

Research Article

A Large Visual, Qualitative, and Quantitative Dataset for Web Intelligence Applications

Christian Mejia-Escobar ¹, **Miguel Cazorla** ², and **Ester Martinez-Martin** ²

¹Central University of Ecuador, P.O. Box 17-03-100, Quito, Ecuador

²Institute for Computer Research, University of Alicante, P.O. Box 99, Alicante 03080, Spain

Correspondence should be addressed to Christian Mejia-Escobar; cimejia@uce.edu.ec

Received 15 May 2022; Revised 15 September 2022; Accepted 5 April 2023; Published 10 October 2023

Academic Editor: Lorenzo Putzu

Copyright © 2023 Christian Mejia-Escobar et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The Web is the communication platform and source of information par excellence. The volume and complexity of its content have grown enormously, with organizing, retrieving, and cleaning Web information becoming a challenge for traditional techniques. Web intelligence is a novel research area to improve Web-based services and applications using artificial intelligence and automatic learning algorithms, for which a large amount of Web-related data are essential. Current datasets are, however, limited and do not combine visual representation and attributes of Web pages. Our work provides a large dataset of 49,438 Web pages, composed of webshots, along with qualitative and quantitative attributes. This dataset covers all the countries in the world and a wide range of topics, such as art, entertainment, economics, business, education, government, news, media, science, and the environment, addressing different cultural characteristics and varied design preferences. We use this dataset to develop three Web Intelligence applications: knowledge extraction on Web design using statistical analysis, recognition of error Web pages using a customized convolutional neural network (CNN) to eliminate invalid pages, and Web categorization based solely on screenshots using a CNN with transfer learning to assist search engines, indexers, and Web directories.

1. Introduction

The modern world relies on the Internet, with many human activities (commerce, education, entertainment, social interaction, etc.) having digital applications supported by this platform. It has allowed the development of such activities despite the paralysis caused by the recent pandemic. The Internet and the Web are closely related concepts, where the first term refers to the large network of networks (the infrastructure), while the second refers to the content, consisting of *Web sites*, which are a collection of interlinked *Web pages* on a specific topic [1]. Since its invention in the 1990s, the *World Wide Web*, or simply the Web, has revolutionized access to large amounts of data and information [2, 3]. Factors such as ease of use, user-friendly interface, popularity, and increased connectivity have made the Web an everyday tool for individuals and organizations in all areas [4]. The Web is a constantly evolving global communication

platform. The volume of information available is enormous and is growing rapidly, becoming more complex, and covering all topics. The effective management of such a quantity and variety of information is an increasingly difficult task for traditional techniques. For example, organizing valid content and filtering invalid content (purifying the Internet) are challenges facing the current Web [5]. The immense presence on the Internet of error Web pages (e.g., under construction, maintenance, domain offer, suspended account, page not found, browser incompatibility, virus, phishing, or service failure), which continue to be indexed and returned by search engines, affecting webmasters and users in general. Once the invalid pages are filtered, the valid content must be organized. To this end, classification is a basic technique, but doing this manually is impractical, with automatic classification of Web pages being the recommended method. This is usually carried out by analyzing both textual content and underlying HTML

code. However, modern Web pages that include multimedia objects, video streaming, and picture sharing are making the extraction of information increasingly complicated [6]. Hence, there is a need to improve the mechanisms to address such problems, which has attracted scientific interest and created new areas of research and development. Web intelligence (WI) is a relatively recent field that fundamentally seeks to improve Web-based services and extract knowledge by using artificial intelligence [6]. In this work, we present applications in the context of WI that efficiently deal with cleaning and organizing Web content, as well as extracting knowledge from Web data. It is worth mentioning that a large amount of Web-related data is an essential resource for WI and its applications. In addition to Web usage data, such as user profiles and preferences, interactions with Web sites and browsing habits [7], data on the structure and content of Web pages are necessary for our purposes. Currently, Web-related datasets only include screenshots, URLs, or images extracted from Web pages. There is no large, quality dataset that integrates the attributes and visual aspect of Web pages. Therefore, the first aim was to create such a dataset, which we collected automatically from Google and a Web directory by leveraging various computational tools, obtaining a total of 49,438 Web pages from all the countries in the world and classified into the following topics: arts and entertainment, business and economy, education, government, news and media, and science and environment. The dataset combines the attributes (structure data) and the visual aspect (content data) of a Web page. The structure data allow us to discover patterns on rules of thumb, trends, and guidelines in current Web design. In this way, a Web of data becomes a Web of knowledge to support beginners and experts, statistically processing and analyzing the qualitative and quantitative attributes of our dataset. Furthermore, the content data were used to develop artificial intelligence applications, based on deep learning, to clean and organize Web content. We implemented automatic recognition of error Web pages and categorization of valid Web pages, both applications based exclusively on screenshots. To this end, we used convolutional neural networks (CNNs), which are state-of-the-art tools in the field of computer vision. Consequently, our work makes the following contributions: (a) a freely available extensive Web page structure and content dataset; (b) a workflow supported by computer tools to automate the process of collecting, organizing, and debugging screenshots and attributes of Web pages, a methodology that can be adapted to other problems where the acquisition of large amounts of data is needed; and (c) WI applications to filter and categorize Web pages using only screenshots, avoiding the analysis of the HTML code and other new technologies. In this regard, it may be useful to save time and cost in information retrieval systems such as search engines (e.g., Google and Bing), classifiers, recommendation systems, Web directories, and crawlers [8, 9], and optimization of Internet resources.

The rest of the work is structured as follows. First we review Web page datasets in Section 2. This is followed by a detailed description of our dataset and the process designed for its creation in Section 3. In Section 4, we then

present Web intelligence applications' three use cases: a statistical analysis of the attributes of Web pages, the implementation of automatic recognition of error Web pages, and Web categorization, based on screenshots and CNNs. Section 5 details our conclusions and future work.

2. Related Work

The availability of a large amount of data underlies the current need for research and development in areas related to the Web. Our review of the literature and existing datasets is summarized in Table 1. For ease of comparison, we have divided them into two groups according to size: small (less than 1000 instances) and large. First, Boer et al. [1] use a tiny dataset, only visual, collected for categorization within four classes (news, hotels, conferences, and celebrities). These categories are considerably different from each other; so, the categorization problem is less complicated. There is no link to download the screenshots. López-Sánchez et al. [10, 11] have datasets with more Web pages, including their respective images and links (URLs). However, these images are not screenshots but elements of the Web page. The URL is used to download the images from the HTML code and analyze them for categorization. Although there are more categories than in the previous work, they still cover very different topics. In neither case is a download link available. Of the smaller ones, Reinecke and Gajos [12] propose the most significant dataset, which might be a useful resource for small-scale research and development works. It covers several countries around the world and various topics and is available for download. However, it is strictly visual and insufficient for current needs, and its purpose is more oriented towards aesthetics analysis and classification. The Computer Incident Response Center Luxembourg (CIRCL) is a government initiative created to respond to computer security threats and incidents. CIRCL [13] offers a dataset of more than 400 screenshots of verified or potential phishing Web sites. Furthermore, an extensive dataset with more than 37000 images is available [14], corresponding to screenshots of Web sites belonging to the dark-Web, the problematic facet of the Web associated with cybercrime, hate, and extremism [15]. Both datasets can be easily downloaded, although, because the images represent fakes or hidden Web pages, these datasets would have a limited application. ImageNet [16], the most popular of the image databases, includes millions of images organized according to the WordNet hierarchy (A large lexical database of English nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), <https://wordnet.princeton.edu>). The Web sites section has 1840 screenshots from different countries and languages without categorization. Some screenshots appear cropped, and download requires registration and authorization. The dataset created by the University of Alicante [17] compiles 8950 screenshots of Web pages for analysis and evaluation of the quality of Web design. Half of the images come from the Awwwards site (<https://www.awwwards.com>) and are labeled with "good design," whereas the other half are extracted from yellow pages, labeled with "bad design." This dataset serves the

TABLE 1: Main characteristics of the datasets reviewed.

Owner and year	Size	Topic	Data type	Purpose
De Boer et al. 2011	Small: 60 screenshots	News, hotels, conferences, and celebrities	Images database	Aesthetics and thematic classification with machine learning
Reinecke et al. 2014	Small: 430 screenshots	Generic	Images database	Aesthetics classification
López et al. 2017	Small: 280 web pages	Food, animals, fashion, nature, home, and vehicles	URL and images extracted from HTML	Thematic classification with machine learning
López et al. 2019	Small: 365 web pages	Food, vehicles, animals, fashion, home design, and landscape	URL and images extracted from HTML	Thematic classification with machine learning
CIRCL, 2019	Small: 460 screenshots	Phishing	Images database	Analysis of security events
ImageNet, 2009	Large: 1840 screenshots	Generic	Images database	Resource for image and vision research field
Nordhoff et al. 2018	Large: 80901 screenshots	Generic	URL, metrics and images	Aesthetics and Web design
CIRCL, 2019	Large: 37500 screenshots	Onion Website (hidden Web, no indexed)	Images database	Analysis of security events
University of Alicante, 2019	Large: 8950 labeled screenshots	Good and bad design	Labeled images dataset	Aesthetics Web categorization

academic work of the institution. Meanwhile, Nordhoff et al. [18] stands out because it covers a larger number of Web pages. However, these come from only 44 countries, the parameters are purely aesthetics, and the image download is not direct. In contrast, our dataset takes into account all the countries in the world, includes attributes related to Web page structure, is general-purpose, and available for download. We created from scratch an extensive and available dataset, which incorporates the visual representation of the Web page, complemented with qualitative and quantitative attributes extracted from the underlying HTML source code, so that a Web page is better characterized. Because manual data collection is a complex task and requires a great deal of time and human effort, we automated the process by writing several programs in Python and R programming languages.

3. Materials and Methods

Current Web page datasets are limited to screenshots and sometimes URLs. Creation of more sophisticated and larger datasets is needed by research on Web-related problems. In this section, we present a methodology for combining visual, textual, and numerical elements into a single dataset on Web pages. The workflow is represented in Figure 1, which considers the example of our large dataset that integrates quantitative and qualitative attributes, along the visual appearance of Web pages. The resulting dataset is the main resource for implementing three applications in the context of Web Intelligence, which we explain in the next section.

3.1. Web Dataset Design. The first step is to define the elements that compose the dataset according to the proposed aim. Our interest is focused on the structure and content of a Web page, so we selected a series of qualitative and quantitative attributes for the structure, and a webshot for the visual aspect. In this way, we created a single mixed dataset, which is designed to combine different data types (visual, textual, and numerical) and is more extensive and descriptive than current Web page datasets. Table 2 shows the list of the elements of our dataset, which are detailed as follows.

A webshot is a digital image of the entire Web page, unlike a screenshot, which may appear cropped because its dimensions exceed the viewing device, forcing the user to scroll. The name given to the webshot is a key element that follows a convention to identify the source, category, and country of the Web page. It is also the link between the webshot and the qualitative and quantitative attributes. A URL (uniform resource locator) is the address of the Web page together with the recovery mechanism (http/https). It is placed in the address bar of browsers, which are the programs that display the content to the user. We collected URLs from around the world to cover different cultural characteristics and preferences, so that the dataset includes attributes related to geographic locations, such as country and continent. The Web pages collected belong to the following categories: arts and entertainment, business and

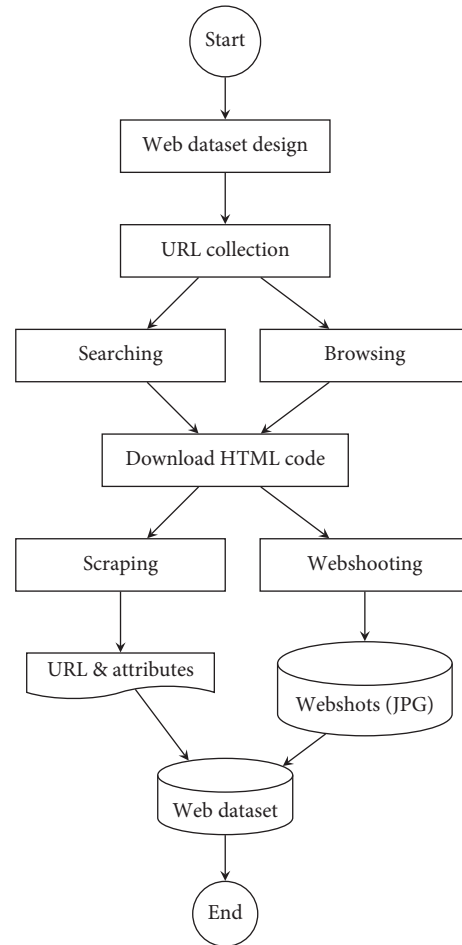


FIGURE 1: Methodology created to produce the Web page dataset.

TABLE 2: Structure of the dataset.

Elements	Type	Description
Webshot	Visual	Web page entire screenshot in JPG format
Name	Text	Identification given to the webshot
URL	Text	Link to locate and display a Web page
Country	Qualitative	National origin of Web page
Continent		Region grouping countries
Category		Main topic of Web page
Time	Quantitative	Web page's source code download time
Bytes		Size in bytes of Web page's source code
Images		Number of images from Web page
Script_files		Number of executable files of Web page
CSS_files		Number of files to layout a Web page
Tables		Number of table tags in the source code
iframes		Number of iframe tags in the source code
Style_tags		Number of style tags in the source code
Img_bytes		Webshot size in bytes
Img_width		Webshot width in pixels
Img_height	Webshot height in pixels	

economy, education, government, news and media, and science and environment. We considered these categories since they are part of the Web directory used here and explained in the section on Browsing. We added the following quantitative parameters, which provide an overview of the structure and quality of a Web page: download time, because users want to wait as little as possible to view a Web page [19], which means reducing the source code download time; size, the larger the size in bytes, the slower are the download and display of the Web page; images, since they will increase the download time. A Web page is not necessarily more attractive because it has more images, a balance between all types of information is recommendable [20]; scripts are external files to provide the Web page with more complex functionality. It is advisable to reduce their quantity because they increase the network traffic and download time; CSS files are style files that cause an extra load and delay the display of the Web page, ideally there should be one; tables are often used to structure the content of a Web page, but this is discouraged due to appropriate elements such as “div” tags; iFrames insert a Web page inside another one, which is not currently good practice; style tags are not recommended since there are CSS files. Finally, we have the image size in bytes, as well as the dimensions (width and height) in pixels of each webshot. Figure 2 presents a small sample of the dataset showing a case of each category.

3.2. URL Collection. Once the dataset had been structured, the next step was to collect URLs worldwide related to the given categories. Each URL was then used to download the HTML code, extract quantitative and qualitative attributes by scraping, and take a screenshot of the entire Web page (webshot). To obtain a larger number of URLs, we used two ways of finding information on the Web: searching and browsing. Searching requires the user to translate a need for information into queries, whereas browsing is a basic, natural human activity, occurring in an information environment where information objects are visible and organized [21]. Next, we describe how both techniques allowed us to collect URLs and, based on these, extract attributes and capture webshots, all through Python and R scripts.

3.2.1. Searching. The Google search engine asks for words or phrases related to the topic of interest. To avoid the repetitive task of typing the query into the Google search page and manually retrieving the response URLs, we automatized the process through a Python script (<https://osf.io/k6yrx>), where

- (1) The country name and its Internet code are extracted iteratively from a plain text file (<https://osf.io/yrmx8>).
- (2) The search query has the following structure: “site: “+ *countrycode* +” business OR economy OR marketing OR computers OR internet OR construction OR financial OR industry OR shopping OR restaurant“ + ” ext :html”

OR, *site* and *ext* are operators or reserved words that can be used in query phrases within the Google search engine. The OR operator concatenates several

search words related to the category. The “site” operator specifies the geographic top-level Internet domain assigned for each country, e.g., “es” for Spain. The “ext:html” operator produces results exclusively with that file extension.

- (3) The request returns the Google results page with the first 100 links, which is used to achieve an approximately uniform distribution of Web pages according to country and category.
- (4) Web page links are extracted by automatically scanning the source code of the results page (*scraping*), generating a text file that contains the URLs and their attributes: country, continent, and category.
- (5) For the rest of the categories, the queries are as follows:

“site: “+ *countrycode* +” arts OR entertainment OR dance OR museums OR theatre OR literature OR artists OR galleries“ + ” ext :html”

“site: “+ *countrycode* +” education OR academy OR university OR college OR school“ + ” ext :html”

“site: “+ *countrycode* +” government OR military OR presidency“ + ” ext :html”

“site: “+ *countrycode* +” news OR media OR magazine OR radio OR television OR newspaper“ + ” ext :html”

“site: “+ *countrycode* +” science OR environment OR archaeology“ + ” ext :html”

3.2.2. Browsing. This technique uses a Web Directory, a specialized Web site consisting of a catalog of links to other Web sites. Building, maintaining, and organizing by category and subcategory is done by human experts, unlike search engines, which do so automatically. To include a URL, the specialists perform a review, analysis, and evaluation process to verify the requirements determined by the Web Directory. A few Web directories have survived the popularity of search engines like Google. We can highlight Best of the Web (BOTW) (Figure 3), one of the most widely recognized due to its quality, global reach, a wide range of categories and subcategories, level of traffic (visits per month), reliability, number of links, and demanding requirements.

Instead of a query or search phrase, it is necessary to know the hierarchical structure of the directories and subdirectories until the URL of interest. We took advantage of the organization by country and category defined by BOTW, e.g., for Greece:

<https://botw.org/top/Regional/Europe/Greece/Arts-and-Entertainment/>

<https://botw.org/top/Regional/Europe/Greece/Business-and-Economy/>

<https://botw.org/top/Regional/Europe/Greece/Education/>

<https://botw.org/top/Regional/Europe/Greece/Government/>

<https://botw.org/top/Regional/Europe/Greece/News-and-Media/>







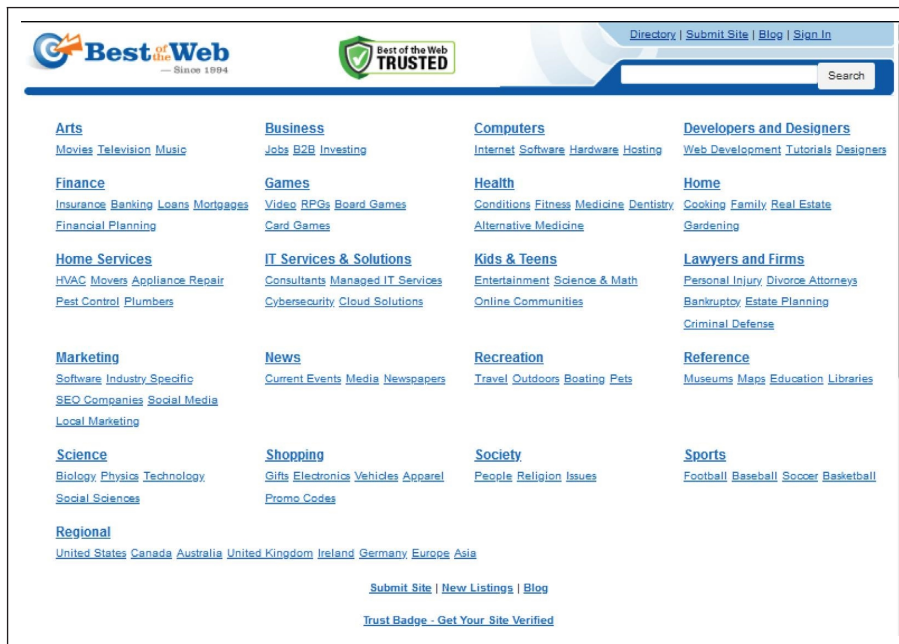
	<table border="1"> <tbody> <tr><td>IMG</td><td>B1Norway_343.jpg</td></tr> <tr><td>URL</td><td>http://www.fotoserch.no/</td></tr> <tr><td>CATEGORY</td><td>Arts and Entertainment</td></tr> <tr><td>COUNTRY</td><td>Norway</td></tr> <tr><td>CONTINENT</td><td>Europe</td></tr> <tr><td>Time</td><td>0.075134</td></tr> <tr><td>Bytes</td><td>36055</td></tr> <tr><td>Images</td><td>12</td></tr> <tr><td>Script_files</td><td>5</td></tr> <tr><td>CSS_files</td><td>47</td></tr> <tr><td>Tables</td><td>0</td></tr> <tr><td>iFrames</td><td>0</td></tr> <tr><td>Style_tags</td><td>1</td></tr> <tr><td>Img_bytes</td><td>170175</td></tr> <tr><td>Img_width</td><td>992</td></tr> <tr><td>Img_height</td><td>1464</td></tr> </tbody> </table>	IMG	B1Norway_343.jpg	URL	http://www.fotoserch.no/	CATEGORY	Arts and Entertainment	COUNTRY	Norway	CONTINENT	Europe	Time	0.075134	Bytes	36055	Images	12	Script_files	5	CSS_files	47	Tables	0	iFrames	0	Style_tags	1	Img_bytes	170175	Img_width	992	Img_height	1464		<table border="1"> <tbody> <tr><td>IMG</td><td>B3Uganda_252.jpg</td></tr> <tr><td>URL</td><td>http://www.excelconstruction.org/</td></tr> <tr><td>CATEGORY</td><td>Business and Economy</td></tr> <tr><td>COUNTRY</td><td>Uganda</td></tr> <tr><td>CONTINENT</td><td>Africa</td></tr> <tr><td>Time</td><td>0.004368</td></tr> <tr><td>Bytes</td><td>25549</td></tr> <tr><td>Images</td><td>2</td></tr> <tr><td>Script_files</td><td>22</td></tr> <tr><td>CSS_files</td><td>29</td></tr> <tr><td>Tables</td><td>0</td></tr> <tr><td>iFrames</td><td>0</td></tr> <tr><td>Style_tags</td><td>0</td></tr> <tr><td>Img_bytes</td><td>143092</td></tr> <tr><td>Img_width</td><td>992</td></tr> <tr><td>Img_height</td><td>951</td></tr> </tbody> </table>	IMG	B3Uganda_252.jpg	URL	http://www.excelconstruction.org/	CATEGORY	Business and Economy	COUNTRY	Uganda	CONTINENT	Africa	Time	0.004368	Bytes	25549	Images	2	Script_files	22	CSS_files	29	Tables	0	iFrames	0	Style_tags	0	Img_bytes	143092	Img_width	992	Img_height	951		<table border="1"> <tbody> <tr><td>IMG</td><td>B3Pakistan_38.jpg</td></tr> <tr><td>URL</td><td>http://va.edu.pk/</td></tr> <tr><td>CATEGORY</td><td>Education</td></tr> <tr><td>COUNTRY</td><td>Pakistan</td></tr> <tr><td>CONTINENT</td><td>Asia</td></tr> <tr><td>Time</td><td>0.004348</td></tr> <tr><td>Bytes</td><td>21413</td></tr> <tr><td>Images</td><td>1</td></tr> <tr><td>Script_files</td><td>1</td></tr> <tr><td>CSS_files</td><td>2</td></tr> <tr><td>Tables</td><td>20</td></tr> <tr><td>iFrames</td><td>0</td></tr> <tr><td>Style_tags</td><td>1</td></tr> <tr><td>Img_bytes</td><td>121227</td></tr> <tr><td>Img_width</td><td>992</td></tr> <tr><td>Img_height</td><td>1007</td></tr> </tbody> </table>	IMG	B3Pakistan_38.jpg	URL	http://va.edu.pk/	CATEGORY	Education	COUNTRY	Pakistan	CONTINENT	Asia	Time	0.004348	Bytes	21413	Images	1	Script_files	1	CSS_files	2	Tables	20	iFrames	0	Style_tags	1	Img_bytes	121227	Img_width	992	Img_height	1007
IMG	B1Norway_343.jpg																																																																																																				
URL	http://www.fotoserch.no/																																																																																																				
CATEGORY	Arts and Entertainment																																																																																																				
COUNTRY	Norway																																																																																																				
CONTINENT	Europe																																																																																																				
Time	0.075134																																																																																																				
Bytes	36055																																																																																																				
Images	12																																																																																																				
Script_files	5																																																																																																				
CSS_files	47																																																																																																				
Tables	0																																																																																																				
iFrames	0																																																																																																				
Style_tags	1																																																																																																				
Img_bytes	170175																																																																																																				
Img_width	992																																																																																																				
Img_height	1464																																																																																																				
IMG	B3Uganda_252.jpg																																																																																																				
URL	http://www.excelconstruction.org/																																																																																																				
CATEGORY	Business and Economy																																																																																																				
COUNTRY	Uganda																																																																																																				
CONTINENT	Africa																																																																																																				
Time	0.004368																																																																																																				
Bytes	25549																																																																																																				
Images	2																																																																																																				
Script_files	22																																																																																																				
CSS_files	29																																																																																																				
Tables	0																																																																																																				
iFrames	0																																																																																																				
Style_tags	0																																																																																																				
Img_bytes	143092																																																																																																				
Img_width	992																																																																																																				
Img_height	951																																																																																																				
IMG	B3Pakistan_38.jpg																																																																																																				
URL	http://va.edu.pk/																																																																																																				
CATEGORY	Education																																																																																																				
COUNTRY	Pakistan																																																																																																				
CONTINENT	Asia																																																																																																				
Time	0.004348																																																																																																				
Bytes	21413																																																																																																				
Images	1																																																																																																				
Script_files	1																																																																																																				
CSS_files	2																																																																																																				
Tables	20																																																																																																				
iFrames	0																																																																																																				
Style_tags	1																																																																																																				
Img_bytes	121227																																																																																																				
Img_width	992																																																																																																				
Img_height	1007																																																																																																				
	<table border="1"> <tbody> <tr><td>IMG</td><td>B6Armenia_220.jpg</td></tr> <tr><td>URL</td><td>http://www.parliament.am/lang-eng</td></tr> <tr><td>CATEGORY</td><td>Government</td></tr> <tr><td>COUNTRY</td><td>Armenia</td></tr> <tr><td>CONTINENT</td><td>Asia</td></tr> <tr><td>Time</td><td>0.005412</td></tr> <tr><td>Bytes</td><td>23423</td></tr> <tr><td>Images</td><td>23</td></tr> <tr><td>Script_files</td><td>2</td></tr> <tr><td>CSS_files</td><td>2</td></tr> <tr><td>Tables</td><td>2</td></tr> <tr><td>iFrames</td><td>0</td></tr> <tr><td>Style_tags</td><td>0</td></tr> <tr><td>Img_bytes</td><td>296180</td></tr> <tr><td>Img_width</td><td>1000</td></tr> <tr><td>Img_height</td><td>1737</td></tr> </tbody> </table>	IMG	B6Armenia_220.jpg	URL	http://www.parliament.am/lang-eng	CATEGORY	Government	COUNTRY	Armenia	CONTINENT	Asia	Time	0.005412	Bytes	23423	Images	23	Script_files	2	CSS_files	2	Tables	2	iFrames	0	Style_tags	0	Img_bytes	296180	Img_width	1000	Img_height	1737		<table border="1"> <tbody> <tr><td>IMG</td><td>B3Barbados_268.jpg</td></tr> <tr><td>URL</td><td>http://www.cbc.bb/</td></tr> <tr><td>CATEGORY</td><td>News and Media</td></tr> <tr><td>COUNTRY</td><td>Barbados</td></tr> <tr><td>CONTINENT</td><td>Caribbean</td></tr> <tr><td>Time</td><td>0.004801</td></tr> <tr><td>Bytes</td><td>147585</td></tr> <tr><td>Images</td><td>6</td></tr> <tr><td>Script_files</td><td>15</td></tr> <tr><td>CSS_files</td><td>19</td></tr> <tr><td>Tables</td><td>0</td></tr> <tr><td>iFrames</td><td>0</td></tr> <tr><td>Style_tags</td><td>2</td></tr> <tr><td>Img_bytes</td><td>592533</td></tr> <tr><td>Img_width</td><td>992</td></tr> <tr><td>Img_height</td><td>4484</td></tr> </tbody> </table>	IMG	B3Barbados_268.jpg	URL	http://www.cbc.bb/	CATEGORY	News and Media	COUNTRY	Barbados	CONTINENT	Caribbean	Time	0.004801	Bytes	147585	Images	6	Script_files	15	CSS_files	19	Tables	0	iFrames	0	Style_tags	2	Img_bytes	592533	Img_width	992	Img_height	4484		<table border="1"> <tbody> <tr><td>IMG</td><td>B6Australia_432.jpg</td></tr> <tr><td>URL</td><td>http://australangeofdinosaurs.com/</td></tr> <tr><td>CATEGORY</td><td>Science and Environment</td></tr> <tr><td>COUNTRY</td><td>Australia</td></tr> <tr><td>CONTINENT</td><td>Oceania</td></tr> <tr><td>Time</td><td>0.045729</td></tr> <tr><td>Bytes</td><td>33356</td></tr> <tr><td>Images</td><td>11</td></tr> <tr><td>Script_files</td><td>18</td></tr> <tr><td>CSS_files</td><td>10</td></tr> <tr><td>Tables</td><td>0</td></tr> <tr><td>iFrames</td><td>1</td></tr> <tr><td>Style_tags</td><td>1</td></tr> <tr><td>Img_bytes</td><td>291150</td></tr> <tr><td>Img_width</td><td>1000</td></tr> <tr><td>Img_height</td><td>1764</td></tr> </tbody> </table>	IMG	B6Australia_432.jpg	URL	http://australangeofdinosaurs.com/	CATEGORY	Science and Environment	COUNTRY	Australia	CONTINENT	Oceania	Time	0.045729	Bytes	33356	Images	11	Script_files	18	CSS_files	10	Tables	0	iFrames	1	Style_tags	1	Img_bytes	291150	Img_width	1000	Img_height	1764
IMG	B6Armenia_220.jpg																																																																																																				
URL	http://www.parliament.am/lang-eng																																																																																																				
CATEGORY	Government																																																																																																				
COUNTRY	Armenia																																																																																																				
CONTINENT	Asia																																																																																																				
Time	0.005412																																																																																																				
Bytes	23423																																																																																																				
Images	23																																																																																																				
Script_files	2																																																																																																				
CSS_files	2																																																																																																				
Tables	2																																																																																																				
iFrames	0																																																																																																				
Style_tags	0																																																																																																				
Img_bytes	296180																																																																																																				
Img_width	1000																																																																																																				
Img_height	1737																																																																																																				
IMG	B3Barbados_268.jpg																																																																																																				
URL	http://www.cbc.bb/																																																																																																				
CATEGORY	News and Media																																																																																																				
COUNTRY	Barbados																																																																																																				
CONTINENT	Caribbean																																																																																																				
Time	0.004801																																																																																																				
Bytes	147585																																																																																																				
Images	6																																																																																																				
Script_files	15																																																																																																				
CSS_files	19																																																																																																				
Tables	0																																																																																																				
iFrames	0																																																																																																				
Style_tags	2																																																																																																				
Img_bytes	592533																																																																																																				
Img_width	992																																																																																																				
Img_height	4484																																																																																																				
IMG	B6Australia_432.jpg																																																																																																				
URL	http://australangeofdinosaurs.com/																																																																																																				
CATEGORY	Science and Environment																																																																																																				
COUNTRY	Australia																																																																																																				
CONTINENT	Oceania																																																																																																				
Time	0.045729																																																																																																				
Bytes	33356																																																																																																				
Images	11																																																																																																				
Script_files	18																																																																																																				
CSS_files	10																																																																																																				
Tables	0																																																																																																				
iFrames	1																																																																																																				
Style_tags	1																																																																																																				
Img_bytes	291150																																																																																																				
Img_width	1000																																																																																																				
Img_height	1764																																																																																																				

FIGURE 2: A sample of the dataset, including one example of each category.

FIGURE 3: BOTW Web Directory (<https://botw.org/>).

<https://botw.org/top/Regional/Europe/Greece/Science-and-Environment/>

We collected the URLs published within each category through a Python script (<https://osf.io/73sc2>) that

- (1) Iteratively reads the name of each country from a flat text file (<https://osf.io/ve986>)
- (2) Sets the path corresponding to the category, which will always have the same structure, where only the country's name changes: "https://botw.org/top/Regional/" + *countryname* + "/Science_and_Environment/"
- (3) A connection to the Web address formed is realized, which obtains the source code of the results page to extract the URLs of each category

- (4) The links are stored in a text file (<https://osf.io/hjwgm>), imported into a spreadsheet where a filter is applied to select only those links belonging to a country in particular

3.3. *Attribute Collection by Scraping.* After collecting and storing the URLs from the two previously described sources, we implemented a Python script (<https://osf.io/de78f>) to (a) sequentially read the URL links stored in the text file (<https://osf.io/gk3p2>); (b) make a connection through the browser to each of these links; and (c) download and analyze the source code of the Web page, obtaining the attributes specified in the dataset: download time in seconds, total size in bytes, number of images, script files, CSS files, tables, iFrames tags, and style tags. This script includes the Web scraping library

known as *Beautiful Soup*, which allows us to define and extract the attributes from the HTML code of each Web page.

3.4. Webshot Collection. We used *Webshot* and *PhantomJS* packages to create an R script (<https://osf.io/6pmyb>) that reads each URL from text files generated in the Searching and Browsing sections, takes a snapshot of the entire Web page, and saves it as a JPG image. The name assigned to the image is an identifier to know the source, category, and country to which the Web page belongs; e.g., *B2Netherlands_791.jpg* indicates a webshot of a page obtained through the technique of browsing, belonging to the category number two (business and economy) and which comes from the Netherlands. The number after the underscore only establishes a sequential order. Note that this identifier is the key to attach the webshot to its respective qualitative and quantitative attributes. In this way, it was possible to link two different storage media, i.e., a data sheet and an image folder. In addition, this script includes functions to obtain both webshot size in bytes and webshot dimensions (width and height) in pixels, in order to better characterize a Web page. All the images and the data sheet are available for viewing and downloading (https://osf.io/7ghd2/?view_only=0bf99589809e4e88b0aa0602c8060b46).

4. Web Intelligence Applications

4.1. Use Case 1: Knowledge from Web Data. The large amount of data collected can provide useful information, which can be converted into knowledge. Through statistical analysis of the qualitative and quantitative attributes of Web pages, we can identify patterns about how they are structured, which is fundamental in Web design. We distinguish between the two sources of Web data: browsing, which is stricter, and searching, which is virtually unrestricted. R was used to compare both sources, and *outliers* were excluded in the calculation of statistical indicators and the creation of graphs due to the great heterogeneity of the variables cited in Table 2. These values differ greatly to those considered common and may cause distortions in mathematical and visual analysis. By using the well-known rule “1.5 times the *Interquartile Range*,” outliers can be identified and omitted. For this reason, the number of values for each of the variables may differ.

4.1.1. Qualitative Attributes. Although we attempted to obtain a uniformly distributed set of URLs with respect to the categories, the errors cited in the section on Automatic Recognition of Error Web Pages generated the results shown in Figure 4(a). In searching, there is less imbalance, in contrast to browsing, where Web pages related to business, economy, and government predominate, possibly due to a greater need for dissemination and economic capacity to register their Web pages in a paid service. Most of the Web pages are geographically located in Europe and Asia, for both browsing and searching, with these continents accounting for a larger number of countries. Moreover, their economic potential could explain why they are at the top (Figure 4(b)).

4.1.2. Quantitative Attributes. The variability of searching is more evident given the number of characters in a URL (Figure 5(a)), resulting in a wider range (the difference between the minimum and maximum) and a higher mean and standard deviation. In both graphs, the values accumulate more at the bottom of the variable and are less frequent at the top, and so, there is a tail to the right. This behavior is desirable when reading or typing a URL in a browser and indicates there are not too many levels to reach a particular Web page. The download time of the source code of the Web pages shown in Figure 5(b) behave similarly for browsing and searching. In both cases, although there is considerable variability due to the width of the range and standard deviation, the values are around 20 milliseconds and mostly low, which is a benefit for the user who wants to view the Web page in the shortest time possible. According to the errors cited in the debugging section, 254 of 3182 URLs (7.98%) belonging to browsing were unavailable, while in searching, 4079 of 46256 (8.82%), were not accessible to download the source code, and hence, the extraction of the quantitative parameters was not possible. The behavior of the size in bytes (Figure 5(c)) is almost identical in browsing and searching, where the average of the Web pages is approximately 50 KB. Both the variability and the tails of the distributions are practically the same, favoring a quick view of the Web page, which is in direct relation to a small size. Nonetheless, there are still some Web pages with a considerable size, which may be due to graphic elements or external objects linked to the page. In Figure 5(d), almost 60% of Web pages include no images within their source code. It seems that the pages present only textual information. However, they might include images through CSS style files, which is a good practice [19]. Although the range of number of images is wide, the average is low, 2 or 3 images per Web page, so there is a tendency to use a few images within a Web page in order to make it lighter. Nonscripted Web pages exceed 50% in both cases. In this sense, scripts are add-on programs that provide additional functions to Web pages. However, their use may cause incompatibilities with browsers and make the page more complex and heavy. In Figure 5(e), the trend is to minimize the presence of scripts, 3 scripts per Web page on average. In Figure 5(f), approximately 50% of the Web pages do not use cascading style sheet (CSS) files, whose use is recommended as good practice in Web design. The ideal number would be one style file per Web page. The average for browsing and searching is 5 and 4, respectively, which is relatively close to the ideal number. The tendency is towards low values, although there are some cases with many style files, which hinders an agile display of the Web page. The graphs in Figure 5(g) have a very similar aspect. The majority of Web pages (about 85%) no longer use tables within the source code. While, in the past, tables were generally used to structure the content, this practice has been replaced by the “div” tag, achieving a more elegant and professional design. More than 85% of Web pages do not use iFrames. For browsing and searching, the bars in the graph are grouped on the left (Figure 5(h)). We can deduce that embedding another document in the current HTML document through

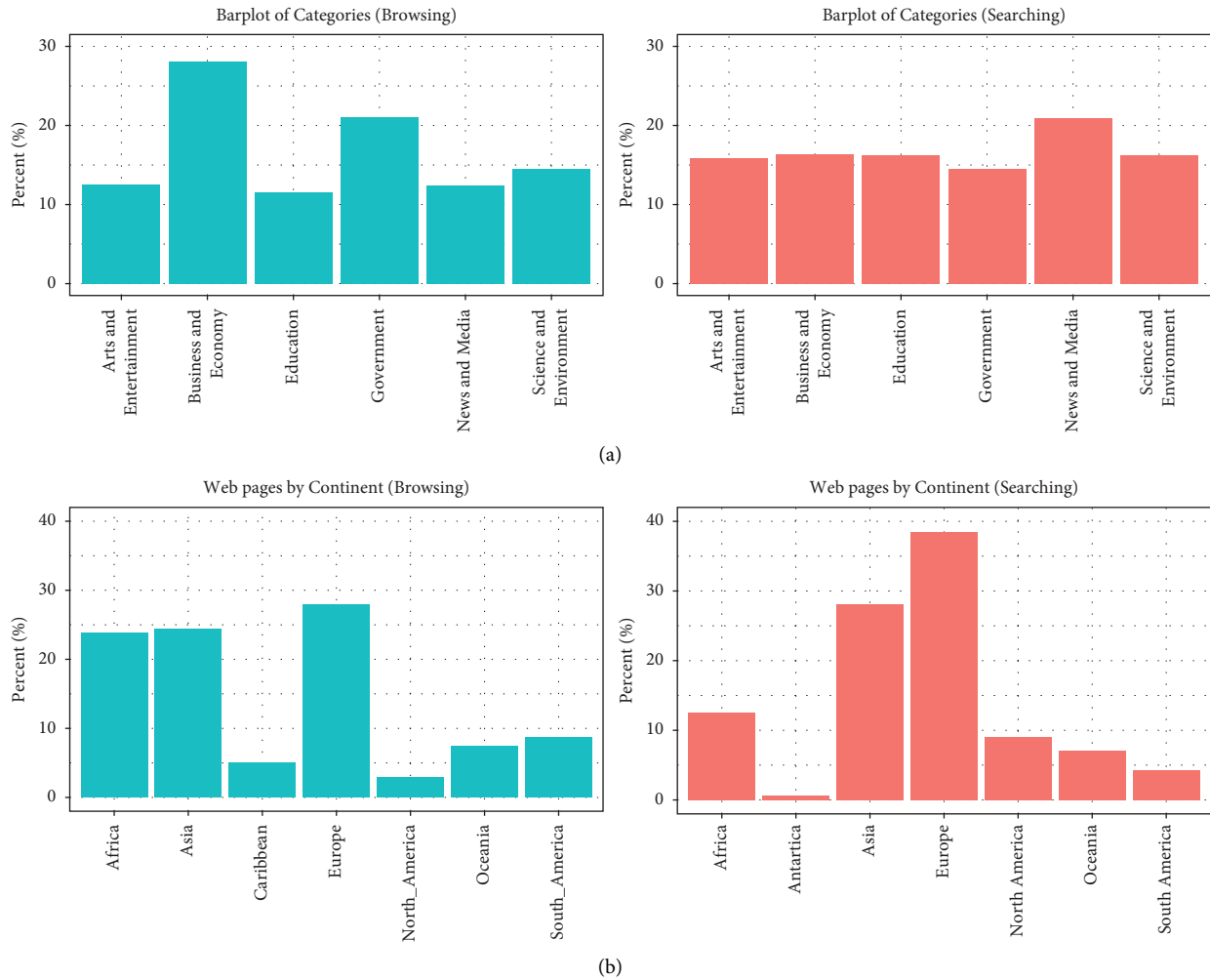
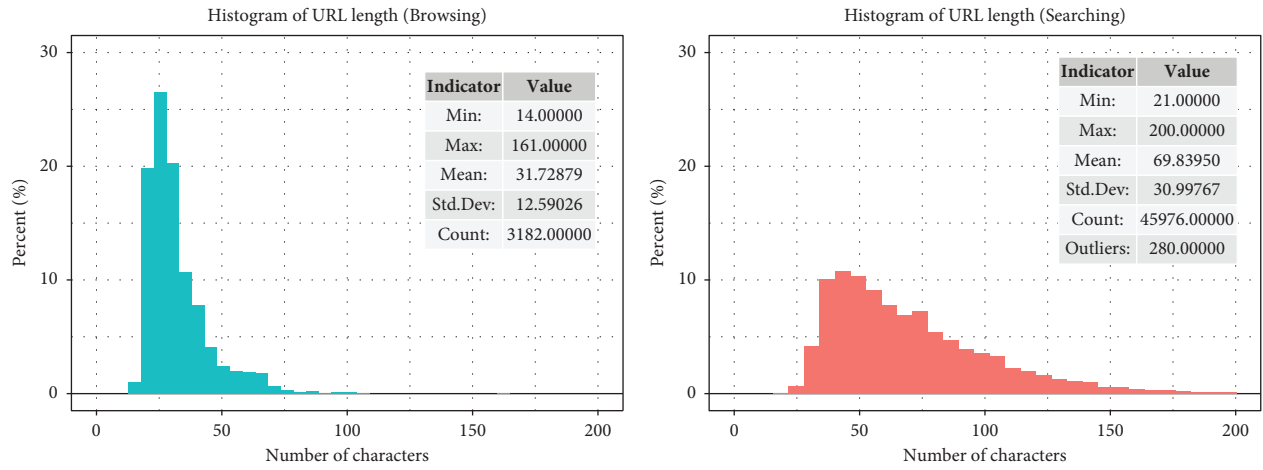


FIGURE 4: Distribution of the qualitative attributes about Web pages: (a) category and (b) continent.

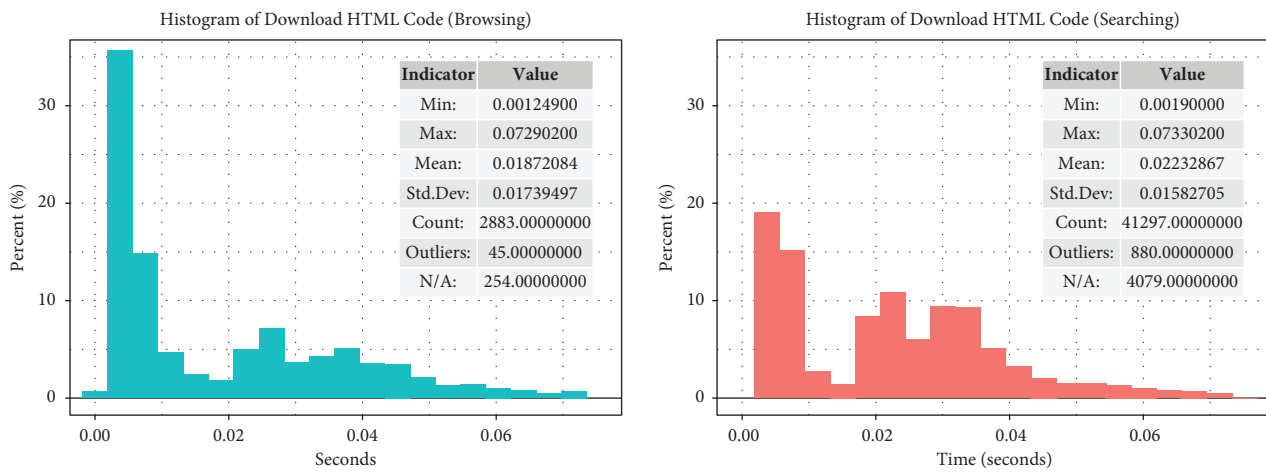
the “iFrame” tag is no longer a common practice, as there are now better options. More than half of the Web pages (close to 60%) no longer use “style” tags in their source code. Both graphs in Figure 5(i) have bars that decrease towards the right. The trend is to minimize the number of such tags, as it is more appropriate to use CSS files. Thus, the source code of a Web page does not overly extend. The webshot size is decisive in determining how much space our dataset will consume on a storage device. In Figure 5(j), the size fluctuates over a wide range of values, with an average of about 300 KB. Most of the values are concentrated in low sizes, but with a considerable presence of images with medium and high size. This behavior would require not only a large amount of space but a preprocessing of the images for machine learning and deep learning applications. The searching and browsing graphs are quite similar for the webshot width (Figure 5(k)). In both cases, the first bar, which shows the minimum, significantly predominates, as the default screenshot sets a width of 992 pixels. The average value is very close to the minimum since almost all images

were captured with this default value (about 85%), although there are also images with a wider width, especially in searching with a maximum of almost 10000 pixels. In the case of the height variable (Figure 5(l)), the minimum value also prevails, albeit to a lesser degree (about 28%), which coincides with the default value set by the screenshot, i.e., 744 pixels. Unlike the previous case, there is a less unbalanced distribution of values, with a wider variability in which the highest accumulation occurs up to 5000 pixels, a considerable accumulation between 5000 and 10000 pixels, and finally, images with a height of up to almost 50000 pixels are obtained. Considering the width and height, most of the Web pages have a vertical layout. These parameters are closely related to the resolution or quality of the image. The more pixels there are, the greater is the resolution and quality of the images, although it does demand more storage space.

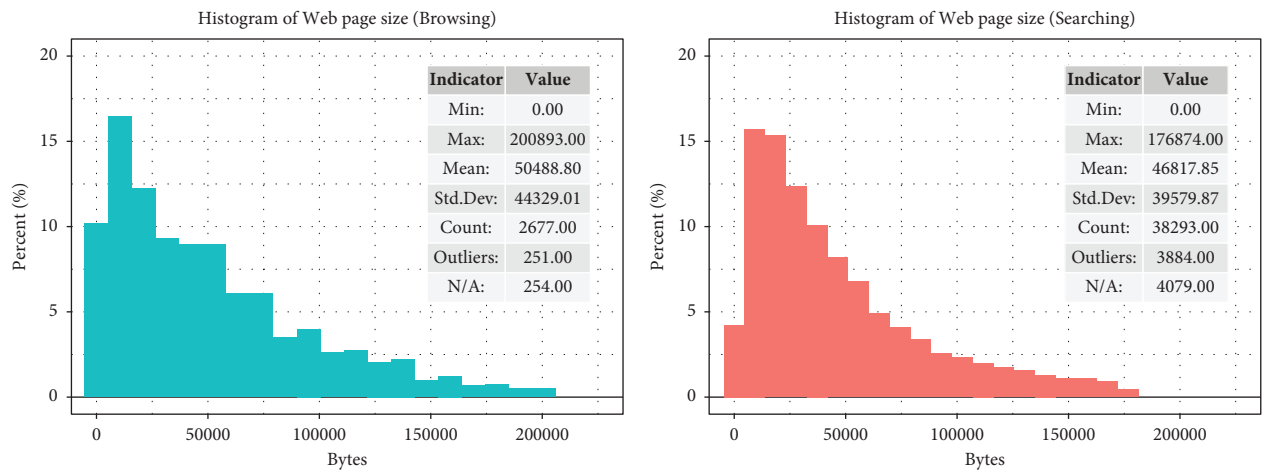
Finally, Table 3 summarizes the main statistical indicators for the quantitative parameters of the Web pages, for both the browsing and searching sets.



(a)

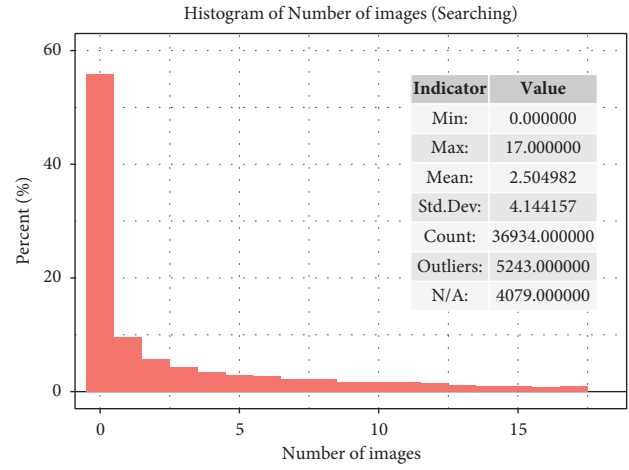
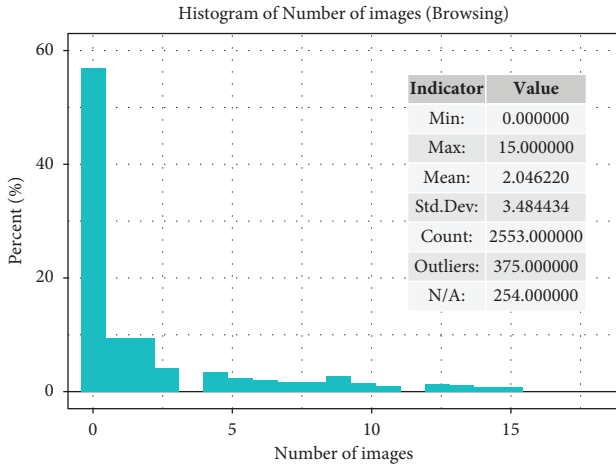


(b)

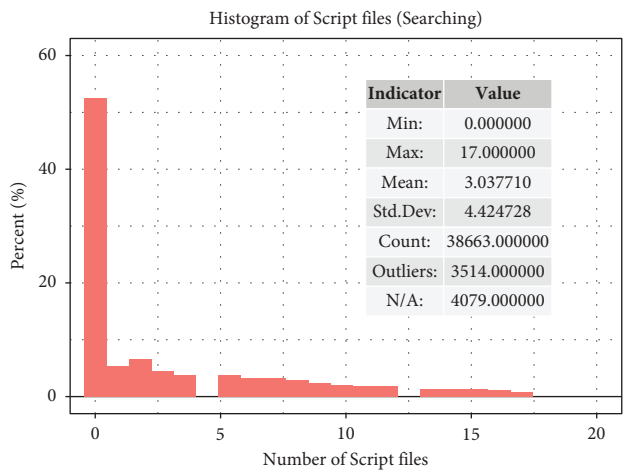
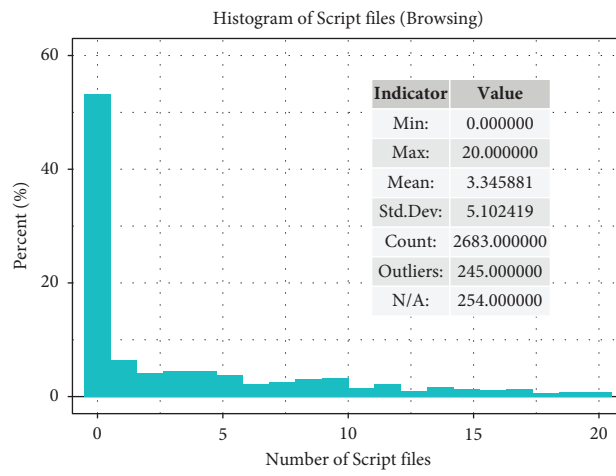


(c)

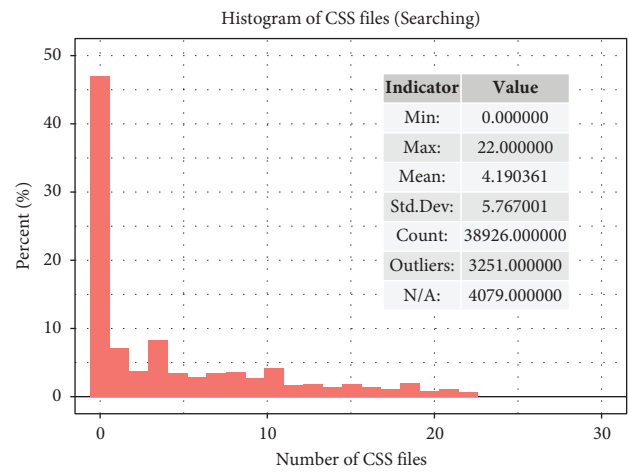
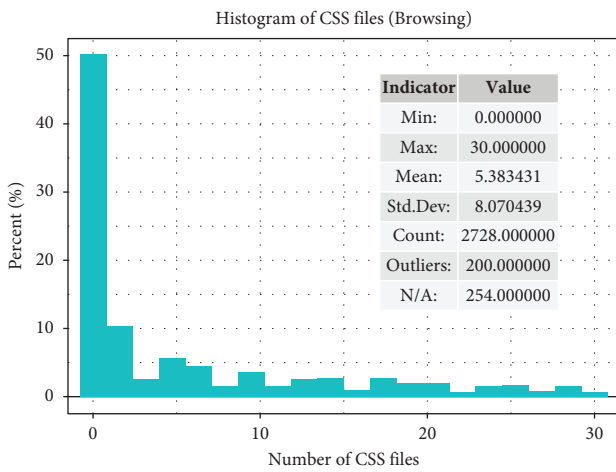
FIGURE 5: Continued.



(d)

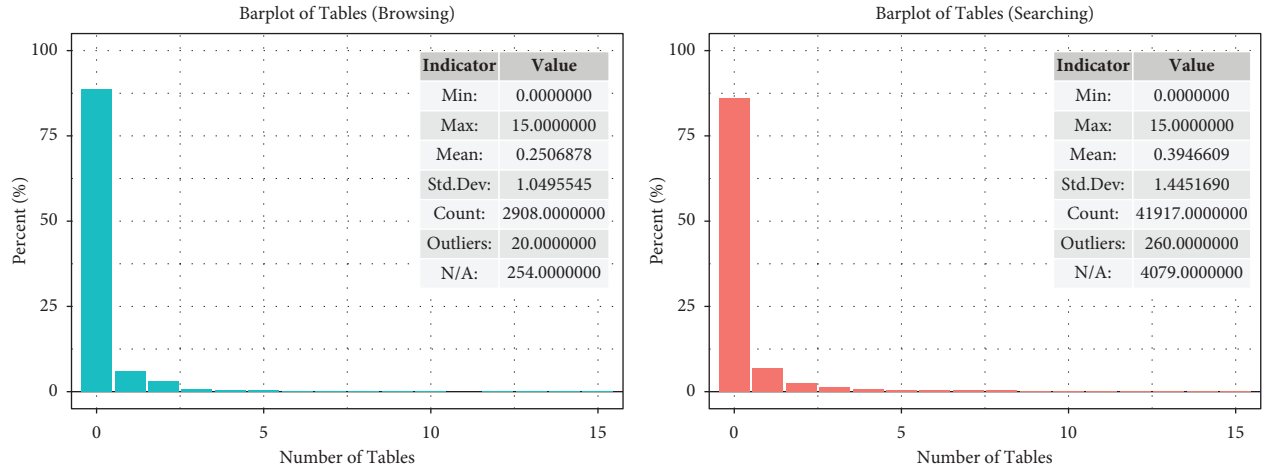


(e)

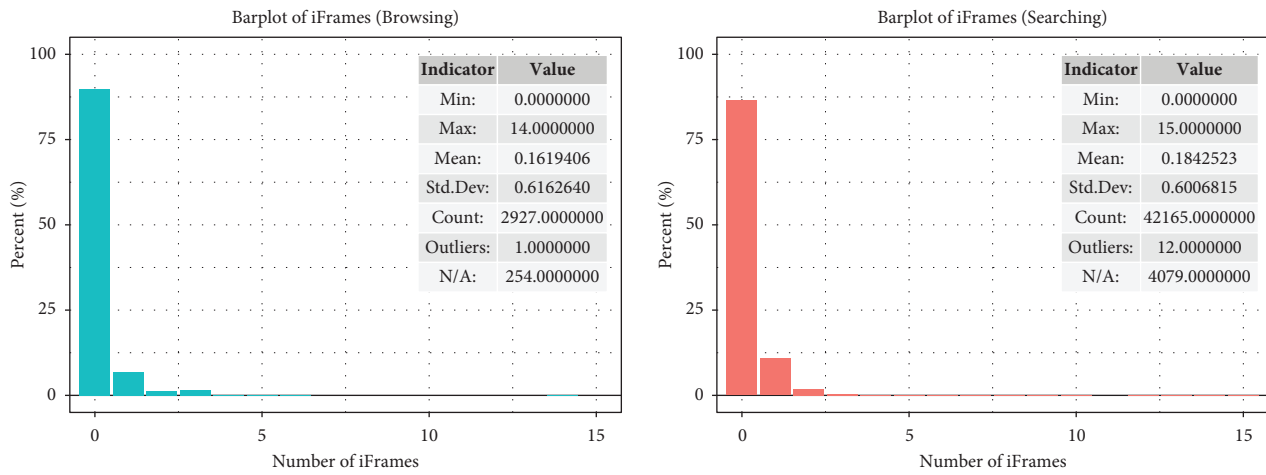


(f)

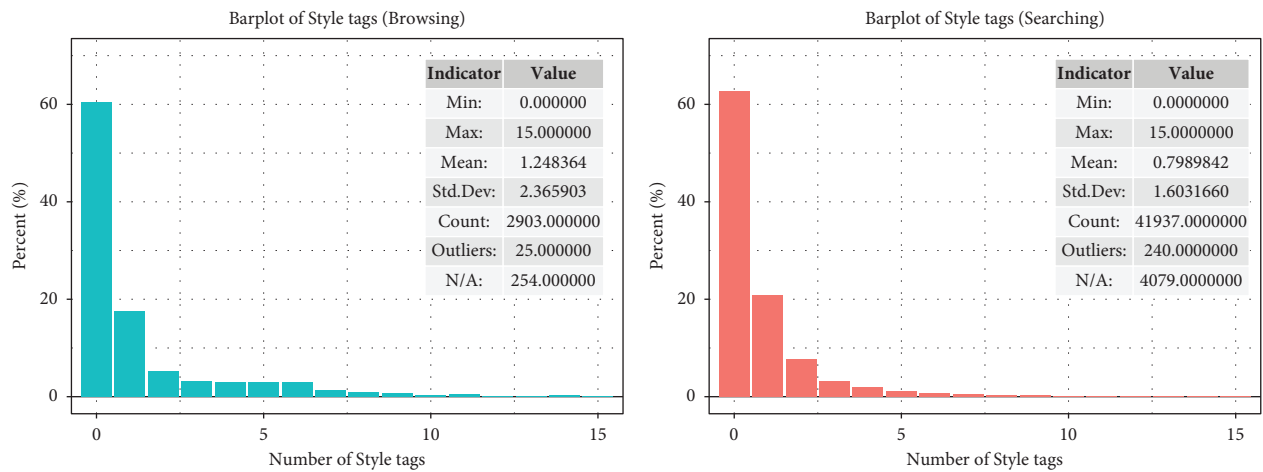
FIGURE 5: Continued.



(g)



(h)



(i)

FIGURE 5: Continued.

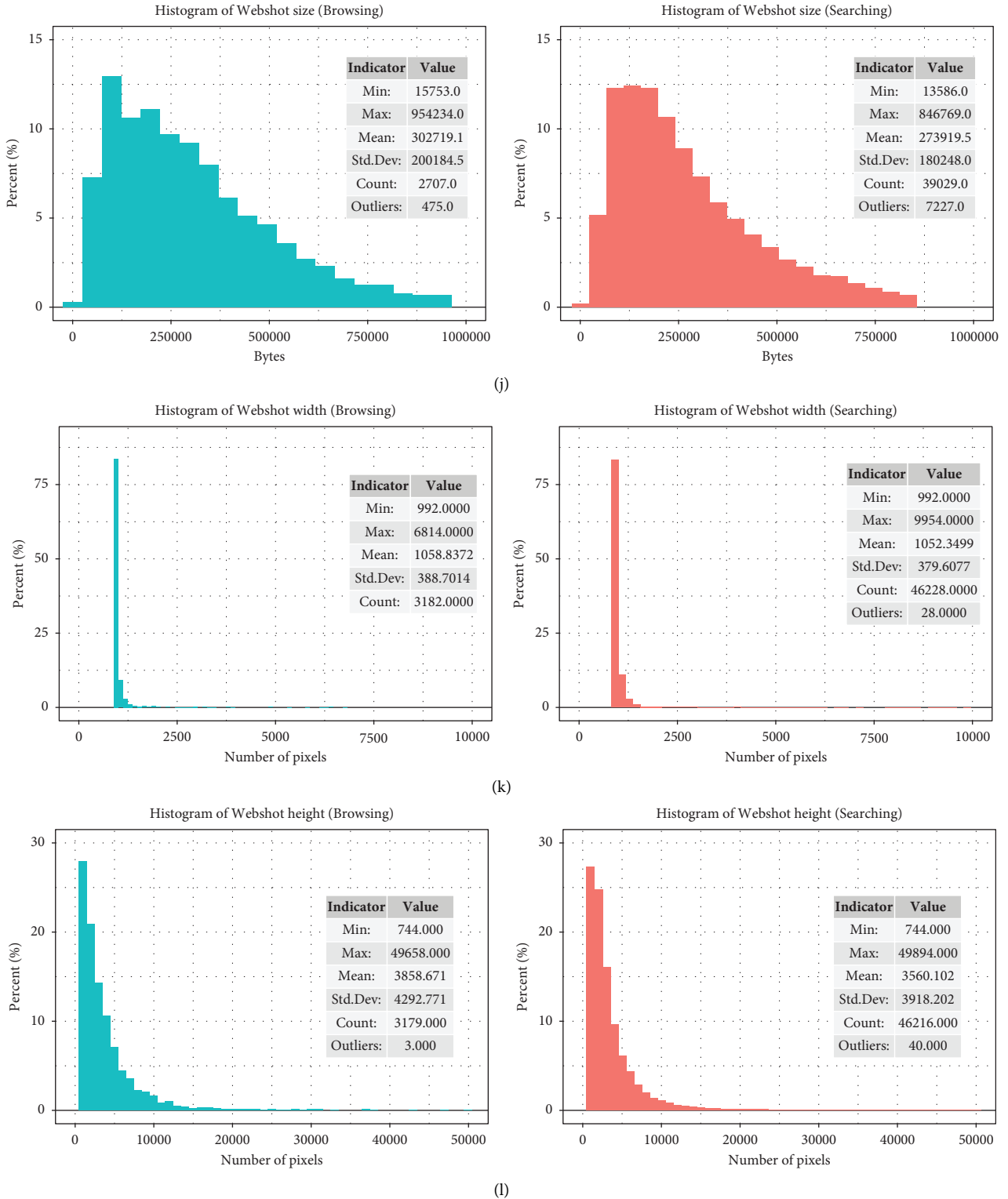


FIGURE 5: Distribution of the quantitative attributes about Web pages.

4.2. Use Case 2: Automatic Recognition of Error Web Pages. During our automatic data collection, some events blocked the download of HTML code or the webshot. These were caused by the following: a request for manual acceptance of cookies and SSL certificates; error messages such as HTTP

403 Forbidden, HTTP 404 Not Found, HTTP 406 Not Acceptable, HTTP 909 Denied permission; and exceeding timeout. We used exception handling inside the scripts to avoid interruptions in the execution of the programs. When an error occurred, the fields associated with the parameters

TABLE 3: Summary of statistical indicators for quantitative attributes.

Parameters	Browsing				Searching			
	Min.	Max.	Mean	Std. dev.	Min.	Max.	Mean	Std. dev.
URL length	14	161	31.73	12.59	21	200	69.84	30.99
Time (ms)	1.25	72.9	18.72	17.39	1.9	73.3	22.33	15.83
Size (KB)	0	200.89	50.49	44.33	0	176.87	46.82	39.58
Images	0	15	2.05	3.48	0	17	2.51	4.14
Scripts	0	20	3.35	5.1	0	17	3.04	4.43
CSS files	0	30	5.38	8.07	0	22	4.19	5.77
Tables	0	15	0.25	1.05	0	15	0.39	1.45
iFrames	0	14	0.16	0.62	0	15	0.18	0.6
Style tags	0	15	1.25	2.37	0	15	0.8	1.6
Size (KB)	15.75	954.23	302.72	200.18	13.59	846.77	273.92	180.25
Width (px)	992	6814	1058.8	388.7	992	9954	1052.3	379.6
Height (px)	744	49658	3859	4293	744	49894	3560	3918

or webshot were assigned the value “-1.” Thus, the programs could continue their execution, and the inexistence of webshots or attributes was solved. For the final dataset, we considered only URLs that had a respective webshot, as this is the most important element of our work. However, after a brief visual review of the webshots in the dataset, several error Web pages were detected, e.g., Web sites under construction, maintenance, domain offer, suspended account, page not found, browser incompatibility, virus, or phishing risks. Some of these are shown in Figure 6.

These webshots are not useful for the dataset, and so, we decided to remove them. Although the size of the final dataset would be smaller, we would obtain a cleaner dataset. Given that the URL connections corresponding to these webshots did not return HTTP 403 or HTTP 404 error messages, nor did the HTML code contain phrases such as “suspended account” or “page under construction,” text analysis was not possible. We implemented an image analyzer to avoid manual and visual verification of thousands of webshots, which requires excessive time and effort. We used a convolutional neural network (CNN), the state-of-the-art tool in computer vision, to detect error Web pages and then separate them into a dedicated folder, all automatically. In this sense, Web pages that do not contain useful information are called “ERROR” Web pages, whereas Web pages that contain valuable information are called “VALID” Web pages [5]. Here, we present an automatic detection of error Web pages based exclusively on their webshots. This consists of determining whether a Web page belongs to a “VALID” category or to an “ERROR” category, i.e., a binary classification problem. To address this, we followed the methodology shown in Figure 7, each phase of which is subsequently explained in detail.

4.2.1. Data Selection. The main resource for automatic learning is the data. In our case, the data are the images that will be the input for a training process that aims, iteratively, to obtain a known output (“valid” or “error”), and if an acceptable accuracy is reached, to make predictions. The training process requires images associated with their respective category: valid or error. Since our dataset consists of two groups of images (browsing and

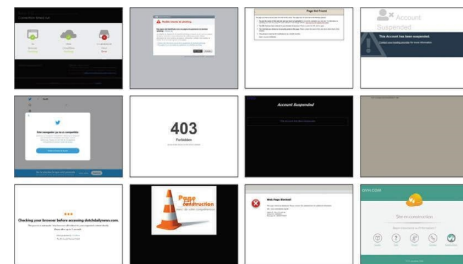


FIGURE 6: A sample of the error web pages.

searching), we selected the webshots of the smallest subset (browsing) to perform an exhaustive visual inspection and classify the images manually, obtaining the results shown in Table 4. Once the neural network model is adjusted, it classifies each webshot in the largest subset (searching) as valid or error.

The dataset for training an error Web page detection model has 3609 images, 427 error webshots, and 3182 valid webshots, which were uploaded to Google Drive in separate folders: “VALID” and “ERROR” (Figure 8).

We used *Google Colaboratory*, a free platform that offers powerful hardware and requires no installation or setup, supports Python through an online notebook and includes the packages and libraries to facilitate automatic learning such as *Tensorflow*, *Keras*, *Sklearn*, and others. Next, we describe the code developed (https://drive.google.com/file/d/1uKRfFb_KtP2KABRCOf1X_Bb847BDi7d/view?usp=sharing). The initial step is the connection to the data source where the images of our dataset have been stored within folders and subfolders named categories (Figure 8). This denomination facilitates the labeling of images with their corresponding category. Two instructions are needed to access *Google Drive*:

```
from google.colab import drive
drive.mount (“/content/drive”)
```

4.2.2. Data Splitting: Training and Validation. One of the tasks that characterize automatic learning is the division of data. Because we have only 3609 images, we consider two subsets: training and validation (Table 5). The training subset

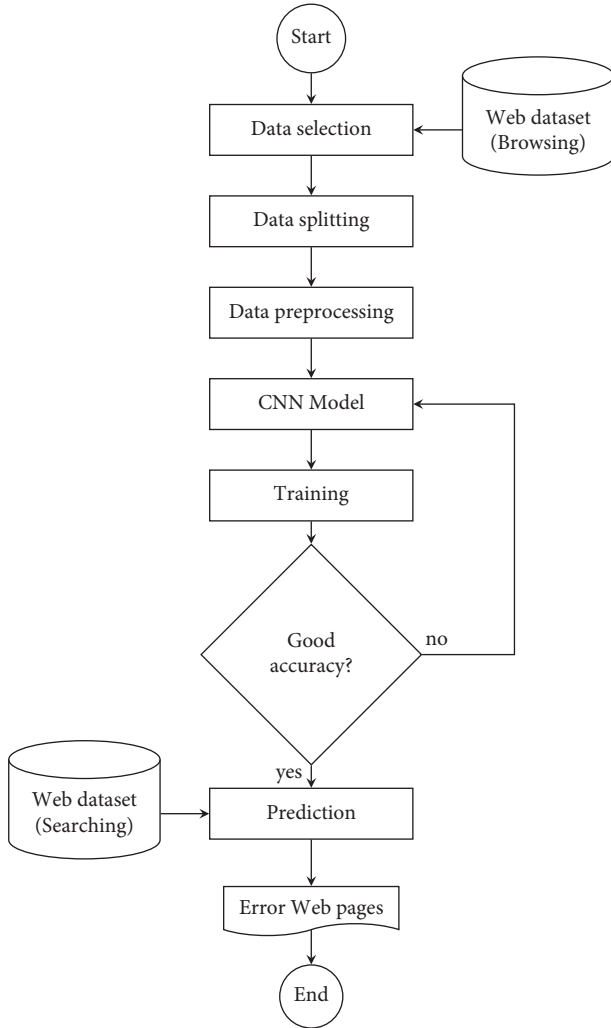


FIGURE 7: Methodology for detecting error Web pages.

TABLE 4: Dataset for binary classification (browsing webshots).

Categories	Browsing		
	Webshots	Valid	Error
Arts and entertainment	447	397	50
Business and economy	1058	892	166
Education	419	368	51
Government	730	669	61
News and media	458	394	64
Science and environment	497	462	35
Total	3609	3182	427

contains the largest number of images (80%) and is used to learn and fit the model parameters, while the validation subset (20%) is used to evaluate the capacity of the model.

Although the most appropriate is a balanced dataset, that is, an equal number of error cases and valid cases, we used all the images to obtain a better generalization. Automatically dividing into training and validation folders is useful to install and import the *split-folders* (<https://pypi.org/project/split-folders/>) package. It is necessary to specify the images directory, output directory, and the proportion to split (80% and 20%, respectively).

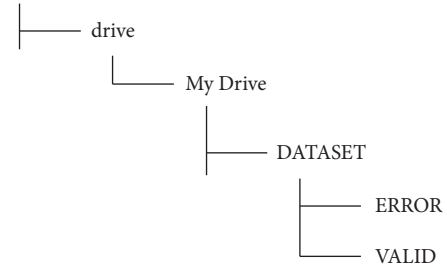


FIGURE 8: Directory structure of the dataset.

TABLE 5: Separation of dataset for training and validation.

Subsets	Webshots	Valid	Error	Percentage (%)
Training	2886	2545	341	80
Validation	723	637	86	20
Total	3609	3182	427	100

`splitfolders.ratio ("/content/drive/My Drive/DATASET", output = "/content/drive/My Drive/SPLIT", seed = 1337, ratio = (0.8, 0.2), group_prefix = None)`

The result is a new directory structure. Within the "SPLIT" folder, the "train" and "val" folders are created, and within each of these, the "ERROR" and "VALID" folders.

4.2.3. Data Preprocessing. The images must be prepared before modeling. First, we normalized the pixel values (integers between 0 and 255) to a scale between 0 and 1. The *ImageDataGenerator* class of the Keras framework divides all the values of pixels by the maximum pixel value (255).

```
train_datagen = ImageDataGenerator (rescale = 1./255)
```

Second, the images have different dimensions (width and height), and so, they were all resized to 256×256 pixels by setting the *target_size* parameter of the *flow_from_directory* method. This operation was performed in groups of 32 images (*batch_size*) that are labeled for binary classification (*class_mode*) according to the folder where they are stored (valid and error) within the training directory.

```
train_generator = train_datagen.flow_from_directory
(train_data_dir, target_size = (256, 256), batch_size
= 32, shuffle = False, class_mode = "binary")
```

In this way, the small values of both pixels and dimensions help speed up the training process. The above code applies to the validation data, with only the directory changing.

4.2.4. CNN Model. The architecture of the model is based on the convolutional neural network proposed by Liu et al. to detect malicious Web sites [22]. Since this is a similar problem, we only applied minor adaptations. Its structure (Figure 9) is composed of the following two parts:

- (i) Convolutional base: for automatic extraction of features. The input image is resized to 256×256

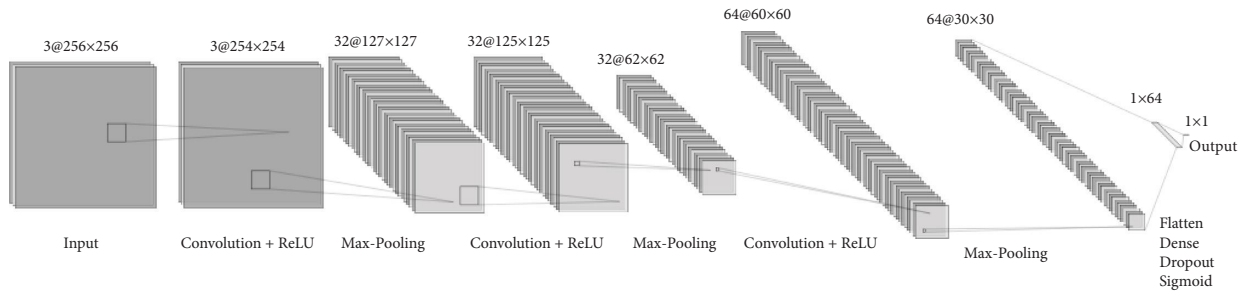


FIGURE 9: CNN architecture for error Web page detection.

pixels and separated into 3 RGB color channels (Red, Green, and Blue). It is then processed by 3 convolutional layers with their respective activation function (ReLU (Rectified Linear Unit)) and max-pooling layers. The first two convolutions use 32 filters (kernels), while the third convolution has 64, with a size of 3×3 , in contrast to the pool size of 2×2 .

- (ii) Binary classifier: features are received in flattened form by a fully connected layer, which applies dropout to reduce overfitting. The sigmoid function generates the prediction as a probability value between 0 and 1. If the value is greater than 0.5, the Web page is valid, and, if not, it is an error Web page.

Keras provides functions to implement this model from scratch in a simple way, just adding in sequence the convolutional, activation, pooling, dropout, flatten, and dense layers, and specifying their respective parameters. For example, the first convolutional block has the following instructions:

```
model = Sequential ()
model.add (Conv2D (32, (3, 3), input_shape = (256, 256, 3)))
model.add (Activation ("relu"))
model.add (MaxPooling2D (pool_size = (2, 2)))
```

Meanwhile, for the classifier part, we have

```
model.add (Flatten ())
model.add (Dense (64))
model.add (Activation ("relu"))
model.add (Dropout (0.5))
model.add (Dense (1))
model.add (Activation ("sigmoid"))
```

4.2.5. Training. Before starting training, we must explicitly define the *hyperparameters* required by the neural network for binary classification. We can establish the loss function that will be minimized by the optimization algorithm and the classification accuracy as the metric that will be collected and reported by the model.

```
model.compile (loss = "binary_crossentropy", optimizer = "rmsprop", metrics = ["acc"])
```

```
history = model.fit (train_generator, steps_per_epoch = train_samples//batch_size, epochs = epochs, validation_data = val_generator, validation_steps = validation_samples//batch_size, callbacks = [checkpoint])
```

After a few hours of computation, 20 iterations (*epochs*) of the training dataset (2886 images) have been executed; each iteration consists of 90 groups (*steps_per_epoch*) of 32 images (*batch_size*). The accuracy achieved is 96.6% (iteration 20), while in the validation stage the accuracy is 97.16% (iteration 16). The evolution of the process is summarized in the graphs of the learning curves (Figure 10).

The training and validation phases reached a high level of accuracy, both progressing to the same level, which is desirable. The model fits very well with the images provided, but how it behaves with new images (generalization) is uncertain. This concern is addressed by analyzing the difference between training and validation losses. The latter, despite oscillating, does not vary greatly from the other one until iteration 17, after which, they start to separate, with the possibility of overfitting. Therefore, the model is saved with the accuracy and parameters of iteration number 16. We can say that the model is capable of acceptably distinguishing error Web pages and valid Web pages and thereby moves to the prediction phase.

4.2.6. Prediction. The images from the largest set (Search) of our dataset become the input of the already trained and validated model. We used the *google.colab* library to select and upload the file (webshot) from the local drive with a click on the "Choose Files" button.

```
from google.colab import files
uploaded = files.upload ()
```

Once the file is 100% uploaded, it is preprocessed using *keras.preprocessing* and *NumPy* libraries to transform the image into an array with a suitable shape and normalized pixel values for the model, which makes the prediction.

```
img = image.load_img (path, target_size = (256, 256))
x = image.img_to_array (img)/255.
x = np.expand_dims (x, axis = 0)
```



FIGURE 10: Accuracy and loss in training and validation phases.

```
images = np.vstack ([x])
classes = model.predict (images)
```

The result, for images selected one at a time, is shown in Figure 11. The resized webshot is displayed and the prediction is a probability value between 0 and 1, less than 0.5, so the category assigned is ERROR (Figure 11(a)), and a case of a valid Web page, with a probability value very close to 1 (Figure 11(b)).

In addition to making predictions one by one, more important for our purpose is to generate predictions for groups of images. To do so, we simply select a list of files using the choose button. As our dataset is organized by topic category, we can select all the images in one category, e.g., “arts and entertainment.” An extract of the results is shown in Figure 12.

This list of predictions is passed to a spreadsheet, and by means of a filter, the Web pages of the error category are selected and saved as a text format file (*list.txt*). This file is the input to execute a command that moves all images from their original folder to the “ERROR” folder. This command line runs in *Windows* (PowerShell interface), although it is easily adaptable to different operating systems such as *Linux*.

```
cat list.txt | ForEach {mv $_ ERROR}
```

As a result of the prediction, 822 error Web pages and 7747 valid Web pages were found. Once the images were classified and separated, the visual verification was much faster, and we were able to manually establish the successes and fails of the classifier. Thus, we identified 1214 real error Web pages and 7355 real valid Web pages. The same procedure was performed for the remaining images in the other categories. The results are summarized in Table 6.

We used the *Confusion Matrix* to evaluate and determine the accuracy of our model. This table compares reality and prediction and, based on successes and failures, calculates an accuracy value. For example, for the “arts and entertainment” category, the classifier predicted 822 error Web pages, but it failed in 32 instances. There were 7747 predictions of valid Web pages, but 424 were incorrect. These values are

placed in Table 7 and substituted into the formula for accuracy.

$$\text{Accuracy} = \frac{790 + 7323}{790 + 32 + 424 + 7323} = 94.68\%. \quad (1)$$

The classifier reached an accuracy of 94.68% for this category, which is good considering the small number of images involved in the training process. Table 8 shows the respective confusion matrix for each category and a total matrix indicating an accuracy of 92.47% for the entire searching dataset.

After running the automatic error Web page detection, the debugging is complete. The composition and final size of our dataset are shown in Table 9. Combining browsing and searching techniques, we managed to collect 49438 valid Web pages that occupy approx. 17 GB.

4.3. Use Case 3: Web Categorization. Here, we demonstrate the use of the presented dataset with a practical case related to Web categorization. The classification of Web pages, also called *Web categorization*, determines whether a Web page or a Web site belongs to a particular category. For example, judging whether a page is about “arts,” “business,” or “sports” is an instance of subject classification [23]. Instead of analyzing complex programming code, visual appearance is also an important part of a Web page, and many topics have a distinctive visual appearance. For example, Web design blogs have a highly designed visual appearance, whereas newspaper sites have a great deal of text and images [1]. In this sense, we present an automatic categorization of Web pages according to topic or subject and based exclusively on their visual appearance. We leverage the dataset generated in this work, formed by webshots belonging to 6 categories: arts and entertainment, business and economy, education, government, news and media, and science and environment. Therefore, the problem becomes a multiclass categorization. We implemented a *deep learning* model with a convolutional neural network. In essence, this is a learning process with the webshots collected in order to achieve an acceptable accuracy

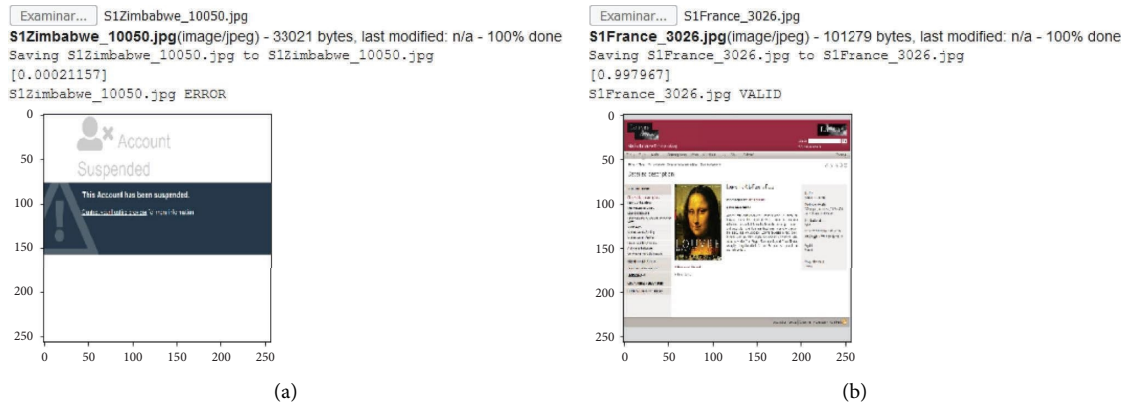


FIGURE 11: Prediction of error (a) and valid (b) Web page.

```
S1Belize_1298.jpg VALID
S1Denmark_2410.jpg VALID
S1Spain_2654.jpg VALID
S1Moldova_5622.jpg ERROR
S1Guadeloupe(France)_3334.jpg VALID
S1Nicaragua_6280.jpg VALID
S1Denmark_2409.jpg VALID
S1Jamaica_4543.jpg VALID
S1Thailand_8478.jpg VALID
S1Taiwan_9033.jpg VALID
S1UnitedArabEmirates_103.jpg VALID
S1NewZealand_6630.jpg VALID
S1Uruguay_9424.jpg VALID
S1Lithuania_5276.jpg VALID
S1Mongolia_5891.jpg VALID
S1Bahamas_1173.jpg VALID
S1Reunion(France)_7335.jpg VALID
S1Zimbabwe_10064.jpg ERROR
S1SaintVincentandtheGrenadines_9571.jpg VALID
S1Iran_4324.jpg VALID
S1Andorra_75.jpg VALID
S1Liechtenstein_5133.jpg VALID
S1Bangladesh_702.jpg ERROR
S1Iran_4334.jpg VALID
```

FIGURE 12: Prediction for a group of Web pages.

TABLE 6: Results of the binary Web categorization.

Categories	Searching					
	Webshots	Valid (prediction)	Error (prediction)	Valid (real)	Error (real)	Accuracy (%)
Arts and entertainment	8569	7747	822	7355	1214	94.68
Business and economy	8699	8004	695	7546	1153	93.79
Education	8742	8083	659	7524	1218	92.80
Government	8088	7363	725	6685	1403	90.85
News and media	11574	10597	977	9650	1924	90.80
Science and environment	8893	8137	756	7496	1397	92.34
Total	54565	49931	4634	46256	8309	92.47

TABLE 7: Confusion matrix for the “arts and entertainment” category.

	Prediction	
	Error	Valid
Real	Error 790	Valid 424
	Valid 32	7323

and then make predictions. We hoped to capture features (difficult to identify manually) that can distinguish categories, predict to which of them a Web page would belong, analyze the difficulty of the topic classification of the Web pages and verify whether there are particular patterns for each category. The following results were selected from a series of several

experiments in which different models and architectures were tested using the entire dataset and parts of it. The code developed, as well as the weights of the adjusted deep learning model, are publicly accessible (<https://osf.io/8zfh2>). The best results were obtained with the *transfer learning* technique and the images of the browsing dataset. This may be because in

TABLE 8: Confusion matrix for the rest of categories and overall result.

		Prediction	
		Error	Valid
<i>Business and economy</i>			
Real	Error	654	499
	Valid	41	7505
<i>Education</i>			
Real	Error	624	594
	Valid	35	7489
<i>Government</i>			
Real	Error	694	709
	Valid	31	6654
<i>News and media</i>			
Real	Error	918	1006
	Valid	59	9591
<i>Science and environment</i>			
Real	Error	736	661
	Valid	20	7476
<i>Overall</i>			
Real	Error	4416	3893
	Valid	218	46038

TABLE 9: Composition and size of the final dataset.

Categories	Browsing Webshots	Searching Webshots	Total
Arts and entertainment	397 (147 MB)	7355 (2.58 GB)	7752
Business and economy	892 (300 MB)	7546 (2.48 GB)	8438
Education	368 (126 MB)	7524 (2.64 GB)	7892
Government	669 (253 MB)	6685 (2.47 GB)	7354
News and media	394 (237 MB)	9650 (3.19 GB)	10044
Science and environment	462 (193 MB)	7496 (2.63 GB)	7958
Total	3182 (1.22 GB)	46256 (15.99 GB)	49438

The values in bold correspond to the total number of webshots collected with the browsing and search techniques, respectively. In parentheses, their sizes in Gigabytes.

a Web directory such as *BOTW*, Web pages go through a rigorous registration process under the supervision of human specialists, so they have a better distinction and categorization. The dataset for the multiclass categorization is composed only of the webshots of the set of Browsing, which was organized and split according to Table 10. For the training process, we have a balanced amount of data, i.e., the same number of images for each category. The category with fewest images (education) was the basis for randomly selecting the same number of images in the other categories. One image of 3.68 MB and resolution of 992×30154 was discarded because the Python imaging library does not open larger images to avoid malicious attacks. Thus, each category has 367 images, a total of 2202 images, 80% for training, while the remaining 20% is used for validation (both sets randomly selected).

The images are stored within the directory structure shown in Figure 13. Within the main folder of the dataset the division in training and validation, and the subfolders represent the categories, which have the same names as the topics considered in this work.

After organizing the images, a preprocessing step is advisable to normalize the image’s pixel values (integers between 0 and 255) to the scale of values between 0 and 1. It

TABLE 10: Dataset split for multiclass categorization.

Categories	Webshots	Dataset	Train (80%)	Val. (20%)
Arts and entertainment	397	367	293	74
Business and economy	892	367	293	74
Education	368	367	293	74
Government	669	367	293	74
News and media	394	367	293	74
Science and environment	462	367	293	74
Total	3182	2202	1758	444

is also necessary to resize to the 224×224 pixels recommended for the model, because the images in the dataset have different dimensions (width and height). Both are common practices that help speed up the process of training. Several models were tested with a variety of options to achieve greater accuracy. The final model exploits the transfer learning technique using *ResNet* [24], a competitive CNN pretrained on the ImageNet dataset (more than

```

|-train
| | -EDUCATION
| | -ARTS-AND-ENTERTAINMENT
| | -NEWS-AND-MEDIA
| | -SCIENCE-AND-ENVIRONMENT
| | -BUSINESS-AND-ECONOMY
| | -GOVERNMENT
|-val
| | -EDUCATION
| | -ARTS-AND-ENTERTAINMENT
| | -NEWS-AND-MEDIA
| | -SCIENCE-AND-ENVIRONMENT
| | -BUSINESS-AND-ECONOMY
| | -GOVERNMENT
    
```

FIGURE 13: Directory structure by category for training and validation.

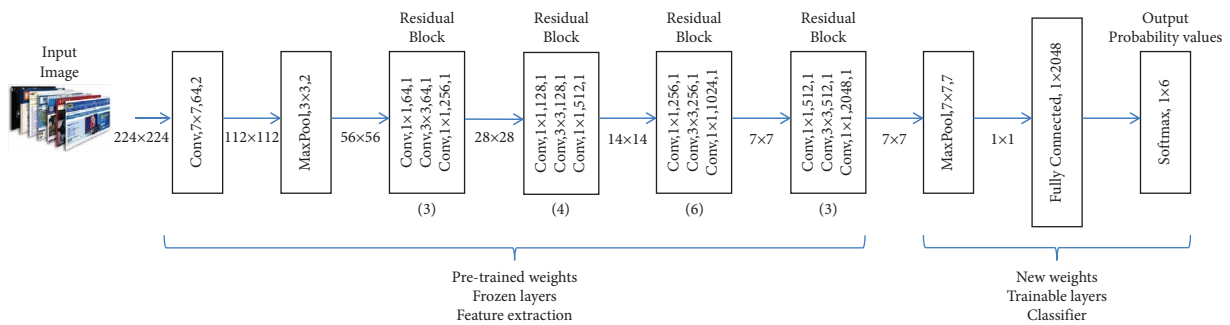


FIGURE 14: Architecture of the model based on ResNet-50 for Web categorization.

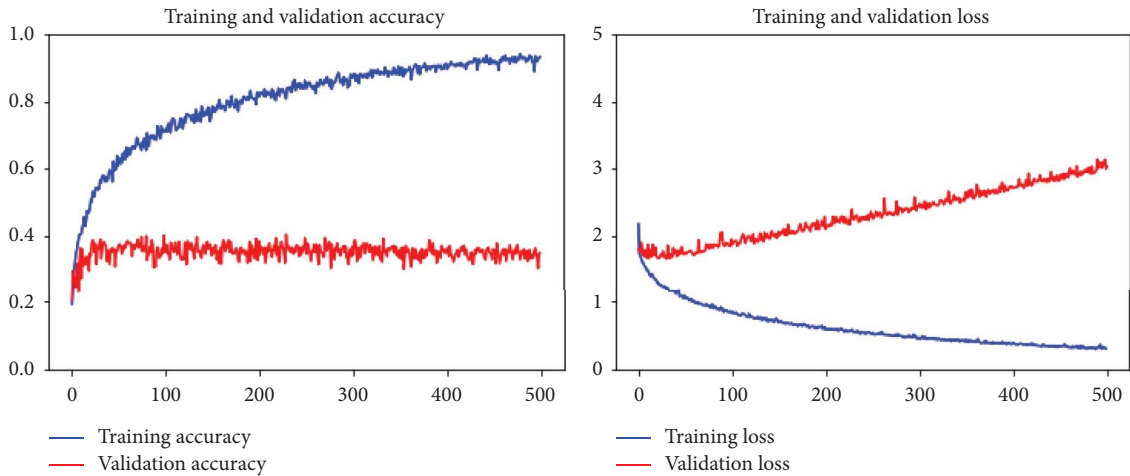


FIGURE 15: Accuracy and loss in training and validation phases.

14 million images belonging to 1000 categories), and which was the winner of the ImageNet challenge in 2015. Despite the more up-to-date models, ResNet is still highly popular for transfer learning implementations. We used *ResNet-50*, which is 50 layers deep, and whose convolutional basis is kept for feature extraction, while the classifier part is replaced by a new one that will predict the probabilities for 6 classes corresponding to our categories (Figure 14). Only the classifier layers are trained in our dataset. After 500 iterations (epochs) of the whole training set (1758 images) in

groups of 32 images (batch_size), an accuracy of 94.26% was obtained and 40.38% in the validation phase. The evolution of the process is summarized in the following learning curves (Figure 15).

A suitable solution to this problem must meet: (a) high training accuracy; (b) validation and training curves that are very close to each other; and (c) small difference between the validation and training error. The graphs show only the first element is accomplished, and so the model learned very well but not for generalization, i.e., to classify new images

TABLE 11: Confusion matrix for validation data.

		Prediction					
		Arts and entertainment	Business and economy	Education	Government	News and media	Science and environment
Real	Arts and entertainment	38	6	7	17	3	3
	Business and economy	16	20	1	25	6	6
	Education	11	8	16	27	5	7
	Government	6	4	9	46	5	4
	News and media	26	2	3	10	31	2
	Science and environment	17	7	4	21	6	19

acceptably. Although we increased the data, tuned the hyperparameters, and applied regularization techniques such as dropout, neither is accuracy improved nor is overfitting significantly reduced. For a better understanding of the results, Table 11 shows the confusion matrix with the validation data.

If we focus on the categories of arts and entertainment, government, and news and media, the model is correct in most cases, although the number of successes is low. This test takes validation data, a total of 444 images, achieving an accuracy of 38.29%, according to the confusion matrix. For the remaining categories, the model becomes significantly confused. Classifying these categories is a complex problem. The composition of today’s Web pages is becoming increasingly complex and the content has a high variability of visual features, even within the same category.

5. Conclusion and Future Work

We created a large dataset on Web pages that combines different types of data: text, numbers, and images. We automated the workflow with scripts in Python and R to collect URLs and their respective webshots, while scraping allowed us to extract attributes from each Web page. The methodology designed can be adapted to problems requiring the collection, organization, analysis, and publication of large amounts of data. In addition, we developed three Web intelligence applications using this dataset. First, the qualitative and quantitative attributes of the dataset allowed us to obtain useful information about the structure of the Web pages. Statistical analysis of these attributes showed a very heterogeneous distribution, high variability, and a tendency towards low values. This suggests that Web design follows an implicit rule of optimization, since the higher the values, the longer the page download and display time, leading to user discomfort. Second, we were able to automatically collect a total of 58174 webshots, although the final dataset was reduced to 49438 due to the elimination of error Web pages. We implemented an automatic detection of error Web pages based on a CNN model from scratch achieving acceptable accuracy. This approach could be a more efficient debugging process to address the significant presence of invalid pages on the Internet, which affects webmasters, search engines and users in general. Third, Web categorization based exclusively on webshots using a multiclass CNN model proved to be a complex problem. The difficulty increases when the

categories cover a wide range of topics and within each topic there is great variability in the visual appearance of Web pages. However, the remarkable accuracy of the model for the government category, as well as the arts and entertainment category, allows us to infer the existence of distinctive visual patterns, which can be a baseline for future research. The results could be improved by increasing the dataset, preprocessing webshots to the same size and resolution, cropping, and scaling. Finally, our work may motivate the extension of the dataset with further categories, URLs, webshots, and other attributes; the exploration of alternative URL sources, i.e., a search engine other than Google and a Web directory other than BOTW; and the improvement of the accuracy achieved in multiclass Web categorization with deeper and more recent convolutional neural networks.

Data Availability

The dataset and code used to support the findings of this study have been deposited in the Open Science Foundation repository (<https://osf.io/7ghd2/>).

Disclosure

A previous version of this paper has been published in arxiv [25].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work has been funded by the grant awarded by the Central University of Ecuador through budget certification No. 34 of March 25, 2022 for the development of the research project with code: DOCT-DI-2020-37.

References

- [1] V. Boer, M. Someren, and T. Lupascu, “Classifying web pages with visual features,” *Journal of Clinical Virology*, vol. 1, pp. 245–252, 2010.
- [2] T. Berners-Lee, R. Cailliau, A. Luotonen, H. F. Nielsen, and A. Secret, “The world-wide web,” *Communications of the ACM*, vol. 37, no. 8, pp. 76–82, 1994.

- [3] M. James, J. Gillies, and R. Cailliau, *How the Web Was Born: The story of the World Wide Web*, Oxford University Press, Oxford, UK, 2000.
- [4] J. Hendler and W. Hall, "Science of the world wide web," *Science*, vol. 354, no. 6313, pp. 703-704, 2016.
- [5] C. Zhou, J. Zhao, T. Ma, and X. Zhou, "Haif: a hierarchical attention-based model of filtering invalid webpage," *IEICE-Transactions on Info and Systems*, no. 5, pp. 659-668, 2021.
- [6] Juan Velasquez and L. Jain, *Advanced Techniques in Web Intelligence- I*, Springer, Berlin, Germany, 2010.
- [7] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*, Springer-Verlag, Berlin, Germany, 2006.
- [8] S. Lassri, E. L. H. Benlahmar, and A. Tragha, "Machine learning for web page classification: a survey," *International Journal of Information Science and Technology*, vol. 3, p. 9, 2019.
- [9] M. Du, Y. Han, and L. Zhao, "A heuristic approach for website classification with mixed feature extractors," in *Proceedings of the 2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, pp. 134-141, Singapore, December, 2018.
- [10] D. López-Sánchez, J. C. Rodríguez, and A. González, "A cbr system for image-based webpage classification: case representation with convolutional neural networks," in *Processings of the Florida Artificial Intelligence Research Society Conference*, p. 5, Marco, Island, FL, USA, May, 2017.
- [11] D. López-Sánchez, A. González, and J. C. Rodríguez, "Visual content-based web page categorization with deep transfer learning and metric learning," *Neurocomputing*, vol. 338, p. 4, 2019.
- [12] K. Reinecke and K. Gajos, "Quantifying visual preferences around the world," in *Proceedings of the Conference on Human Factors in Computing Systems*, p. 4, Hamburg, Germany, April, 2014.
- [13] F. Vincent, "Circl images phishing dataset," 2019, <https://www.circl.lu/opendata/circl-phishing-dataset-01/>.
- [14] F. Vincent, "Circl images ail dataset," 2019, <https://www.circl.lu/opendata/circl-ail-dataset-01/>.
- [15] T. Fu, A. Ahmed, and H.-C. Chen, "A focused crawler for dark web forums," *Journal of the Association for Information Science and Technology*, vol. 61, pp. 1213-1231, 2010.
- [16] D. Jia, W. Dong, and R. Socher, "Imagenet: a large-scale hierarchical image database," in *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248-255, IEEE, Miami, FL, USA, June, 2009.
- [17] Institute for Computer Research, *Web Image Dataset*, Institute for Computer Research, Alicante, Spain, 2019.
- [18] M. Nordhoff, T. August, N. Oliveira, and K. Reinecke, "A case for design localization: diversity of website aesthetics in 44 countries," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 04, Montreal, Canada, April, 2018.
- [19] A. B. K. Website, *Optimization: Speed, Search Engine and Conversion Rate Secrets*, O'Reilly Media, Sebastopol, CA, USA, 2008.
- [20] E. Michailidou, S. Harper, and S. Bechhofer, "Visual complexity and aesthetic perception of web pages," in *Proceedings of the 26th Annual ACM International Conference on Design of Communication, SIGDOC '08*, pp. 215-224, New York, NY, USA, March, 2008.
- [21] R. White, Ian Ruthven, and D. Kelly, *Interactive Information Seeking, Behaviour and Retrieval*, Facet Publishing, London, UK, 2018.
- [22] D. Liu, J.-H. Lee, W. Wang, and Y. Wang, "Malicious websites detection via cnn based screenshot recognition," in *Proceedings of the 2019 International Conference on Intelligent Computing and its Emerging Applications (ICEA)*, pp. 115-119, Tainan, Taiwan, September, 2019.
- [23] X. Qi and B. D. Davison, "Web page classification: features and algorithms," *ACM Computing Surveys*, vol. 41, no. 2, pp. 1-31, 2009.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015, <https://arxiv.org/abs/1512.03385>.
- [25] C. Mejia-Escobar, M. Cazorla, and E. Martinez-Martin, "A large visual, qualitative and quantitative dataset of web pages," 2021, <https://arxiv.org/abs/2105.07113>.