.

D. Métricas de evaluación

- **Precision, Recall, F1-score:** para la clase positiva (phishing).
- **ROC-AUC y PR-AUC:** dado que la clase positiva puede ser minoritaria, PR-AUC es útil para valorar el rendimiento en detección.
- **Matriz de confusión:** análisis de falsos positivos y falsos negativos (importante por coste distinto de ambos errores).

V. RESULTADOS

A. Tabla comparativa de métricas

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

B. Matrices de confusión

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

C. Figuras

VI. DISCUSIÓN

A. Consideraciones Éticas

- El análisis de correos electrónicos implica acceso a información potencialmente sensible (mensajes privados, datos personales). Es obligatorio aplicar principios de minimización de datos, anonimización/pseudonimización y políticas de retención.

- La sobre-representación de ciertos idiomas puede traducirse en mayor tasa de falsos positivos para grupos específicos, siendo idiomas minoritarios, esto añade un sesgo en los datos de entrenamiento.
- Para tratar el sesgo en los datos se recomendaría evaluar métricas por subgrupos, usar muestreo estratificado y técnicas de debiasing.
- La aparición de falsos positivos puede bloquear comunicaciones legítimas y afectar operaciones como pérdida de información demora. Ante esto se deberían aplicar políticas de cuarentena, notificaciones o vías de apelación, umbrales ajustables según el criterio del usuario u organización.

VII. CONCLUSIONES

A. Conclusiones Principales

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdier mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

B. Trabajo a Futuro

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdier mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

REFERENCES

- [1] Abu-Nimeh, S., Nappa, D., Wang, X., & Nair, S. (2007). A comparison of machine learning techniques for phishing detection. *eCrime Researchers Summit*.
- [2] Fette, I., Sadeh, N., & Tomasic, A. (2007). Learning to detect phishing emails. *Proceedings of the 16th International World Wide Web Conference (WWW)*.
- [3] Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- [4] Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- [5] Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.
- [6] Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall.
- [7] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
- [8] Bergholz, A., De Beer, J., Glahn, S., Moens, M.-F., Paaf, G., & Strobel, S. (2010). New filtering approaches for phishing email. *Journal of Computer Security*, 18(1), 7–35.
- [9] Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84, 317–331.
- [10] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*.
- [11] Garera, S., Provos, N., Chew, M., & Rubin, A. D. (2007). A framework for detection and measurement of phishing attacks. *Proceedings of the 2007 ACM Workshop on Recurring Malcode*.
- [12] Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR)*.
- [13] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems (NeurIPS)*.
- [14] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *ICLR Workshop*.
- [15] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. *EMNLP*.
- [16] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- [17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*.
- [18] European Parliament and Council. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). *Official Journal of the European Union*.

ANEXOS

- **Anexo A:** Código (notebook Colab) con pipeline completo (preprocesamiento, vectorización, entrenamiento y evaluación).
- **Anexo B:** Parámetros exactos de los modelos y resultados de las 5 corridas cross-validation (media y desviación estándar).
- **Anexo C:** Plots (Figura 1, Figura 2, Figura 3) y matrices de confusión.
- **Anexo D:** Tabla detallada de distribución de trabajo y cronograma (según requisitos del curso).