Data in Brief 67 (2026) 112890

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Spanish phishing and legitimate email dataset with technical and psychological annotations

Lázaro Bustio-Martínez[a,*], Viviana Inés Fuentes-Fuentes[b], Luisa Fernanda Agudelo Fuentes[a], Vitali Herrera-Semenets[c], Darián Llanes-Guilarte[d], Felipe Antonio Trujillo-Fernández[a], Antonio Carlos Cardeña-Matamoros[a], Carlos Francisco Betancourt-Moreno[a], Andrés Guillermo Molano-Jiménez[a], Jan van den Berg[e]

[a] *Department of Engineering, Universidad Iberoamericana Mexico City (IBERO), Prolongación Paseo de la Reforma 880, Colonia Lomas de Santa Fe C.P. 01219, Álvaro Obregón, Mexico City, Mexico*
[b] *Independent researcher, Estado de Mexico, C.P. 52005, Mexico*
[c] *On the Dime S.r.l., 555 Via Casine di Paterno, 122/a 60131 Ancona, Italy*
[d] *Advanced Technologies Application Center (CENATAV), 7ma A #21406 e/ 214 y 216, Reparto Siboney, Playa C.P. 12200 La Habana, Cuba*
[e] *Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS), Delft University of Technology, Mekelweg 4 2628 CD Delft, The Netherlands*

ARTICLE INFO

Article history:
Received 3 March 2026
Revised 12 May 2026
Accepted 22 May 2026
Available online 4 June 2026

Dataset link: SpaPhish: A Spanish Dataset for Phishing and Psychological Pattern Detection (Original data)

Keywords:
Phishing emails
Spanish language
Email dataset
Technical metadata
Persuasion annotations
Human annotation

ABSTRACT

The SpaPhish dataset is a curated corpus of 1395 anonymized Spanish-language emails collected from the personal and institutional inboxes of the dataset authors. The collection comprises 731 phishing messages and 664 legitimate communications, spanning 2014–2025. Each record integrates raw textual content (subject and body), derived technical metadata, and psychological annotations.

Technical variables include extracted URLs (url_count, urls), routing depth (hops_count), and attachment metadata (attachments_count, types, and size-related fields). All personally identifying elements were anonymized through manual redaction and controlled substitution to preserve readability while preventing re-identification.

A central component of SpaPhish is the persuasion-annotation layer aligned with Ana Ferreira's Principles

* Corresponding author.
E-mail address: lazaro.bustio@ibero.mx (L. Bustio-Martínez).

https://doi.org/10.1016/j.dib.2026.112890

of Persuasion framework. Three independent annotators performed a triple-blind protocol, assigning binary presence labels (0/1) for five dimensions (Authority, Social Proof, Liking/Similarity/Deception, Commitment/Integrity/Reciprocation, and Distraction), accompanied by brief Spanish justifications and consolidated consensus labels.
SpaPhish supports Spanish-language phishing research, hybrid text–metadata modeling, and annotation reliability studies, and enables explainable analyses grounded in human-provided evidence. The dataset is publicly available in Mendeley Data as "SpaPhish: A Spanish Dataset for Phishing and Psychological Pattern Detection" (version 5, doi: 10.17632/hz2d6gz7pc.5; https://data.mendeley.com/datasets/hz2d6gz7pc/5).

Specifications Table

Subject	Computer Sciences
Specific subject area	Phishing and social engineering in Spanish-language email communications with technical and persuasion-based annotations
Type of data	Table (CSV), Raw, Analyzed, Filtered, Processed.
Data collection	Real Spanish-language emails (phishing and legitimate) were collected from multiple inboxes managed by the dataset authors (2014–2025). Duplicate messages were identified and removed using SHA-256 hashes computed over the raw email message prior to parsing and anonymization. Messages were retained if they contained at least a subject line or a message body; non-Spanish messages were excluded. Phishing and legitimate labels were assigned using a triple-annotator protocol. Technical variables were derived using deterministic rule-based procedures: URL extraction (urls, url_count), routing depth (hops_count), and attachment metadata (count, types, individual and total size). All personally identifiable information was anonymized through manual redaction and controlled substitution. Psychological annotations were produced using the same triple-blind protocol by three independent annotators with domain expertise in cybersecurity and social engineering analysis. Each annotator assigned binary presence labels (0/1) for five persuasion dimensions and provided a brief textual justification. Consolidated consensus labels were obtained by majority voting with adjudication in cases of full disagreement.
Data source location	Institution: Universidad Iberoamericana Mexico City (IBERO) City/Town/Region: Mexico City (Álvaro Obregón) Country: Mexico Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: Not applicable (digital dataset; emails collected via multiple inboxes and anonymized). Primary data sources (if secondary data): Not applicable (primary data collected by the dataset creators).
Data accessibility	Repository name: Mendeley Data Data identification number (DOI or persistent identifier): 10.17632/hz2d6gz7pc.5 Direct URL to data:https://data.mendeley.com/datasets/hz2d6gz7pc/5 Instructions for accessing these data: Open access (CC BY 4.0). No login required. Download the files from the landing page using "Download all" or by selecting individual files in the "Files" section. No access restrictions apply beyond the CC BY 4.0 license terms; all personally identifiable information has been removed through manual anonymization prior to release. The dataset is versioned on Mendeley Data; the current release corresponds to version 4 (v4). Future updates, if any, will be published as new versions under the same DOI.
Related research article	None

1. Value of the Data

- Provides a native Spanish-language email corpus for benchmarking phishing detection models and for evaluating cross-lingual transfer from English-trained systems [1].
- Offers a clean binary distinction between phishing and legitimate emails, enabling supervised classification, class-imbalance studies, and evaluation of task-specific feature engineering without contamination from generic spam.
- Supplies message-level persuasion labels for five principles, supporting multi-label learning, inter-annotator agreement analysis, and studies on the relationship between technical indicators and social-engineering strategies.
- Includes human-written justification fields that can be used to train and evaluate explainable NLP systems and to benchmark rationale-generation methods against expert evidence.
- Combines raw text with structural features (URLs, attachments, routing depth), allowing the development and comparison of text-only, metadata-only, and hybrid detection pipelines.
- Covers an 11-year collection period, enabling temporal drift analysis and longitudinal studies of phishing tactics in Spanish-language communications.
- Provides a fully documented tabular structure with schema and open access, facilitating direct reuse for reproducible experiments and educational purposes.

2. Background

SpaPhish was compiled to address the limited availability of publicly accessible Spanish-language email corpora that separate phishing from legitimate communication and preserve native linguistic content [2]. Many widely used phishing email resources are English-centric, such as the Nazario phishing corpus [3] and Enron-derived email collections [4], and they typically lack message-level behavioural annotations. This limits their reuse for studies focused on Spanish-language social engineering, persuasion strategies, and the interaction between linguistic and technical indicators.

Table 1 compares SpaPhish with representative publicly available phishing email datasets. To the best of our knowledge, no publicly available phishing email corpus in Spanish existed prior

Table 1
Comparison of SpaPhish with representative phishing email datasets.

Dataset	Language	Phishing/ Legitimate distinction	Psychological Annotations	Manual Verification	Availability
Nazario Phishing Corpus [3]	English	Predominantly phishing (no guaranteed purity)	No	Yes (manual, single annotator, error-prone)	Public
Enron Corpus [4]	English	Unlabelled real-world email corpus (contains ham, spam, unknown)	No	No (no class annotation protocol)	Public
IWSPA-AP [5]	English	Phishing + Legitimate (multi-source, includes synthetic emails)	No	No (automated preprocessing only; no per-message manual verification)	Public (after request)
SpaPhish	**Spanish**	**Phishing + Legitimate (spam excluded)**	**Yes (5 PoP dimensions, multi-annotator)**	**Yes**	**Public (CC BY 4.0)**

to this release. SpaPhish is, to our knowledge, the only publicly available Spanish-language corpus that maintains a strict separation between phishing and legitimate emails excluding generic spam, includes expert-validated psychological annotations based on Principles of Persuasion, and provides multi-annotator labels with individual justifications for each message.

The dataset follows a combined technical and behavioural perspective. In addition to raw text, SpaPhish provides deterministic structural variables extracted from the original email artifacts, including URL-related features, attachment metadata, and routing depth. A persuasion-annotation layer aligned with Ana Ferreira's Principles of Persuasion framework [6] was incorporated to capture the presence of social-engineering strategies at the message level. Each email was evaluated under a triple-blind protocol by three independent annotators, who assigned binary labels and produced brief Spanish justifications. Consolidated consensus labels are also provided to support direct reuse in supervised and agreement-based studies.

The 2014–2025 temporal coverage enables longitudinal analyses of phishing activity in Spanish-language communications, including the study of temporal drift in technical artifacts and persuasion strategies. By combining native text, structural indicators, and human-produced behavioural annotations in a single tabular resource, SpaPhish provides a reusable dataset for reproducible research in phishing detection, explainable natural language processing, and annotation reliability.

3. Data Description

SpaPhish comprises 1395 real-world email messages collected between June 2014 and October 2025, with 731 phishing (52.4%) and 664 legitimate (47.6%) instances. Each record follows a three-layer schema: (i) raw textual content (subject and body), (ii) structural technical variables derived from the original email artifacts (e.g., URLs, attachments, and routing depth), and (iii) message-level persuasion annotations.

The persuasion layer was produced under a triple-blind protocol by three independent annotators with domain expertise in cybersecurity and social-engineering analysis. For each message, binary presence labels (0/1) are provided for five persuasion principles together with short Spanish justifications and a consolidated consensus label.

This layered organization allows the independent or combined reuse of textual, technical, and behavioral variables for reproducible analyses and benchmarking.

The analyses and summary statistics reported in this article correspond to SpaPhish version 4, publicly available in Mendeley Data (doi:10.17632/hz2d6gz7pc.5; https://data.mendeley.com/datasets/hz2d6gz7pc/5).

3.1. Repository inventory and file integrity

SpaPhish is released as a structured repository designed for programmatic access and cross-disciplinary reuse. The repository includes the following artifacts:

- SpaPhish dataset-DiB.csv: The primary tabular file containing 1395 unique email records encoded in UTF-8 and formatted as a semicolon-delimited table. It integrates raw textual fields, derived technical variables, and persuasion-annotation fields across 47 columns.
- dataset_schema.json: A machine-readable schema specification defining data types, nullability constraints, and field-level descriptions for every CSV column. This file supports automated validation and consistent ingestion. This file serves as the complete reference schema for all 47 variables in the dataset, including data types, field-level descriptions, and nullability constraints.
- README.txt: Technical documentation providing dataset-loading instructions, the complete data dictionary, licensing information (CC BY 4.0), column naming conventions, example usage, and guidance for interpreting the 'justif_*' fields and the consolidated consensus labels.

- SpaPhish_html_report.zip: An HTML report summarizing dataset characteristics and descriptive statistics for exploratory inspection.

3.2. Multi-Layered data schema (47 variables)

The released SpaPhish CSV contains 1395 emails (rows) described by 47 variables grouped into three layers: (i) identification and raw text, (ii) structural technical indicators (URLs, attachments, routing depth), and (iii) persuasion-annotation variables based on Principles of Persuasion (PoP). This organization allows users to work with text-only, metadata-only, or combined representations, and to reuse the annotation layer for supervised learning and agreement analyses. A complete schema of all 47 variables, including data types and field-level descriptions, is available in the dataset_schema.json file provided in the Mendeley Data repository.

3.2.1. Layer 1: identification and text

Each row represents a single email and includes five primary variables: a unique identifier, subject, body, timestamp, and class label:

- **hash:** (String) A unique SHA-256 identifier computed from the raw email message prior to parsing and anonymization ($N = 1395$). This value enables record linkage and duplicate detection without exposing the original message.
- **subject:** (String) Raw subject line exactly as extracted from the email header, without cleaning or normalization. Subject lines are short (mean = 51.7 characters, median = 47, IQR = 30–67; mean = 7.9 words, median = 7, IQR = 4.5–10). Fig. 1 shows their right-skewed distribution.
- **body:** (String) Plain-text message content. When the original format was HTML, tags were removed and only visible text was retained. No additional preprocessing was applied. Message length is highly variable (mean = 1205.5 characters, median = 706, IQR = 450.5–1276; mean = 158.8 words, median = 102, IQR = 66–184). Fig. 2 summarizes this variability.

date: (String) Original message timestamp used for temporal analyses. Values were not discretized or transformed. The dataset spans June 2014–October 2025. Fig. 3 shows the monthly distribution of messages. The dataset was finalized and released as a fixed snapshot; therefore, no messages dated after October 2025 are included.

Table 2 reports the annual distribution of emails by class. The corpus exhibits a temporal imbalance, with 57.7% of messages concentrated in 2024–2025. The proportion of phishing messages increases substantially in recent years, reflecting the growing volume of phishing campaigns targeting Spanish-language users during this period. The uneven temporal distribution

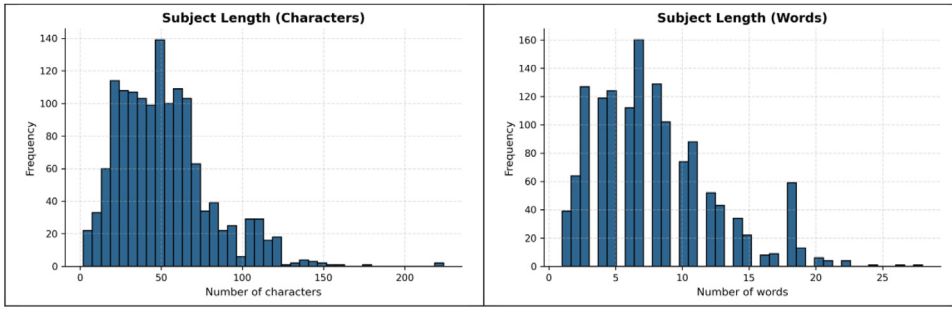

Fig. 1. Distribution of email subject length. Left: The histogram of subject length measured in characters. Right: The histogram of subject length measured in words. Both panels summarize the empirical length of distribution of subject lines in the dataset, highlighting their compact and right-skewed nature.

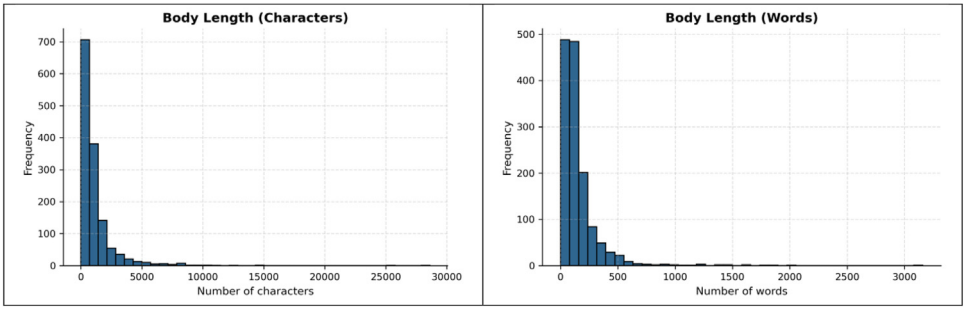

Fig. 2. Distribution of email body length. Left: Histogram of body length measured in characters. Right: The histogram of body length measured in words. The distributions exhibit strong right skewness, reflecting substantial variability in message body size across the dataset.

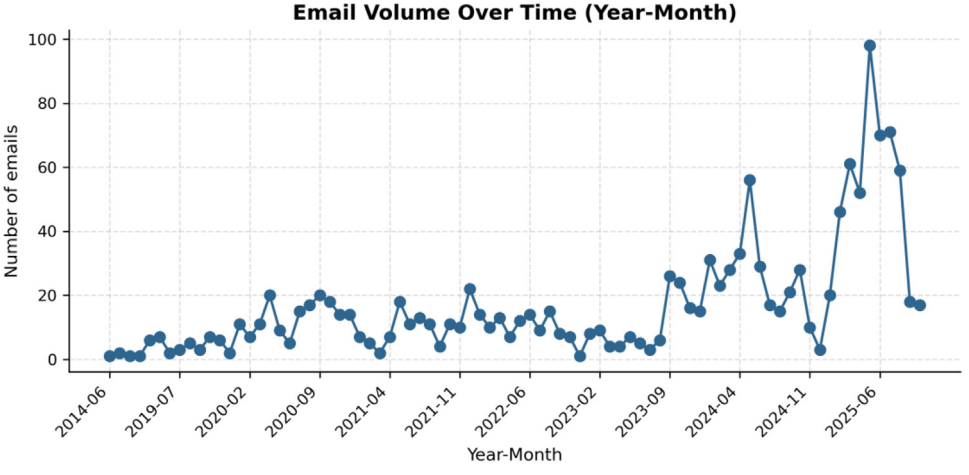

Fig. 3. Email volume over time (year–month). Monthly aggregation of email messages across the full observation period (June 2014–October 2025), showing the temporal distribution of samples and variations in email volume over time.

reflects the heterogeneous coverage of individual contributor inboxes: not all contributors had messages spanning the full 2014–2025 period, and the absence of messages in certain years (e.g., 2015–2017) is attributable to gaps in the available inbox archives rather than to systematic exclusion criteria. Twenty-four records have missing or unparseable date fields and are excluded from the annual breakdown but retained in the dataset.

- **label:** (Integer) Binary class label (0 = legitimate, 1 = phishing) assigned to all 1395 emails using the same triple-annotator protocol applied to the persuasion dimensions. Each message was independently evaluated by three annotators, consensus labels were obtained by majority voting, and cases of complete disagreement were resolved by adjudication from a fourth expert. The final dataset contains 731 phishing (52.4%) and 664 legitimate (47.6%) emails. Fig. 4 presents the class distribution.

3.2.2. Layer 2: forensic layer (Structural indicators)

This layer contains quantitative variables extracted from the original email content and headers. These fields describe the structural characteristics of embedded links, attachments, and message routing:

Table 2
Annual distribution of emails by class.

Year	Legitimate	Phishing	Total	% Phishing
2014	1	0	1	0.0
2018	3	0	3	0.0
2019	41	1	42	2.4
2020	160	1	161	0.6
2021	105	16	121	13.2
2022	90	20	110	18.2
2023	72	55	127	43.3
2024	103	191	294	65.0
2025	89	423	512	82.6
No date	—	—	24	—
Total	**664**	**731**	**1395**	**52.4**

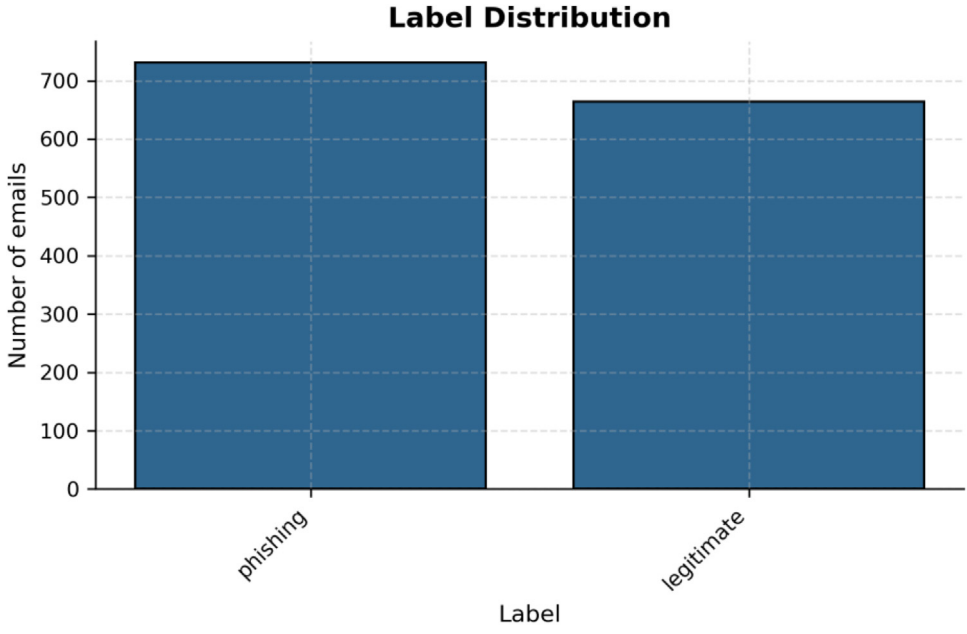

Fig. 4. Class label distribution in the dataset. Bar chart showing the number of email messages per class: 731 phishing (52.4%) and 664 legitimate (47.6%), illustrating the near-balanced distribution between classes.

- **url_count:** (Integer) Number of URLs detected in the email. Values range from 0 to 117 (mean = 6.62, median = 4), with a strongly right-skewed distribution. HTTPS links represent 89.68% of all extracted URLs. Fig. 5 shows the distribution of URLs per message and the most frequent domains.
- **urls:** (List) Serialized list of extracted URLs preserving the original scheme and domain. Emails without links are represented by an empty list and have url_count = 0.
- **attachments_count:** (Integer) Number of attached files per email (mean = 0.36, median = 0, range 0–29), indicating a sparse distribution. Fig. 6 shows the attachment-count distribution.
- **attachments_types:** (List) Serialized list of file extensions associated with the attachments. Emails without attachments are represented by an empty list. Fig. 7 shows the frequency of attachment types.
- **attachments_total_size:** (Integer) Total cumulative size of all attachments in bytes. Values range from 0 to approximately 16.2 MB (mean approximately 107 KB), with a highly right-

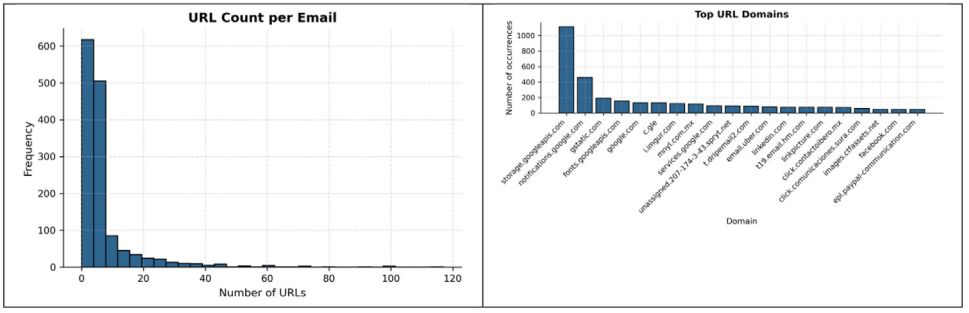

Fig. 5. Distribution of URL counts per email. Left: distribution of the number of URLs per email, showing a strongly right-skewed pattern with most messages containing few links and a small number of link-heavy outliers. Right: frequency of the most common URL domains, highlighting the concentration of links on a limited set of recurrent hosting and notification infrastructures.

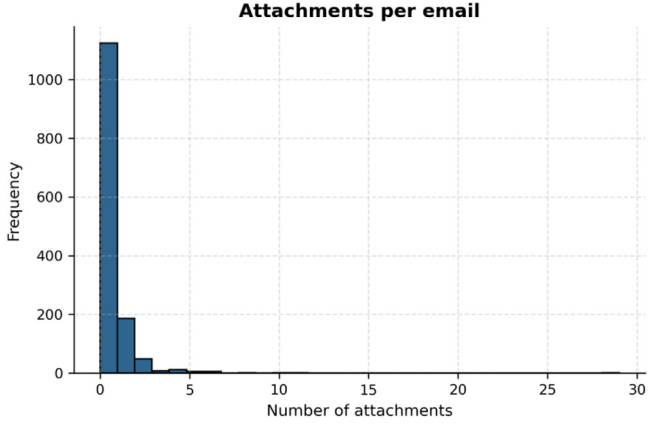

Fig. 6. Distribution of the number of attachments per email. Histogram of attachment counts across messages, showing a highly sparse and right-skewed distribution in which most emails contain no attachments, while a small number of messages include multiple files, with extreme cases reaching up to 29 attachments.

skewed distribution dominated by messages without attachments. Fig. 8 summarizes the distribution of individual attachment sizes as recorded in the attachments_sizes field.

- **attachments_sizes:** (List) Serialized list of individual attachment sizes in bytes. Emails without attachments have an empty list and attachments_total_size = 0.
- **hops_count:** (Integer) Routing depth computed as the number of Received headers (mean = 4.28, median = 3, max = 19). Fig. 9 shows the distribution of routing depth.

3.3. Layer 3: psychological annotation layer (Principles of persuasion)

This layer contains human annotations of persuasion strategies expressed at the message level, following the framework of Ferreira et al. [6]. Each email was independently evaluated by three annotators with domain expertise in cybersecurity and social-engineering analysis across five dimensions: Authority, Social_Proof, Liking/Similarity/Deception, Commitment/Integrity/Reciprocation, and Distraction. These annotations were applied to the entire corpus (both phishing and legitimate emails) to support comparative studies. In the context of legitimate communications, positive labels for dimensions such as Authority or Liking/Similarity reflect standard professional or personal interactions (such as a supervisor providing instructions

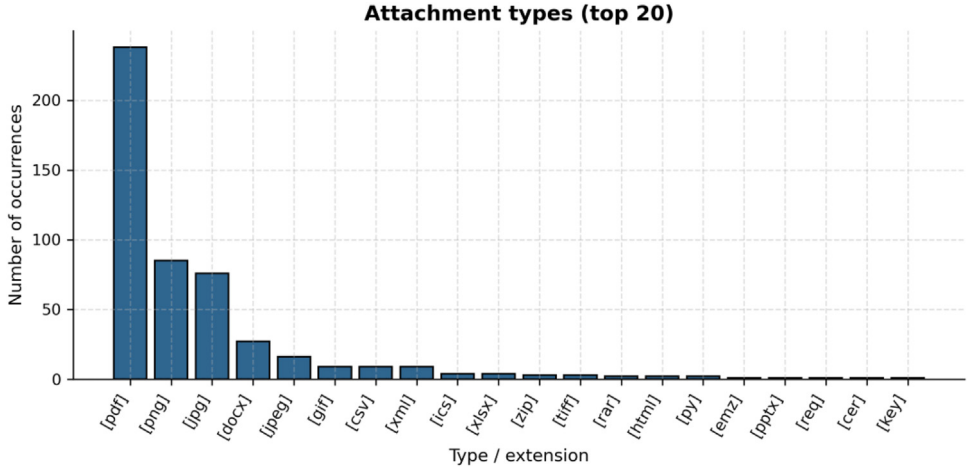

Fig. 7. Distribution of attachment file types. Bar chart showing the frequency of the 20 most common file extensions extracted from the attachments_types field across all emails containing at least one attached file, illustrating the relative prevalence of different attachment formats in the corpus.

Fig. 8. Distribution of individual attachment sizes. Histogram of per-attachment file sizes (in bytes), showing a highly right-skewed distribution in which most attachments are small, while a limited number of files exhibit substantially larger sizes, reaching values on the order of several megabytes.

or a cordial greeting) rather than malicious social engineering or deceptive intent. For instance, while the Liking/Similarity/Deception dimension captures deceptive rapport in phishing, in legitimate emails it simply identifies instances of interpersonal similarity or affinity expressed in the text.

For each dimension, three binary variables are provided (*_A, *_B, *_C), where 0 indicates absence and 1 indicates presence of the corresponding persuasion principle. For example, authority_A, authority_B, and authority_C store the independent decisions of the three annotators.

In addition, qualitative justification fields (justif_*) contain short textual rationales describing the evidence supporting each decision. These fields preserve the annotation trace and enable qualitative inspection, auditing, and explainability-oriented analyses. Justifications consist of short, structured rationales written in Spanish following conventions agreed upon during the kickoff session, ensuring terminological consistency across annotators while preserving individual evaluative judgment. Representative examples include: for Authority, a positive label justified

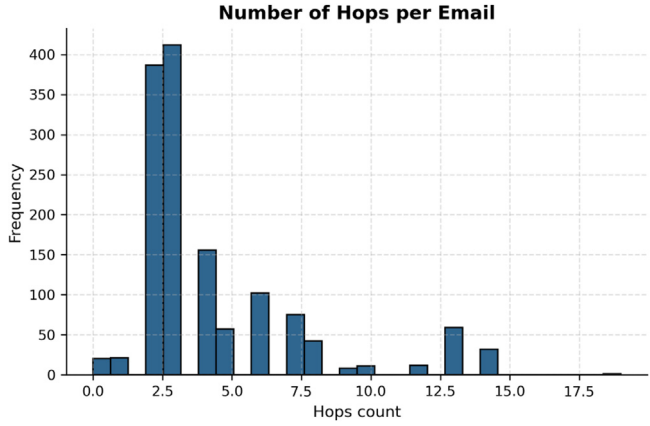

Fig. 9. Distribution of email routing depth. Histogram of the number of transmission hops per email, computed from Received headers, showing the frequency of routing depths across messages and the presence of a right-skewed tail with a small number of emails traversing a higher number of intermediate relays.

Table 3
Number and percentage of emails with positive consensus label for each persuasion principle.

Principle	N positive	% of dataset
Authority	824	59.07
Social Proof	142	10.18
Liking / Similarity / Deception	918	65.81
Commitment / Integrity / Reciprocation	311	22.29
Distraction	777	55.70

as "*Se hace pasar por una autoridad (servicio de entrega de paquetes)*" [It impersonates an authority (parcel delivery service)] and a negative as "*No se hace pasar por una figura de autoridad*" [It does not impersonate an authority figure]; for Distraction, a positive label justified as "*Crea preocupación y distracción por un paquete no recibido, impulsando una acción inmediata*" [It creates concern and distraction about an undelivered package, prompting immediate action] and a negative as "*No crea urgencia, escasez ni distracción emocional*" [It does not create urgency, scarcity, or emotional distraction]".

A consolidated consensus label is provided for each persuasion dimension using majority voting over the three individual annotations. In cases of full disagreement, a final label was assigned through adjudication by a fourth expert, a senior cybersecurity researcher who did not participate in the initial annotation round.

The availability of individual annotations, consensus labels, and justification texts enables reuse for supervised multi-label learning, inter-annotator agreement measurement, and explainable natural language processing.

Fig. 10 shows the proportion of positive labels for each persuasion principle in the dataset.

Table 3 reports the number and percentage of emails with a positive consensus label for each persuasion principle.

3.4. Data quality assessment

Inter-annotator agreement was assessed using Krippendorff's alpha (α) and Cohen's kappa (κ) computed from the individual annotations of the three annotators prior to adjudication. Table 4 reports the agreement metrics for each persuasion dimension. Agreement levels range from moderate to substantial, which is consistent with the inherently subjective nature of psychological annotation tasks. Lower agreement values for principles such as Liking/Similarity/Deception

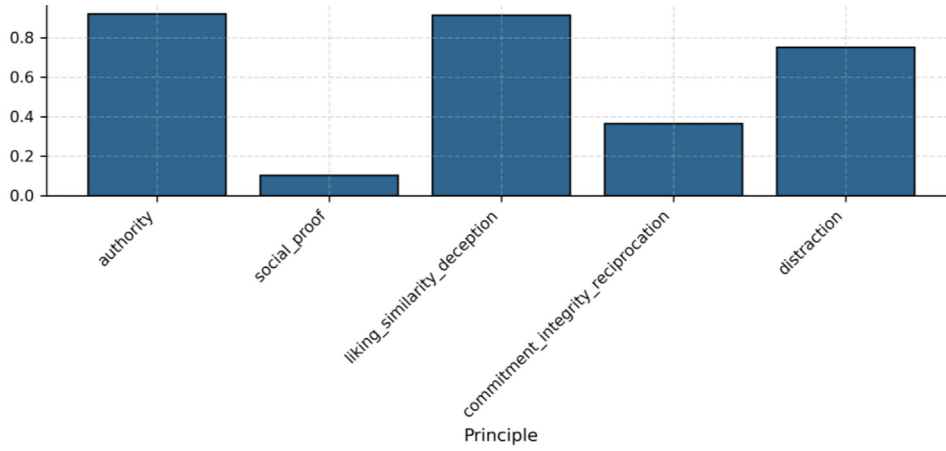

Fig. 10. Positive rate by persuasion principle. Proportion of emails with a positive consensus label for each of the five Principles of Persuasion (Authority, Social Proof, Liking/Similarity/Deception, Commitment/Integrity/Reciprocation, and Distraction), computed from the majority-voting consensus labels stored in the dataset.

Table 4
Inter-annotator agreement metrics per persuasion dimension, computed from the individual annotations of the three annotators before adjudication. Krippendorff's alpha (α) is computed simultaneously over the three annotators. Cohen's kappa (κ) is reported for each annotator pair (A–B, A–C, B–C) and as the mean across the three pairs.

Principle	Krippendorff's α	Cohen's κ (A–B)	Cohen's κ (A–C)	Cohen's κ (B–C)	Mean κ
Authority	0.356	0.604	0.269	0.336	0.403
Social Proof	0.498	0.574	0.418	0.526	0.506
Liking/Similarity/Deception	0.287	0.359	0.273	0.248	0.293
Commitment/ Integrity/ Reciprocation	0.516	0.563	0.564	0.443	0.523
Distraction	0.655	0.741	0.611	0.610	0.654

and Authority reflect the greater interpretive ambiguity of these constructs compared to more explicitly marked principles such as Distraction.

Dataset completeness was evaluated through missing-value diagnostics. For each variable, the number and percentage of missing entries were computed, and their distribution was inspected to distinguish structural absence from extraction or annotation gaps.

Missing values are concentrated in attachment-related fields and URL lists, which are only populated when the corresponding artifact is present in the message. Core variables (hash, subject, body, date, label, hops_count, and persuasion annotations) are nearly complete. No imputation was performed; missing values are preserved to maintain fidelity to the extracted records.

Fig. 11 shows the missingness matrix for variables with incomplete coverage, and Fig. 12 reports the percentage of missing values per column.

4. Experimental Design, Materials and Methods

4.1. Data acquisition and corpus construction

SpaPhish was constructed from a curated collection of real-world Spanish-language emails collected between June 2014 and October 2025 from the personal and institutional inboxes of all dataset contributors. The dataset does not originate from a single organization or from any previously published public corpus. Instead, it was constructed by aggregating emails contributed

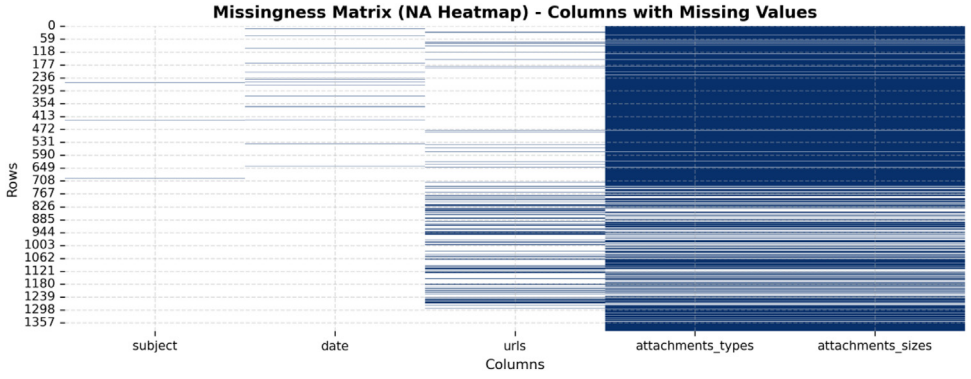

Fig. 11. Missingness matrix across variables with incomplete coverage. Binary heatmap indicating the presence and absence of values by row. Missingness is primarily structural and associated with messages without attachments or URLs.

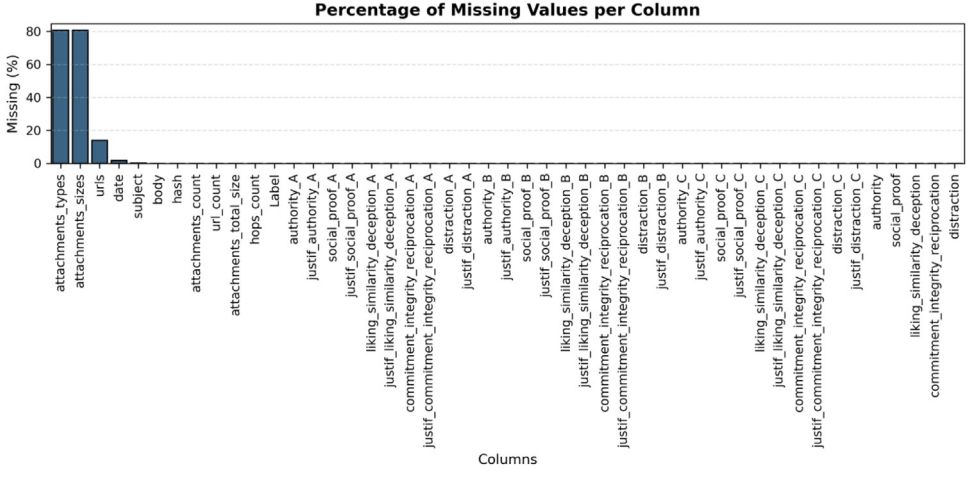

Fig. 12. Percentage of missing values per column. Proportion of missing entries for each variable, showing that incompleteness is confined to structurally dependent fields.

by the co-authors from their personal and institutional email accounts, which were hosted on a variety of common email platforms and service providers.

Each contributor provided a selection of messages from their personal and institutional inboxes without applying formal inclusion criteria. All contributed messages were pooled into a raw collection and subsequently labeled by the annotation committee through the triple-annotator protocol described below. No pre-filtering or sampling was applied at the collection stage.

The anonymization process was performed manually in a controlled and documented manner. The following entity types were treated: personal names, organizations, email addresses, phone numbers, account/credential identifiers, physical addresses, and location-specific references. Sensitive spans were replaced with fictitious surrogates belonging to the same semantic category as the original entity (e.g., a personal name was replaced by a different invented name, an organization by a fictitious organization name, and a location by an invented location), preserving grammatical role and surface form while preventing re-identification. Substitutions were applied consistently within each email (i.e., the same original entity was mapped to the same surrogate within that message), without enforcing global cross-email identity linking. No masking, hashing, or automated placeholder substitution was applied to the textual content.

A final quality-control pass verified that no residual direct identifiers remained and that the resulting text preserved readability and coherent discourse structure. Sensitive entities were identified through manual inspection of each message by the dataset contributors. No automated tools, heuristics, or named entity recognition systems were used at any stage of the anonymization process. The impact of anonymization on downstream NLP tasks is expected to be minimal, as substitutions preserve the semantic category, grammatical role, and surface form of the original entities.

Each record corresponds to a single email instance and is treated as an independent observational unit. The corpus includes phishing and legitimate communications. Generic spam was excluded and phishing labels were assigned following the NIST definition of phishing [7], reserving the label for targeted social-engineering attempts intended to deceive individuals into disclosing sensitive information or executing harmful actions (e.g., credential disclosure or malicious link execution), rather than bulk unsolicited advertising.

Prior to labeling, duplicate messages were identified and removed using SHA-256 hashes computed over the raw email message prior to parsing and anonymization. Messages were assigned a binary label (phishing or legitimate) using the same triple-annotator protocol applied to the persuasion dimensions described below. Each annotator independently assessed message intent, content, and contextual cues to determine whether a message constituted a targeted social-engineering attempt or a legitimate communication. Consensus labels were derived by majority voting, with adjudication by a fourth expert in cases of full disagreement. Emails were retained if they contained at least a subject line or a message body. No synthetic data, automated generation, or data augmentation was used. The corpus was filtered to Spanish-language emails; messages containing minor non-Spanish fragments (e.g., automated footers, disclaimers, or signatures) were retained when the primary communicative content was in Spanish.

4.2. Message parsing and textual field extraction

Raw email artifacts were parsed to extract the subject line (subject), message body (body), and message date (date). Subject and body are stored as raw text exactly as extracted, without normalization, tokenization, or linguistic preprocessing. When the original payload contained HTML, tags were removed and only the visible text was retained, preserving the original wording and ordering; no further transformations were applied.

The date field was extracted from the email header and stored as provided by the source metadata to support temporal analyses. Keeping textual fields unaltered enables downstream users to apply task-specific preprocessing pipelines as needed.

4.3. Technical and structural feature extraction

A set of structural technical variables was derived from the parsed message content and headers using deterministic, rule-based extraction procedures:

- URL-related features were obtained by identifying all hyperlinks present in the message body. For each email, url_count stores the number of detected URLs and urls stores the serialized list of extracted links. URL schemes and domains were parsed directly from the extracted strings; no URL resolution, crawling, or external network interaction was performed.
- Attachment-related features were obtained from message metadata. attachments_count stores the number of attached files, attachments_types stores the serialized list of file extensions, and attachments_sizes stores per-attachment sizes in bytes. attachments_total_size stores the cumulative attachment size computed as the sum of individual sizes. Missing values in attachment-related fields reflect structural absence (emails without attachments) rather than extraction failure.
- Routing depth was quantified as hops_count, computed as the number of Received headers in the email trace.

These procedures are fully deterministic to support consistent regeneration of the released technical fields.

Technical variables were extracted from raw .eml files using a deterministic processing pipeline implemented in Python 3.12. Email messages were parsed with the standard email library (BytesParser, policy.default), and HTML content was converted to plain text using BeautifulSoup when no text/plain MIME part was available. No tokenization, stemming, lemmatization, lowercasing, stopword removal, or other text normalization procedures were applied during feature extraction. The complete source code used for parsing, feature extraction, and dataset construction is publicly available in the SpaPhish GitHub repository (https://github.com/lbustio/spa_phish). The software dependencies required to reproduce the processing pipeline are documented in the accompanying requirements.txt and environment.yml files.

4.4. Psychological annotation protocol: principles of persuasion

The dataset includes message-level human annotations of persuasion strategies following the framework proposed by Ana Ferreira et al. [6]. This taxonomy was selected because it was specifically developed and validated in the context of phishing and social engineering attacks, making it more appropriate for this domain than general-purpose persuasion frameworks. Each email was independently evaluated across five dimensions: Authority, Social Proof, Liking/Similarity/Deception, Commitment/Integrity/Reciprocation, and Distraction. These dimensions are not mutually exclusive; annotators evaluated each principle independently, so a single email may carry positive labels for multiple persuasion dimensions simultaneously.

A triple-annotator protocol was applied for each dimension. The annotation panel comprised three independent domain experts: a Threat Intelligence Analyst with over 20 years of industry expertise, a psychologist specializing in Social Psychology with over 15 years of experience, and a senior cybersecurity researcher. Given their established expertise, no formal training procedures were required prior to annotation. A structured kickoff session was conducted in which annotation guidelines were presented and discussed. These guidelines consisted of the NIST definition of phishing [7] for the binary label and the Principles of Persuasion framework of Ferreira et al. [6] for the persuasion dimensions. Annotators then worked independently, assigning binary presence labels (0 = absence, 1 = presence). These labels are stored using the suffix convention *_A, *_B, and *_C, corresponding to the three annotators.

In addition to binary labels, each annotator provided a short qualitative justification (justif_*) describing the textual evidence supporting the decision. Justifications were written in Spanish to preserve linguistic consistency with the source material.

A consolidated consensus label was computed for each persuasion dimension using majority voting over the three individual judgments. In cases of full disagreement, the final label was assigned through adjudication by a fourth expert, a senior cybersecurity researcher who did not participate in the initial annotation round. In addition, when a numerical majority was obtained but the accompanying justifications revealed substantial conceptual disagreement among annotators, the case was also referred to the fourth expert for adjudication to ensure that the final label reflected substantive agreement rather than coincidental numerical consensus. Consensus labels are stored without suffixes and are provided alongside individual annotations to enable agreement analyses and supervised learning under different label assumptions.

4.5. Data quality control and consistency checks

Multiple quality-control procedures were applied throughout dataset construction. Structural validation verified that numerical fields fall within expected ranges and that serialized list fields follow consistent formatting conventions.

Missing values were audited across all variables. Missingness is predominantly structural, arising from the absence of URLs or attachments in many messages rather than from extrac-

tion failures. No imputation or synthetic filling was performed; missing values are preserved to reflect the observed record structure.

Missing values were audited across all variables. Missingness is predominantly structural, arising from the absence of URLs, attachments, or parseable dates rather than from extraction failures. No imputation or synthetic filling was performed, and no records were removed due to missing values after corpus construction. Missing entries were retained as observed to preserve the original structure of each email record.

Annotation consistency is supported by retaining all individual annotator labels and justification texts, enabling post hoc auditing and inter-annotator agreement analyses. As a post-annotation quality control step, a random sample of 10% of the records was reviewed by the research team to verify the consistency and correctness of the assigned labels and justifications. No systematic errors or inconsistencies were identified during this review.

4.6. Software environment and reproducibility

Data parsing, feature extraction, annotation aggregation, and quality-control procedures were implemented in Python. The released dataset is provided as UTF-8 encoded CSV together with an explicit machine-readable schema. An interactive HTML report (SpaPhish_html_report.zip) is included in the Mendeley Data repository. The report was generated automatically from the released dataset using the accompanying analysis code and provides variable-level descriptive statistics, frequency tables, missing-value summaries, and graphical visualizations of the principal distributions reported in the manuscript. Its purpose is to facilitate exploratory inspection of the dataset and to enable direct verification of the summary statistics presented in this article. Because the report is generated programmatically from the shared dataset, it can be reproduced by any user using the provided source code. The processing pipeline is deterministic and does not rely on external services or stochastic components.

Limitations

SpaPhish spans a long temporal window (2014–2025) but contains 1395 emails. This reflects practical constraints in collecting and manually curating real-world phishing messages suitable for multi-annotator labeling.

The corpus was built from volunteer-contributed inboxes and therefore follows a convenience sampling strategy rather than a statistically representative sampling design. Therefore, the dataset may reflect demographic, regional, or provider-specific biases (e.g., variations in Spanish usage, common brands, and service infrastructures). In addition, phishing tactics evolve over time; models trained on older portions of the corpus may not transfer optimally to more recent campaigns without temporal validation. Finally, manual anonymization and controlled substitution preserve readability but may introduce residual artifacts; users should consider this when training models sensitive to entity patterns.

No limitations affecting data integrity or file-level completeness were identified in the released repository.

Ethics Statement

The authors confirm compliance with the ethical requirements for publication in Data in Brief. The dataset was constructed from emails obtained from inboxes managed by the dataset contributors, all of whom are co-authors of this work and provided explicit consent to process their messages and release anonymized derivatives for public research use. No external participants were involved in data collection.

Because the dataset was constructed exclusively from materials voluntarily contributed by the co-authors, involved no external human participants, and only anonymized derivatives were released, no institutional ethics committee approval was required.

The anonymization procedure was designed to comply with applicable data protection regulations, including the Mexican Federal Law on Protection of Personal Data Held by Private Parties (Ley Federal de Protección de Datos Personales en Posesión de los Particulares). All personally identifiable information was removed prior to release.

Potential misuse is mitigated by the fact that the dataset contains historical phishing messages released solely for research and educational purposes and does not include executable malicious content, active payloads, or personally identifiable information.

Before inclusion, all messages underwent a strict manual anonymization procedure. Personally identifiable information (PII), financially identifiable information, and other potentially sensitive content (including location-specific references) were removed. When replacement was necessary to preserve readability and contextual coherence, original entities were substituted with fully fictitious but semantically consistent surrogates (e.g., personal names and locations were replaced by invented alternatives while preserving grammatical and discourse consistency). No real identities, addresses, account numbers, or other identifying traces were retained.

This work does not involve experiments with human participants or animals, and it does not include data collected from social media platforms.

The SpaPhish dataset and all associated annotations are original contributions that have not been previously published, deposited in any other repository, or included in any other dataset or publication. No part of the data or annotations has been disseminated elsewhere prior to this submission.

CRediT Author Statement

Lázaro Bustio-Martínez: Conceptualization, Methodology, Software, Data curation, Investigation, Visualization, Writing – Original Draft, Supervision, Project administration. **Viviana Inés Fuentes-Fuentes:** Conceptualization, Methodology, Data curation, Investigation, Supervision. **Luisa Fernanda Agudelo Fuentes:** Software, Data curation. **Vitali Herrera-Semenets:** Conceptualization, Methodology, Software, Data curation, Investigation, Writing – Original Draft, Writing – Review & Editing, Supervision, Project administration. **Darián Llanes-Guilarte:** Software, Data curation, Visualization. **Felipe Antonio Trujillo-Fernández:** Conceptualization, Methodology, Data curation, Investigation. **Antonio Carlos Cardeña-Matamoros:** Conceptualization, Methodology, Data curation. **Carlos Francisco Betancourt-Moreno:** Conceptualization, Methodology, Data curation, Visualization. **Andrés Guillermo Molano-Jiménez:** Data curation, Funding acquisition. **Jan van den Berg:** Conceptualization, Methodology, Investigation, Writing – Review & Editing, Supervision, Funding acquisition.

Data Availability

SpaPhish: A Spanish Dataset for Phishing and Psychological Pattern Detection (Original data) (Mendeley Data).

Acknowledgements

This work was supported by the Departamento de Estudios en Ingeniería para la Innovación, Universidad Iberoamericana, Ciudad de México, through the project “Creación de un dataset de phishing en español”. During the preparation of this work, the authors used ChatGPT to improve

the grammar and clarity of the manuscript. Following the use of this tool, the authors thoroughly reviewed and edited the content as needed and take full responsibility for the final version of the publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] N. Altwaijry, I. Al-Turaiki, R. Alotaibi, F. Alakeel, Advancing phishing email detection: a comparative study of deep learning models, Sensors 24 (7) (2024) 2077 Art. no..
[2] A.I. Champa, M.F. Rabbi, M.F. Zibran, Curated datasets and feature analysis for phishing email detection with machine learning, in: Proceedings of the IEEE International Conference on Computing and Machine Intelligence (ICMI), 2024, pp. 1–7.
[3] J. Nazario, "Phishing corpus," 2006. [Online]. Available: http://monkey.org/~jose/phishing/. [Accessed: Apr. 2026].
[4] B. Klimt, Y. Yang, The Enron corpus: a new dataset for email classification research, in: Proc. European Conference on Machine Learning (ECML), 2004, pp. 217–226.
[5] A.E. Aassal, L.F.T. Moraes, S. Baki, A. Das, R.M. Verma, Anti-phishing pilot at ACM IWSPA 2018: evaluating performance with new metrics for unbalanced datasets, in: Proc. Anti-Phishing Shared Task at ACM IWSPA (IWSPA-AP 2018), CEUR Workshop Proc, 2124, 2018 [Online]. Available:.
[6] A. Ferreira, L. Coventry, G. Lenzini, Persuasion: how phishing emails can influence users and bypass security measures, Int. J. Human-Comput. Stud. 125 (2019) 19–31.
[7] National Institute of Standards and Technology (NIST), "Phishing," NIST computer security resource center. [Online]. Available: https://csrc.nist.gov/glossary/term/phishing. [Accessed: Apr. 2026].