Detector de phishing en correos electrónicos

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Juan de Dios Lerzundi
Escuela de Ciencia de la Computación
Universidad Nacional de Ingeniería
Lima, Perú
juan.lerzundi.r@uni.pe

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Resumen*—El phishing por correo electrónico sigue siendo una de las principales amenazas a la seguridad informática. En este trabajo se presenta un sistema de detección de correos electrónicos de phishing basado en técnicas de aprendizaje supervisado y procesamiento de lenguaje natural. Partiendo de un dataset preprocesado, se compararon dos clasificadores clásicos: Multinomial Naive Bayes y Random Forest. El preprocesamiento incluyó limpieza, normalización, tokenización, eliminación de stopwords, lematización y vectorización mediante TF-IDF. La evaluación se realizó mediante validación cruzada y mediciones de precisión, recall, F1-score, ROC-AUC y PR-AUC. Los resultados (placeholders) muestran que Random Forest supera a Naive Bayes en F1 y recall, sugiriendo que los modelos de conjunto aprovechan mejores interacciones entre características textuales. Se discuten implicancias, limitaciones y direcciones futuras, incluyendo el uso de embeddings contextualizados (p. ej. BERT) y la incorporación de metadatos del correo (remitente, dominio, encabezados).**

***Palabras clave*—phishing, detección de spam, procesamiento de lenguaje natural, Random Forest, Naive Bayes, TF-IDF, clasificación supervisada**

I. Introducción

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).

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).

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.

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. 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 alfabéticos), reemplazo de direcciones y números por tokens especiales, minúsculas, eliminación de stopwords, lematización/stemming.
2) **Extracción de características:** representación mediante TF-IDF (n-grams unigrama 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.
7)
8)

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).

III. Resultados

A. Tabla comparativa de métricas

Lorem ipsum dolor sit amet, consectetuer 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, consectetuer 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, consectetuer 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, consectetuer 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

IV. Discusión

A. Comparación de Modelos

- Lorem ipsum dolor sit amet, consectetuer 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, consectetuer 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. Preprocesamiento

Lorem ipsum dolor sit amet, consectetuer 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, consectetuer 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. Limitaciones

Lorem ipsum dolor sit amet, consectetuer 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, consectetuer 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.

V. Conclusiones

A. Conclusiones Principales

Lorem ipsum dolor sit amet, consectetuer 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, consectetuer 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, consectetuer 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, consectetuer 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.

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).