Phishing URL Detection Dataset: Feature Descriptions

This document contains detailed descriptions of the features present in the phishing URL detection dataset used to train a hybrid LSTM-CNN deep learning model.

URL Structure Features

- **url**: The actual web address being analyzed.
- **length_url**: Total number of characters in the URL.
- **length_hostname**: Length of the domain name (hostname).
- **ip**: Presence of an IP address instead of a domain name.
- **nb_dots**: Number of "." (dots) in the URL.
- **nb_hyphens**: Number of hyphens in the URL.
- **nb_at**: Number of "@" symbols.
- **nb_qm**: Number of "?" symbols.
- **nb_and**: Number of "&" symbols.
- **nb_or**: Number of "|" symbols.
- **nb_eq**: Number of "=" symbols.
- **nb_underscore**: Number of "_" symbols.
- **nb_tilde**: Number of "\~" symbols.
- **nb_percent**: Number of "%" symbols.
- **nb_slash**: Number of forward slashes "/".
- **nb_star**: Number of "*" symbols.
- **nb_colon**: Number of ":" characters.
- **nb_comma**: Number of commas.
- **nb_semicolumn**: Number of semicolons.
- **nb_dollar**: Number of "\$" symbols.
- **nb_space**: Number of space characters.
- **nb_www**: Number of times "www" appears.
- **nb_com**: Number of times ".com" appears.
- **nb_dslash**: Number of "//" substrings.
- **http_in_path**: Indicates if "http" appears in the path.
- **https_token**: Indicates if "https" is used incorrectly in the domain or path.

Ratio and Encodings

- **ratio_digits_url**: Ratio of digits in the entire URL.
- **ratio_digits_host**: Ratio of digits in the hostname.
- **punycode**: Checks if the domain uses punycode (internationalized domain names).
- **port**: Presence of a port in the URL.

Domain/Path Based Features

- **tld_in_path**: Top-Level Domain (TLD) appears in the path.
- **tld_in_subdomain**: TLD appears in the subdomain.
- **abnormal_subdomain**: Irregular subdomain structure.
- **nb_subdomains**: Number of subdomains.
- **prefix_suffix**: Checks for prefixes/suffixes separated by "-" in domain.
- **random_domain**: Domain contains random strings.
- **shortening_service**: URL is shortened via services like bit.ly.
- **path_extension**: Path has suspicious extensions.
- **nb_redirection**: Number of redirections in URL.
- **nb_external_redirection**: Number of redirections to external domains.

Word-Based Features

- **length_words_raw**: Total characters in all words in URL.
- **char_repeat**: Frequency of character repetition.
- **shortest_words_raw**: Length of the shortest word in raw URL.
- **shortest_word_host**: Shortest word in host.
- **shortest_word_path**: Shortest word in path.
- **longest_words_raw**: Length of the longest word in raw URL.
- **longest_word_host**: Longest word in host.
- **longest_word_path**: Longest word in path.
- **avg_words_raw**: Average word length in full URL.
- **avg_word_host**: Average word length in hostname.
- **avg_word_path**: Average word length in path.

Suspicious Patterns

- **phish_hints**: Presence of known phishing hints.
- **domain_in_brand**: Brand name present in domain.
- **brand_in_subdomain**: Brand present in subdomain.
- **brand_in_path**: Brand present in path.
- **suspecious_tld**: TLDs often associated with phishing (e.g., .tk).
- **statistical_report**: Matches known phishing patterns from stats.

Hyperlink and Resource Analysis

- **nb_hyperlinks**: Total number of hyperlinks.
- **ratio_intHyperlinks**: Internal hyperlink ratio.
- **ratio_extHyperlinks**: External hyperlink ratio.
- **ratio_nullHyperlinks**: Ratio of empty/null hyperlinks.
- **nb_extCSS**: Number of external CSS links.
- **ratio_intRedirection**: Ratio of internal redirection.
- **ratio_extRedirection**: Ratio of external redirection.
- **ratio_intErrors**: Internal error links.

- **ratio_extErrors**: External error links.

HTML and JavaScript Behavior

- **login_form**: Presence of login forms.
- **external_favicon**: External favicon source.
- **links_in_tags**: Hyperlink tags in HTML.
- **submit_email**: Form submits to email address.
- **ratio_intMedia**: Internal media links ratio.
- **ratio_extMedia**: External media links ratio.
- **sfh**: Server Form Handler — suspicious values (e.g., blank).
- **iframe**: Use of iframes.
- **popup_window**: Popup window detection.
- **safe_anchor**: Proportion of anchors with safe hrefs.
- **onmouseover**: Use of onmouseover JavaScript.
- **right_clic**: Right-click is disabled.
- **empty_title**: Page has empty title.
- **domain_in_title**: Domain appears in title.

Domain and Reputation

- **domain_with_copyright**: Domain has a copyright symbol.
- **whois_registered_domain**: WHOIS lookup succeeded.
- **domain_registration_length**: Time until domain expires.
- **domain_age**: Domain age in days.
- **web_traffic**: Site's web traffic rank.
- **dns_record**: Presence of DNS records.
- **google_index**: Whether site is indexed by Google.
- **page_rank**: Page rank of site.

Labels

- **status**: Class label — `legitimate` or `phishing`.