CEA Comité Español de Automática

XLV JORNADAS AUTOMÁTICA MÁLAGA 2024

UNIVERSIDADE DA CORUÑA

Jornadas de Automática

Mejoras en extracción de URLs en smishing mediante text spotting

Blanco-Medina, P.[a,b,*], Biswas, R.[a,b], González-Castro, V.[a,b], Alaiz Rodríguez, R.[a,b], Fidalgo, E.[a,b], Alegre, E.[a,b]

[a] *Departamento de Ingeniería Eléctrica y de Sistemas y Automática, Universidad de León, Campus de Vegazana, 24007, Leon, España*
[b] *Investigador Colaborador en INCIBE*

To cite this article: Blanco-Medina, P., Biswas, R., González-Castro, V., Alaiz Rodríguez, R., Fidalgo, E., Alegre, E. 2024. Enhancing Smishing URL Extraction with Text Spotting.
Jornadas de Automática, 45. https://doi.org/10.17979/ja-cea.2024.45.10954

Resumen

Los Equipos de Respuesta ante Emergencias Informáticas (CERT) reciben comúnmente capturas de pantalla de *Smishing*, que tratan de suplantar a distintos tipos de organizaciones, con el objetivo de apropiarse de información personal de usuarios o malversar fondos de sus cuentas mediante enlaces maliciosos. Los CERTs buscan soluciones automatizadas que permitan recuperar URLs de capturas de pantalla. Para extraer texto pueden utilizarse métodos basados en el reconocimiento óptico de caracteres (OCR), pero su rendimiento es bajo debido a problemas como la baja calidad de la imagen o textos divididos en múltiples frases. Proponemos un proceso para la extracción de URL de *Smishing* basado en técnicas de Text Spotting, complementado con una reconstrucción de URL personalizada utilizando características resaltadas en la imagen. Aplicamos la metodología propuesta a un conjunto personalizado de 244 capturas y 262 URLs, obteniendo como resultado un aumento de la precisión de reconocimiento de 3, 05 % a 22, 90 %, tras lo cual puede continuarse procesando el texto extraído en *Smishing*.

Palabras clave:
Seguridad, Aprendizaje Profundo, Apoyo a Operadores Humanos, Redes Sociales, Automatización para la Ayuda Internacional

Enhancing Smishing URL Extraction with Text Spotting.

Abstract

Computer Emergency Response Teams (CERTs) often get screenshots showcasing brief texts with doubtful content. *Smishing* attempts to mimic reputable organizations, urging individuals to act promptly by clicking on a link, aiming to hijack personal information or illicitly debit funds from their accounts. CERTs may find value in automated solutions that can retrieve URLs from screenshots for subsequent validation. Approaches based on Optical Character Recognizers (OCRs) could be used to extract text. However, their performance is low due to the poor performance of OCR in certain images. In this work, we propose a pipeline for *Smishing* URL extraction based on Text Spotting, which will later be applied to a custom URL reconstruction using highlighted features. We applied the proposed pipeline to a custom set of 117 screenshots containing 121 URLs, resulting in a precision increase on the URL recovery task from 3,05 % to 22,90 %. This allows the original URL to be restored for subsequent processing in the analysis of *Smishing* messages.

Keywords:
Security, Deep Learning, Human operator support, Social networking, Control and Automation Systems for International Aid

1. Introducción

En los tiempos actuales, los servicios de mensajería instantánea (SMS) han adquirido un rol secundario respecto a otras plataformas como Whatsapp, Telegram o Line, debido a sus capacidades de personalización y facilidad de uso. En su lugar, los SMS pueden utilizarse para recibir comunicaciones oficiales o notificaciones de servicios personalizados por

*Autor para correspondencia: pblanm@unileon.es

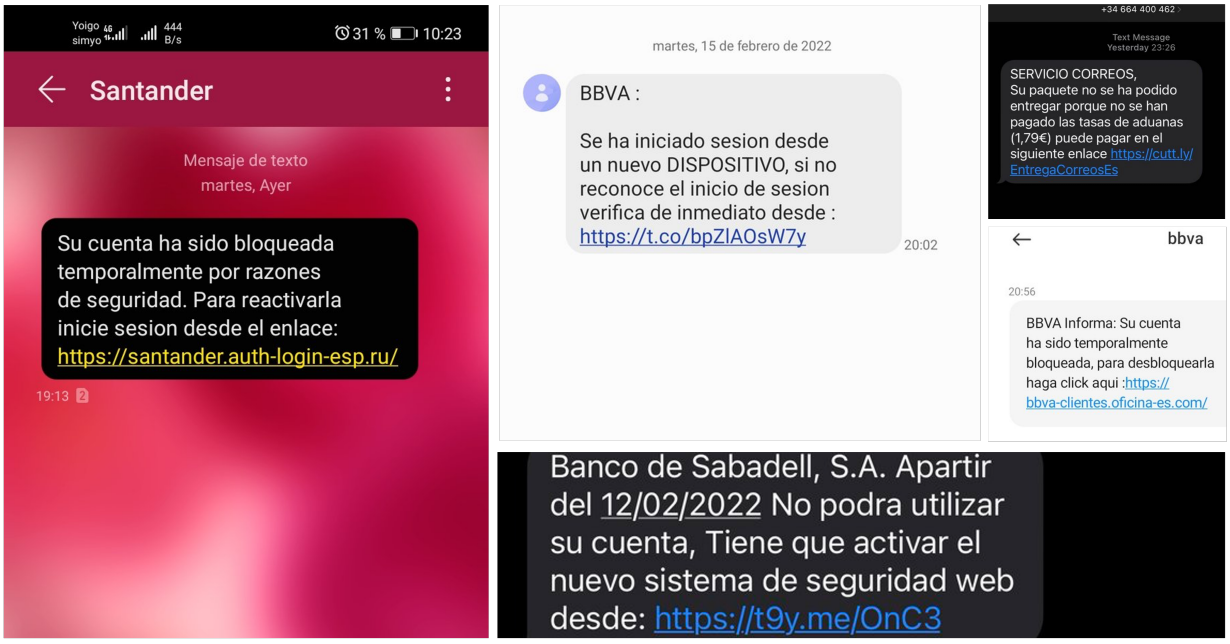

Figura 1: Ejemplos de capturas de pantalla de smishing. Los usuarios pueden confundir mensajes legítimos con campañas maliciosas. Además, los múltiples fondos, colores, fonts y disposiciones de texto dificultan la recuperación automática de la información en la captura.

organismos oficiales (Church and De Oliveira, 2013).

Uno de los riesgos más comunes de los SMS es la suplantación de entidades oficiales y envío de enlaces maliciosos, conocidos como *Smishing* (Al-Qahtani and Cresci, 2022). Este tipo de estafas buscan que los usuarios accedan, a través de los enlaces incluidos en los SMS, a plataformas que imitan organismos oficiales u otras empresas con las que se relacionan con el objetivo de robar sus datos (Sánchez-Paniagua et al., 2022; Jáñez-Martino et al., 2023).

A pesar del desarrollo de diversas técnicas y filtros automáticos para reducir el impacto de este tipo de mensajes, los usuarios no expertos pueden ser coaccionados con mensajes de urgencia a interactuar con este tipo de estafas (Rahman et al., 2023). La figura 1 ilustra varios ejemplos de campañas de *Smishing* con objetivos maliciosos, alentando al usuario a acceder a un enlace mediante el cual se suplanta a un banco o un servicio de correos con el objetivo final de robar sus credenciales.

Para distinguir mensajes legítimos de mensajes de *Smishing*, los métodos más actuales emplean técnicas de procesamiento de lenguaje natural, que analizan tanto el contenido del mensaje como la estructura de la URL (Timko and Rahman, 2023). Sin embargo, estos métodos no permiten analizar capturas de pantalla, una de las formas más comunes de compartir estos mensajes sin comprometer la seguridad de otros usuarios.

Las capturas de pantalla de estos mensajes, que típicamente formar pante de campañas de *Smishing*, pueden utilizarse con fines de preservación, así como de sistemas de registro (logging) para reconstruir ataques pasados y prever futuras campañas de *Smishing* (Vadrevu et al., 2017). En el caso de los Equipos de Respuesta ante Emergencias Informáticas (CERTs en Inglés), al no disponer del texto plano contenido en la captura del mensaje, este tipo de campañas deben ser analizadas mediante la extracción manual del texto y las URLs contenidas en la imagen. Además, debido a las múltiples configuraciones de texto en los dispositivos móviles, el texto puede presentarse en fuentes y tamaños distintos, con distribuciones variables, o incluso partido en múltiples líneas, lo que dificulta la lectura automática del contenido (Maneriker et al., 2021).

Para solucionar estos problemas, en este artículo proponemos una solución basada en técnicas de Text Spotting, consistentes en la localización y reconocimiento de texto presentes en una imagen (Blanco-Medina et al., 2022), con el fin de recuperar URLs contenidas en capturas de pantalla de mensajería instantánea. Nuestra propuesta permite mejorar la calidad del texto recuperado y las URLs completas, facilitando las tareas posteriores de procesado del lenguaje.

Verificamos nuestra propuesta en un conjunto de datos de *smishing*, formado por 244 capturas de pantalla que continene 262 URLs, analizando la aplicación de técnicas de detección y reconocimiento de texto, estudiando las diferencias entre las URLs y analizando la precisión que se obtendría con los métodos de Text Spotting sin aplicar las mejoras propuestas.

Proponemos también una metodología para unir y reconstruir las componentes de las URLs, que permite recomponer URLs separadas en dos o más múltiples líneas, a través de la unión de componentes destacados, los cuales incluyen texto subrayado o en un color distinto del resto del mensaje. Gracias a esta propuesta, conseguimos aumentar la precisión total del reconocimiento de URLs en mensajes *Smishing*, a partir de la cual pueden realizarse operaciones de post-procesado y análisis de la información (Joshi et al., 2023)

El resto del artículo se organiza de la siguiente forma. En la sección 2 revisamos brevemente trabajos similares aplicados a la detección y reconocimiento de textos en el campo de *Smishing*. En la sección 3 detallamos la metodología de nuestra experimentación, presentando los resultados obtenidos en la sección 4. Finalmente, la sección 5 concluye el artículo detallando nuestras conclusiones y futuras líneas de trabajo.

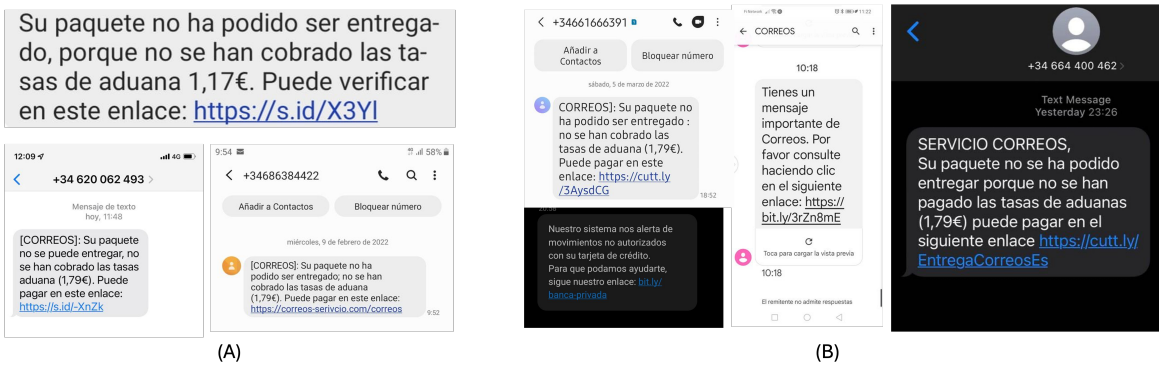

Figura 2: Capturas de Smishing presentes en nuestro conjunto de datos. (A) URLs completas (URL-C) y (B) divididas (URL-D). Las URLs divididas tienen que ser reconstruidas antes de ser comparadas con el etiquetado del conjunto de datos.

2. Trabajos Relacionados

Diversos estudios sobre *Smishing* referencian las múltiples propuestas de prevención de este tipo de estafas, que se centran en actuar sobre los servicios de envíos masivos, los proveedores de servicio SMS, y filtros o aplicaciones antismishing disponibles (Timko and Rahman, 2023). La mayoría de los métodos propuestos se centran en el análisis del contenido del texto, aunque hay otros que analizan los componentes de red que ofrezcan detalles sobre la procedencia del mensaje (Ulfath et al., 2022).

Las propuestas de solución más recientes utilizan técnicas de Aprendizaje Profundo para analizar los componentes del mensaje de texto, URLs, o nombres destacados del mensaje original para categorizarlos como posibles campañas de *Smishing* (Ulfath et al., 2022; Blanco-Medina et al., 2022). Estos trabajos citan la necesidad constante de conjuntos de datos actualizados para poder prevenir futuras campañas de *Smishing*, dada la rápida pérdida de relevancia y representatividad de los datasets previos (Mishra and Soni, 2022), lo que suscita la necesidad de herramientas y conjuntos de datos novedosos como es el caso de Smishtank [1].

Los trabajos revisados no incluyen análisis sobre la eficacia de métodos de detección y reconocimiento de texto en imágenes *Smishing*. En los últimos años, estos métodos han aumentado su rendimiento mediante el uso de Transformers (Bautista and Atienza, 2022), modelos de redes neuronales que utilizan mecanismos de atención. Además, gracias a la optimización de sus técnicas de entrenamiento, no requieren una gran cantidad de imágenes para reentrenarlos de extremo a extremo para una tarea (Baek et al., 2021), lo que los convierte en un método ideal para recuperar texto en el contexto reducido de las imágenes *Smishing*.

3. Metodología

Para realizar la experimentación utilizamos un conjunto de 244 imágenes de *Smishing*, recopiladas de manera automatizada de plataformas como Twitter y FaceBook, y otras proporcionadas por el Instituto Nacional de Ciberseguridad (INCIBE). Para cada imagen, etiquetamos tanto la localización espacial de la URL como la transcripción del texto contenido en la misma. Este proceso se realizó de manera semiautomatizada, combinando el uso de un detector y reconocedor de caracteres seguido de una comprobación y corrección realizadas por un etiquetador humano.

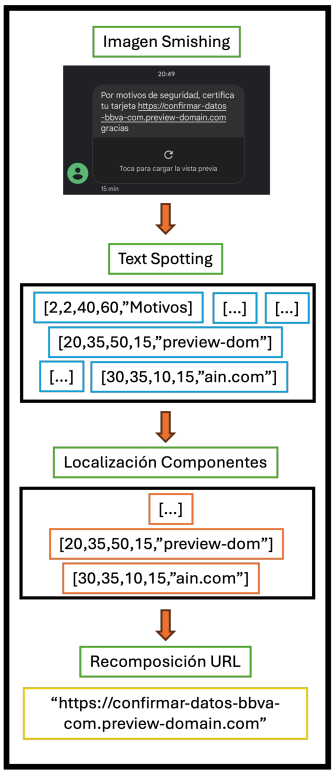

Figura 3: Resumen gráfico del proceso seguido para reconstruir las URLs según los componentes recuperados utilizando Text Spotting. Tras aplicar Text Spotting, obtenemos los puntos de localización espacial de las regiones de texto y sus transcripciones. Comprobamos aquellas regiones que tienen componentes con texto resaltado y las unimos en una sola región en base a su posición relativa vertical, formando la URL final y comparándola con la URL documentada para evaluar el método propuesto.

Al analizar las imágenes, comprobamos que estas direcciones pueden agruparse en dos tipos principales: las que se presentan en una única línea y las que se sitúan en múltiples

[1] https://smishtank.com/

líneas, disponiendo de 121 y 141 ejemplos respectivamente de cada tipo en nuestro dataset. La figura 2 ilustra estos dos tipos de imágenes presentes en nuestro conjunto de datos.

Observamos que las URLs de las imágenes recopiladas siempre se encuentran destacadas de dos maneras; o bien con un color distinto al del resto del texto (azul o lila), o mediante el subrayado de sus componentes. A pesar de que estos formatos de resalte de texto también pueden darse en textos que no pertenecen a URLs, consideramos útil su reconocimiento al ser posibles componentes de aquellas URLs que se dividen en múltiples líneas, siendo necesario en esos casos unir estos componentes para reconstruir la URL completa. La tabla 1 resume la división de nuestras imágenes en las categorías mencionadas, las cuales serán accesibles públicamente en el futuro a través nuestro sitio web [2].

Tabla 1: División de Imágenes y URLs en el conjunto de datos utilizado para nuestra experimentación. Una misma captura puede contener varias URLs documentadas.

Datos	Texto Completo	Texto Dividido	Total
Capturas	117	127	244
URLs	121	141	262

Para recuperar el texto presente en imágenes de *Smishing* seleccionamos Pixel Aggregation Network (PAN++) (Wang et al., 2021), un método del estado del arte basado en redes neuronales profunas que obtiene buenos resultados en conjuntos de datos que combinan textos de escena con textos escritos a máquina, siendo también aplicable a imágenes cuyo texto incluye fuentes personalizadas (Bautista and Atienza, 2022). Frente a los sistemas de reconocimiento óptico de caracteres (OCR), estos modelos resultan más eficaces en entornos diversos, lo cual puede beneficiarse en el contexto de las URLs debido a múltiples fonts y fondos que ofuscan y dificultan la recuperación del texto, en una distribución más desordenada que el texto plano de los documentos.

Utilizamos este método con sus pesos ya pre-entrenados en datasets del estado del arte de detección y reconocimiento de texto (Karatzas et al., 2015)

Una vez recuperado el texto, comprobamos si las regiones detectadas corresponden con las URLs etiquetadas, reportando la precisión final del método. Observamos que no sólo las URLs divididas en múltiples líneas pueden aparecer separadas en distintos componentes. Algunas de las URLs categorizadas como "completas" pueden ser localizadas en dos o varias regiones según el método de Text Spotting utilizado, siendo necesario reconstruir también este tipo de URLs.

Para ello, proponemos categorizar como regiones candidatas de URLs aquellas regiones que se encuentran subrayadas o tienen un color distinto al resto del texto que aparece en la imagen que está siendo procesada. Mediante operaciones de umbralización y otros procesamientos en escala de grises, recuperamos aquellas regiones detectadas por PAN++ y las unimos, considerando que las regiones seguirán un orden vertical, de arriba a abajo y en el sentido de lectura estándar, obteniendo la URL reconstruida, reportando la URL final. La figura 3 ilustra gráficamente la metodología seguida.

[2] https://gvis.unileon.es/datasets_main/

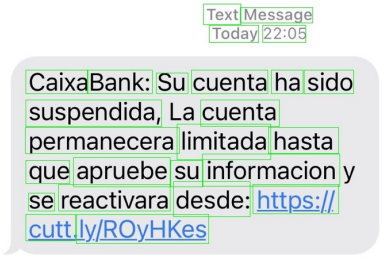

Figura 4: Resultado visual de la aplicación del método PAN++ a un mensaje de Smishing. Algunas regiones pueden omitirse o localizarse de manera separada, lo cual repercute en la precisión final, siendo necesaria una lógica de unión de regiones.

4. Experimentación

La experimentación ha sido realizada en un contenedor con 128 GB de RAM, dos procesadores Intel Xeon E5-2630v3 de 2,4GHz, y dos GPUs NVIDIA Titan X.

En la tabla 2 presentamos los resultados obtenidos al aplicar el método pre-entrenado PAN++, seguido de la ejecución con la lógica de reconstrucción, reportando tanto la precisión global en nuestro conjunto de datos como el número de URLs reconstruidas que son idénticas a las documentadas.

Tabla 2: Precisión en el reconocimiento de URLs en nuestro conjunto de datos. URL-C y URL-D, donde *C* y *D* indican Completa y Dividida respectivamente. El método propuesto utiliza PAN++ y el pre y post procesado de imagen explicado.

Ejecución	URL-C	URL-D	Total
Usando sólo PAN++	6,61 %	0,00 %	3,05 %
Con método propuesto	25,62 %	20,57 %	22,90 %

Al utilizar PAN++ de forma directa, recuperamos correctamente un 6,61 % de las URLs que aparecen en una única línea. Esta baja precisión se debe a que estas URLs de una única línea, cuando son detectadas por el algoritmo de texto, pueden verse "separadas" en múltiples líneas o incluidas en otros bloques de texto mayores, que incluyen texto normal, lo que disminuye la precisión obtenida.

Al incorporar el preprocesado y la lógica de unión, la precisión en esta categoría aumentó al 25,62 %, pudiendo unir los componentes que habían quedado separados por errores en la detección. Además, aumentamos la precisión de la segunda categoría (URLs separadas en varias líneas) a un 20,57 %, desde un 0 % inicial, lo que permitió reconstruir las URLs separadas inicialmente por la disposición del mensaje.

Analizando empíricamente los resultados obtenidos, se observa que los errores más comunes se producen en la fase de reconocimiento del método. Se pueden confundir caracteres similares y generar errores que reducen la precisión global, así como obtener reconocimientos erróneos en áreas que no son caracteres, los cuales pueden incluirse también en la fase de unión de URLs.

Otros errores que hemos observado son la omisión de ciertos caracteres en las cadenas finales o la inclusión de regiones similares a las de URLs, como pueden ser fechas o palabras

subrayadas intencionalmente, los cuales pueden verse corregidos mediante la aplicación de léxicos específicos del contexto de *Smishing* o a través de técnicas de postprocesado de texto. Finalmente, localizamos algunos problemas de superposición entre regiones candidatas, los cuales hacen que cierta información de los componentes de la URL se repita y se generen resultados erróneos.

En resumen, la aplicación conjunta de algoritmos de Text Spotting y nuestra propuesta de unión de componentes de URLs aumentan la precisión de un 3,05 % a un 22,90 % en el contexto de la recuperación de URLs en capturas de pantalla de mensajes *Smishing*. La figura 5 presenta algunos de los errores observados en el uso del método.

Figura 5: Errores de localización de URLs en el método utilizado. La separación de URLs se une a la omisión de ciertos componentes, además de la confusión entre caracteres similares, que disminuye la precisión final del modelo.

5. Conclusiones

En este trabajo analizamos el problema de la recuperación de texto asociado a capturas de pantalla de mensajes *Smishing*. Buscamos extraer la URL contenida en el mensaje mediante técnicas de detección y reconocimiento de texto, detectando los problemas en su recuperación por las distintas disposiciones de texto presentes en dispostivos móviles, para ayudar a Equipos de Respuesta ante Emergencias Informáticas (CERTs) a recuperar la información de estas imágenes y prevenir posibles campañas de smishing.

Para ello hemos utilizado un método de Text Spotting, PAN++, que permite detectar y reconocer regiones de texto localizadas en una imagen. Además, proponemos el uso de una lógica de unión de componentes, destacando aquellos con un color distinto del resto del mensaje o que contengan subrayado. Tras unir todas las regiones destacadas, reconstruimos la URL original, la cual puede ser post-procesada con otros métodos de análisis de texto.

Evaluamos nuestra propuesta utilizando un conjunto de datos de 244 capturas de pantalla que contenían 262 URLs separadas en las categorías completas (cuando la URL está presente en una única línea) y divididas (la URL se distribuye en múltiples líneas de texto), siguiendo un proceso de etiquetado semi-automatizado y anotando tanto la transcripción de las URLs contenidas como su localización espacial en la imagen.

Nuestra propuesta obtiene una precisión de 22,90 % en el reconocimiento de URLs en nuestro conjunto de datos, mejorando la precisión del método original de 3,05 % . Detallamos los errores más comunes en el reconocimiento de texto que se deben a caracteres similares y división de frases en múltiples niveles, tanto por el contenido del mensaje como por la localización de texto del método, lo cual perjudica la precisión global del sistema.

En trabajos futuros, trataremos de combinar ambas mejoras en una lógica más avanzada que considere múltiples características de los componentes de las URLs. La experimentación con técnicas adicionales de pre y postprocesado también son líneas interesantes de investigación, así como el uso de un comprobador automático de vigencia de la URL con el objetivo de avisar al usuario de la posible estafa en tiempo real.

Agradecimientos

Este trabajo ha sido realizado gracias al Plan de Recuperación, Transformación y Resiliencia, financiado por la Unión Europea (Next Generation) gracias al Proyecto LUCIA (Lucha contra el Cibercrimen utilizando Inteligencia Artificial) concedido por INCIBE a la Universidad de León.

Referencias

Al-Qahtani, A. F., Cresci, S., 2022. The covid-19 scamdemic: A survey of phishing attacks and their countermeasures during covid-19. IET Information Security 16 (5), 324–345.

Baek, J., Matsui, Y., Aizawa, K., 2021. What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3113–3122.

Bautista, D., Atienza, R., 2022. Scene text recognition with permuted autoregressive sequence models. In: European conference on computer vision. Springer, pp. 178–196.

Blanco-Medina, P., Fidalgo, E., Alegre, E., González-Castro, V., 2022. A survey on methods, datasets and implementations for scene text spotting. IET Image Processing 16 (13), 3426–3445.

Church, K., De Oliveira, R., 2013. What's up with whatsapp? comparing mobile instant messaging behaviors with traditional sms. In: Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services. pp. 352–361.

Jáñez-Martino, F., Alaiz-Rodríguez, R., González-Castro, V., Fidalgo, E., Alegre, E., 2023. A review of spam email detection: analysis of spammer strategies and the dataset shift problem. Artificial Intelligence Review 56 (2), 1145–1173.

Joshi, A., Fidalgo, E., Alegre, E., Fernández-Robles, L., 2023. Deepsumm: Exploiting topic models and sequence to sequence networks for extractive text summarization. Expert Systems with Applications 211, 118442.

Karatzas, D., Gomez-Bigorda, L., Nicolaou, A., Ghosh, S., Bagdanov, A., Iwamura, M., Matas, J., Neumann, L., Chandrasekhar, V. R., Lu, S., et al., 2015. Icdar 2015 competition on robust reading. In: 2015 13th international conference on document analysis and recognition (ICDAR). IEEE, pp. 1156–1160.

Maneriker, P., Stokes, J. W., Lazo, E. G., Carutasu, D., Tajaddodianfar, F., Gururajan, A., 2021. Urltran: Improving phishing url detection using transformers. In: MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM). IEEE, pp. 197–204.

Mishra, S., Soni, D., 2022. Sms phishing dataset for machine learning and pattern recognition. In: International Conference on Soft Computing and Pattern Recognition. Springer, pp. 597–604.

Rahman, M. L., Timko, D., Wali, H., Neupane, A., 2023. Users really do respond to smishing. In: Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy. pp. 49–60.

Sánchez-Paniagua, M., Fernández, E. F., Alegre, E., Al-Nabki, W., Gonzalez-Castro, V., 2022. Phishing url detection: A real-case scenario through login urls. IEEE Access 10, 42949–42960.

Timko, D., Rahman, M. L., 2023. Commercial anti-smishing tools and their comparative effectiveness against modern threats. In: Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks. pp. 1–12.

Ulfath, R. E., Sarker, I. H., Chowdhury, M. J. M., Hammoudeh, M., 2022. Detecting smishing attacks using feature extraction and classification techniques. In: Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021. Springer, pp. 677–689.

Vadrevu, P., Liu, J., Li, B., Rahbarinia, B., Lee, K. H., Perdisci, R., 2017. Enabling reconstruction of attacks on users via efficient browsing snapshots. In: NDSS.

Wang, W., Xie, E., Li, X., Liu, X., Liang, D., Yang, Z., Lu, T., Shen, C., 2021. Pan++: Towards efficient and accurate end-to-end spotting of arbitrarily-shaped text. IEEE Transactions on Pattern Analysis and Machine Intelligence 44 (9), 5349–5367.