Automatic Evaluation of Web Pages Quality through Deep Learning

Christian Mejia Escobar
cme26@alu.ua.es

Tutor: PhD. Miguel Angel Cazorla Quevedo
Director: PhD. Ester Martinez Martin

University of Alicante (Spain)

2021

Content

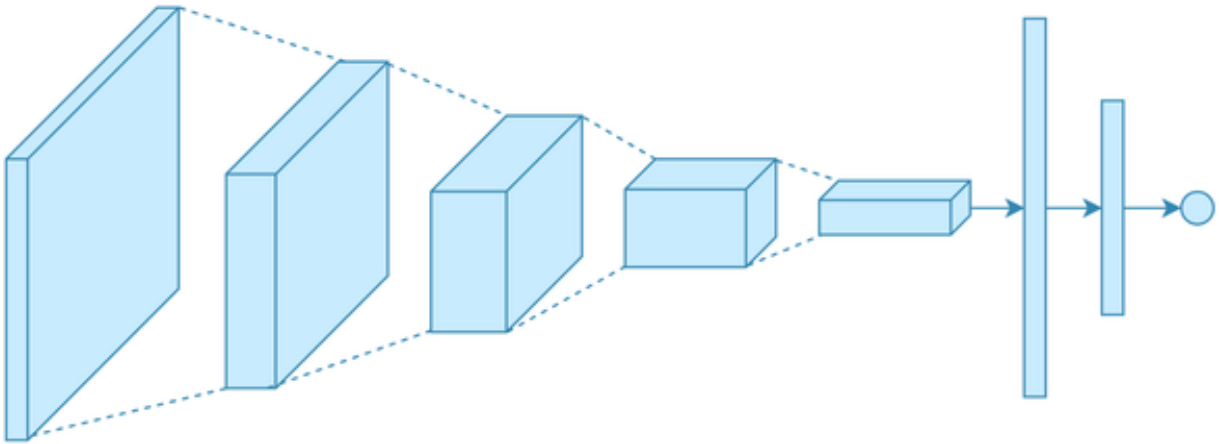
Importance

- We focus on the World Wide Web (just **Web**), because it is the most important global communication platform nowadays.
- Technological, social, and economic factors have made the Web a field of interest for the business sector and **scientific research**.
- A **quality** *Web site*[1] can bring significant benefits to a growing number of individuals and organizations.
- Therefore, the evaluation of the quality of a Web site (particularly, its homepage) is a **relevant** problem.
- This task is **difficult and subjective** for humans beings, as there is no universal agreement on the correct quality guidelines.
- Using AI through ML is identified as the most promising method for determining the quality of Web pages.

[1] An organized collection of related Web pages

Hypothesis

The problem of evaluation and categorization of the quality of a Web site can be addressed by a **Deep Learning** model based on a Convolutional Neural Network, which can automatically extract features that would be difficult to establish manually.



Goals

Main

To develop a Deep Learning system for the evaluation and multiclass categorization of the quality of Web pages.

1. To create a large, reliable and publicly available dataset of Web page screenshots and their respective tags.
2. To implement a Deep Learning model based on a convolutional neural network for the evaluation and classification of Web pages.
3. To define quality characteristics identified by the model as determinants for the quality of a Web page, in order to recommend them as design guidelines.
4. To implement an online application for the analysis, evaluation and categorization of the quality of a Web page.

Challenges

- **Absence of a Web pages dataset**
  - Large and reliable
  - Available, public for download
  - Not only screenshots, design parameters too
- **How quality is evaluated?**
  - There are no universal standard rules
  - Quality $\neq$ Design $\neq$ Aesthetics
  - Quality $=$ Design $+$ Aesthetics $+$ Functionality $+$ Content $+$ Technology $+$ Popularity $+$ ...
  - Define quality categories
- **Analysis based on screenshots**
  - Complex mixture of elements: textual, graphic and multimedia
  - Features not well defined
  - Why analyze the image and not the HTML code?

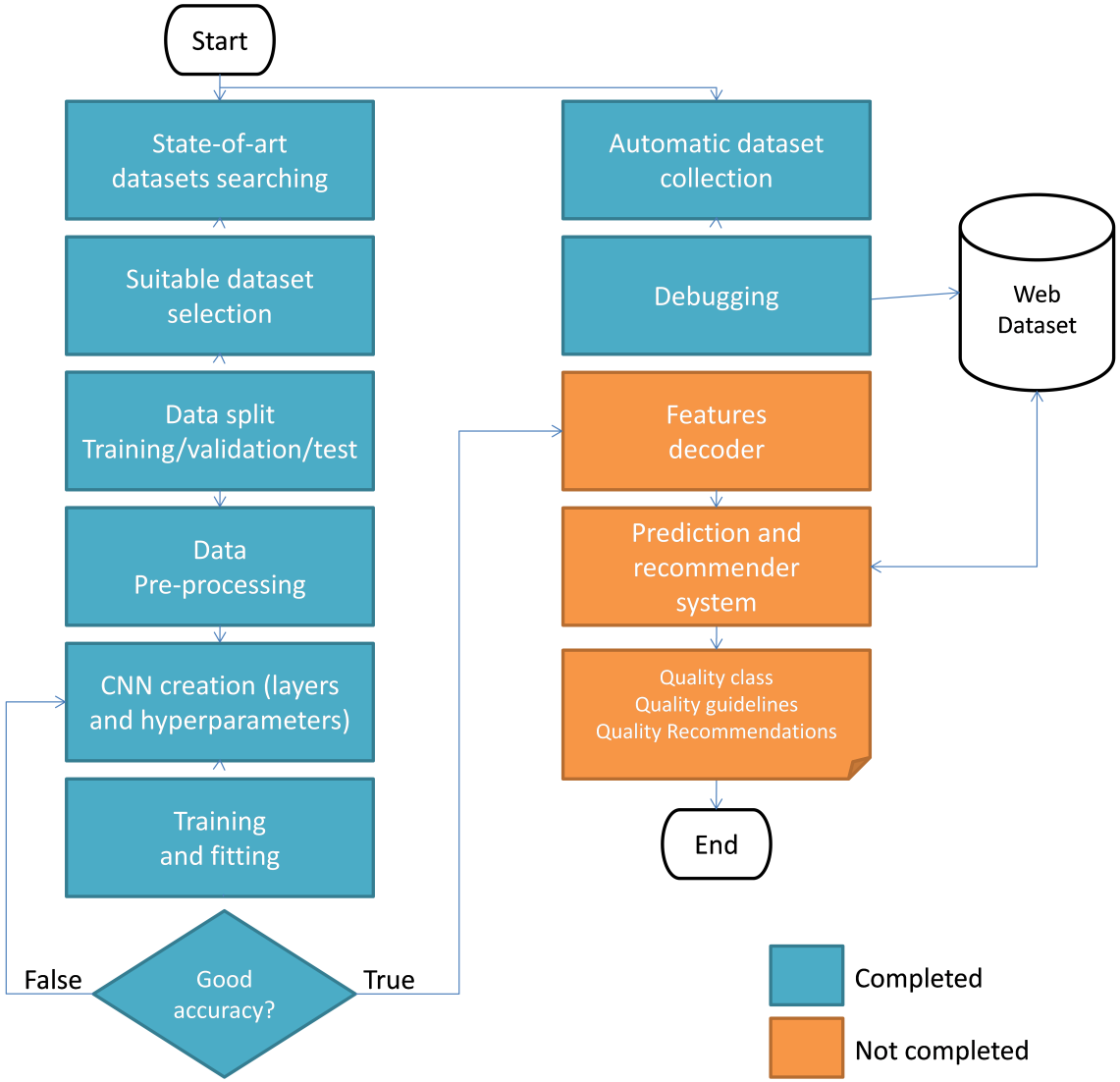
Contributions

- A dataset of 49,438 Web pages from all countries worldwide and classified in the following topics: arts and entertainment, business and economy, education, government, news and media, and science and environment.
- A combination of different types of data: images, text, and numbers, which represent the visual aspect of the full Web page (webshot), and qualitative and quantitative attributes, respectively.
- Our classifier and recommender system can have multiple uses, e.g. Web design and redesign, analysis of aesthetics, optimization of search engines, Web categorization, security, accessibility, programming, ML projects, competitions, etc.

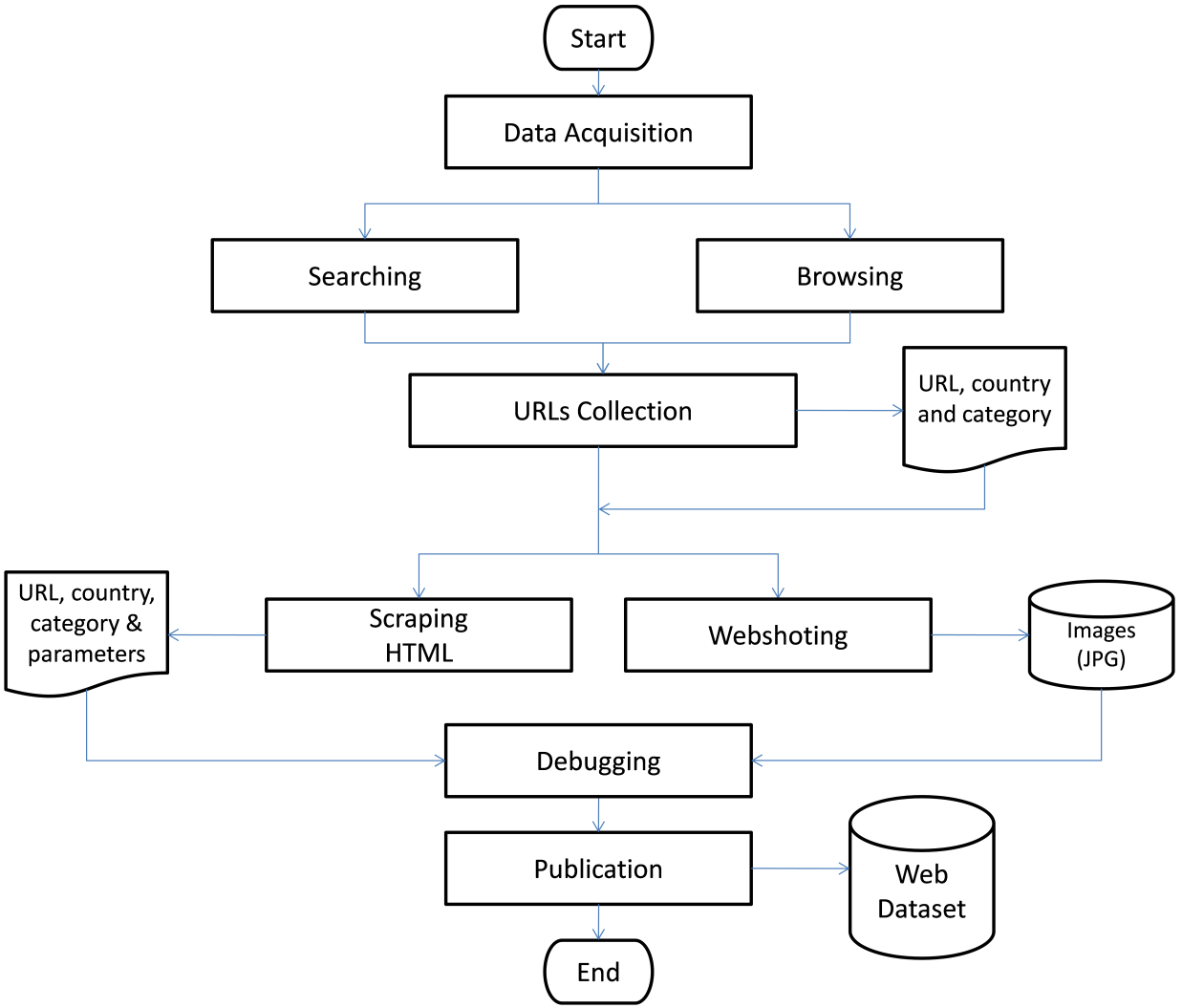
State-of-the-art datasets

Owner & 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

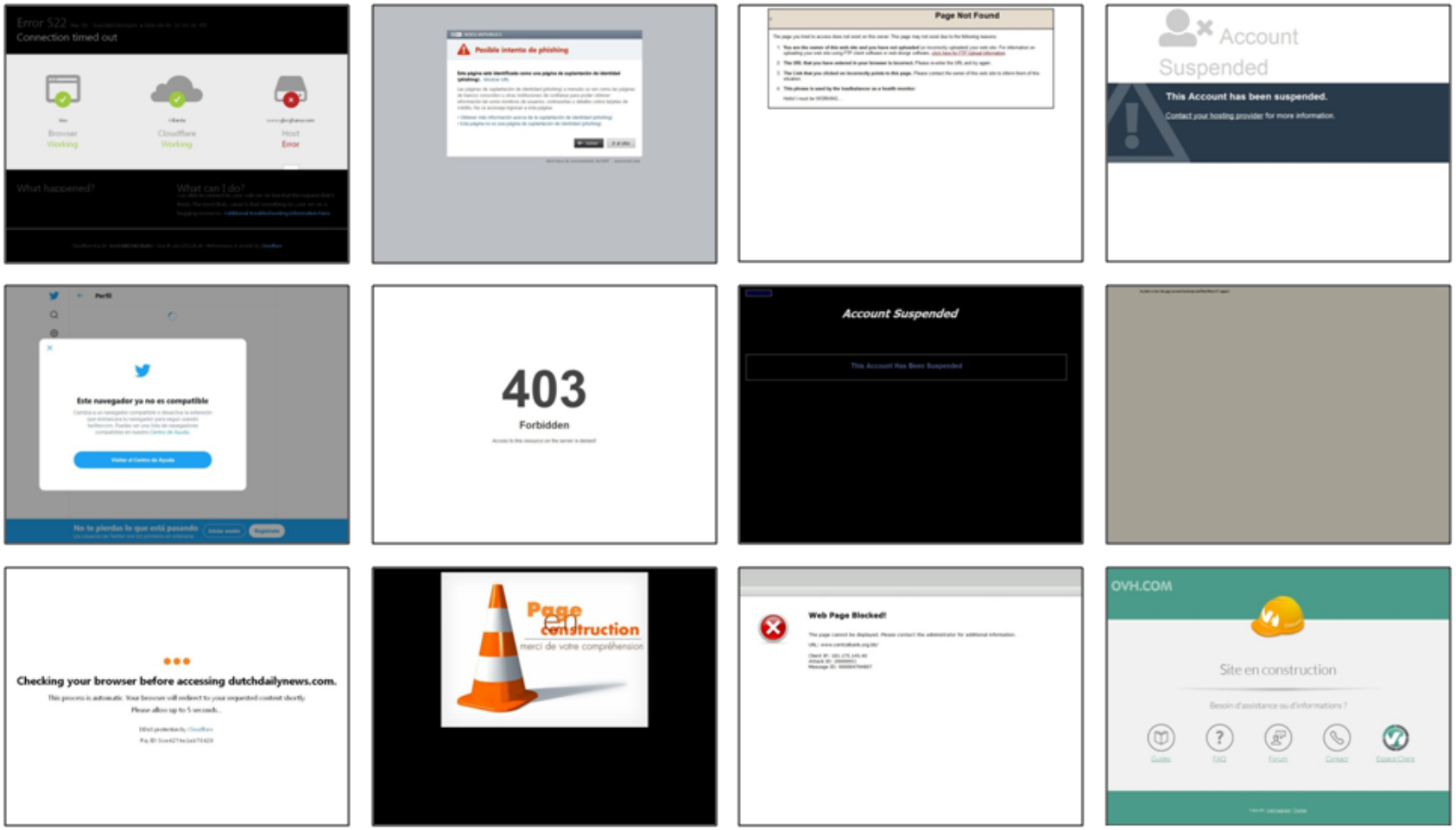
Flow chart


Start
State-of-art datasets searching
Suitable dataset selection
Data split Training/validation/test
Data Pre-processing
CNN creation (layers and hyperparameters)
Training and fitting
Good accuracy?
False
True
Automatic dataset collection
Debugging
Web Dataset
Features decoder
Prediction and recommender system
Quality class
Quality guidelines
Quality Recommendations
End
Completed
Not completed

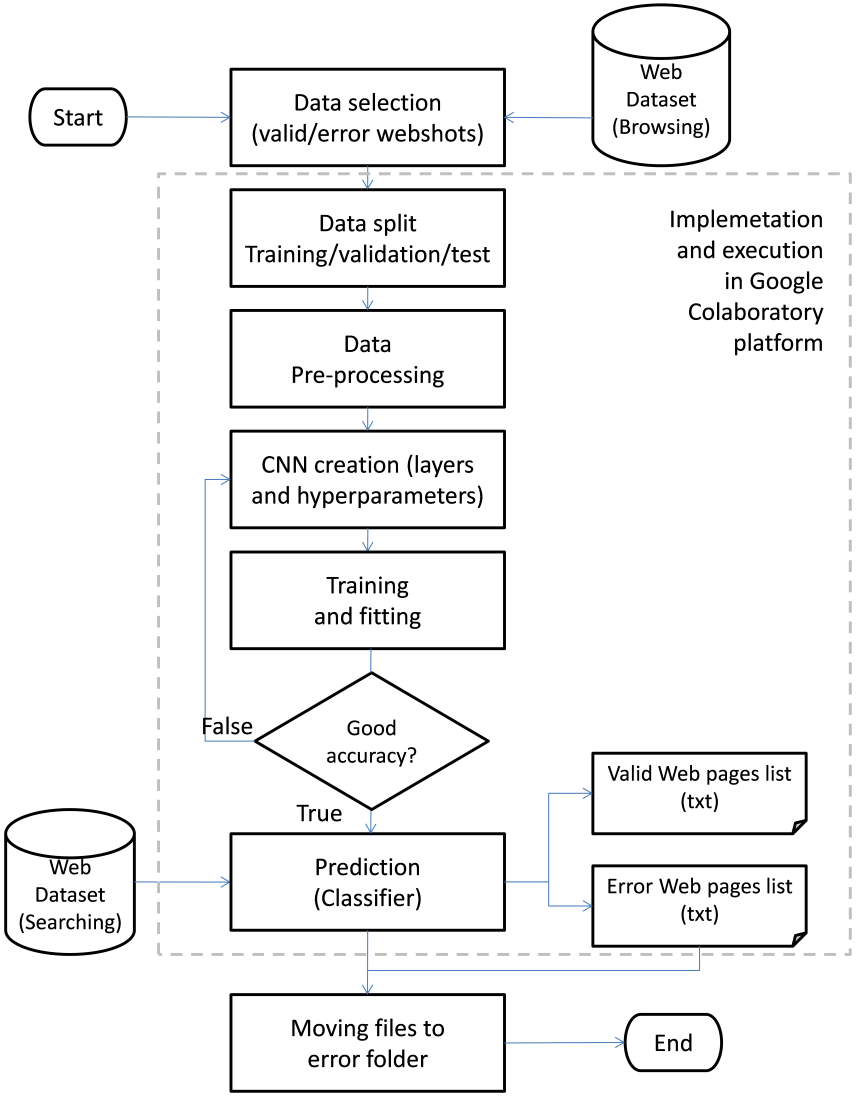
Methodology to collect Web pages dataset



Error Web pages



Methodology to detect error Web pages



Experiments

- Web dataset collection
  - Binary categorization: detection of error Web pages
  - Multiclass categorization: by topic of the Web pages
- **Quality classification of Web pages**

Computational platforms and tools

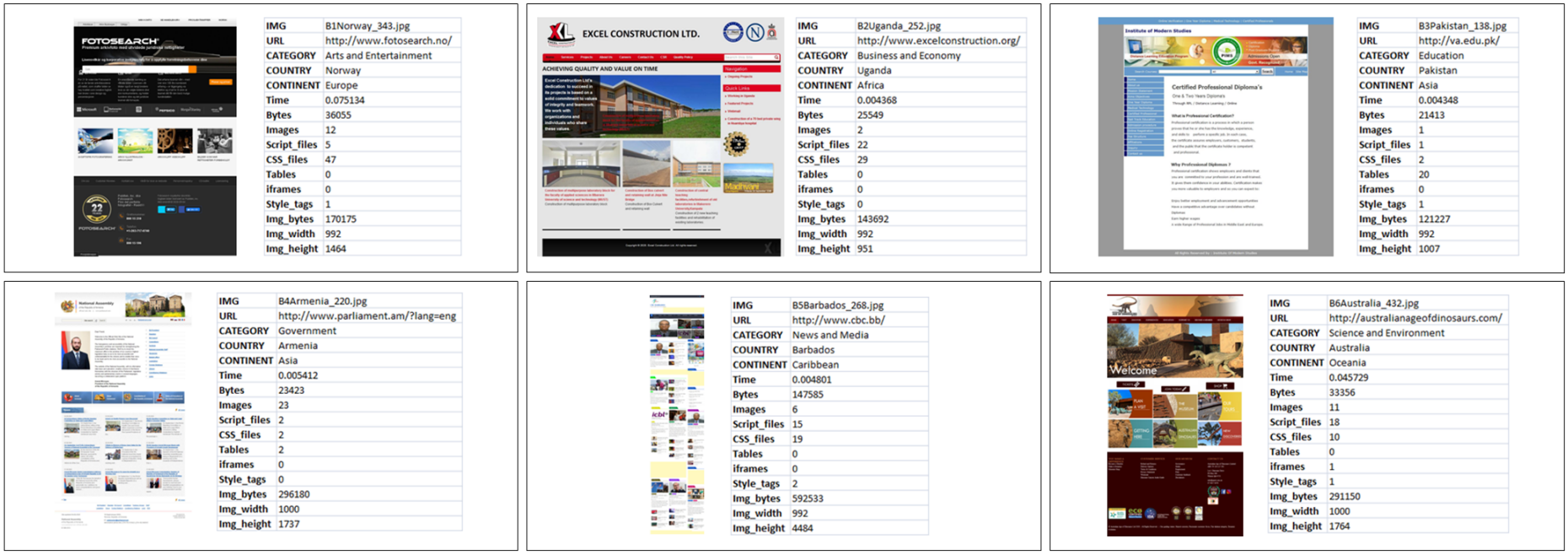
		Local (Laptop)	Remote (Servergpu)	Cloud (Google Colab)
Hardware	**CPU**	Intel Core i3 2.53 GHz	Intel(R) Core(TM) i9-7920X CPU @ 2.90GHz	Intel(R) Xeon(R) CPU @ 2.30GHz
	Cores / Threads	1	12 / 24	4 / 2
	RAM	4 GB	64 GB	25.51 GB
	Disk	500 GB	Disk array storage	68.40 GB
	GPU	Intel HD Graphics integrated	NVIDIA GeForce RTX208 (4)	NVIDIA Tesla P100
	GPU RAM	_	12 GB	16 GB
Software	**Operating system**	Windows 7	Ubuntu 18.04.5 LTS	Ubuntu 18.04.5 LTS
	Programming language	Python 2.7	Python 2.7.17	Python 3.7.10
	IDE	Anaconda Spyder	vi Editor	Jupyter Notebook

Web dataset collection

Table: Structure of the dataset.

Element	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 the Web page
Continent		Region grouping countries
Category		Main thematic of the 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 on the 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

Web dataset collection



IMG	B1Norway_343.jpg
URL	http://www.fotosearch.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	B2Uganda_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	143692
Img_width	992
Img_height	951

IMG	B3Pakistan_138.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	B4Armenia_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	B5Barbados_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

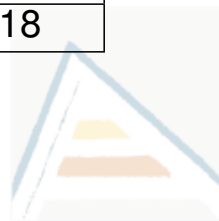
IMG	B6Australia_432.jpg
URL	http://australianageofdinosaurs.com/
CATEGORY	Science and Environment
COUNTRY	Australia
CONTINENT	Oceania
Time	0.045729
Bytes	33356
Images	11
Script_files	18
CSS_files	10
Tables	0
iframes	1
Style_tags	1
Img_bytes	291150
Img_width	1000
Img_height	1764

A small sample of the dataset, including one example of each category.

Web dataset collection

Table: Summary of statistical indicators for quantitative parameters.

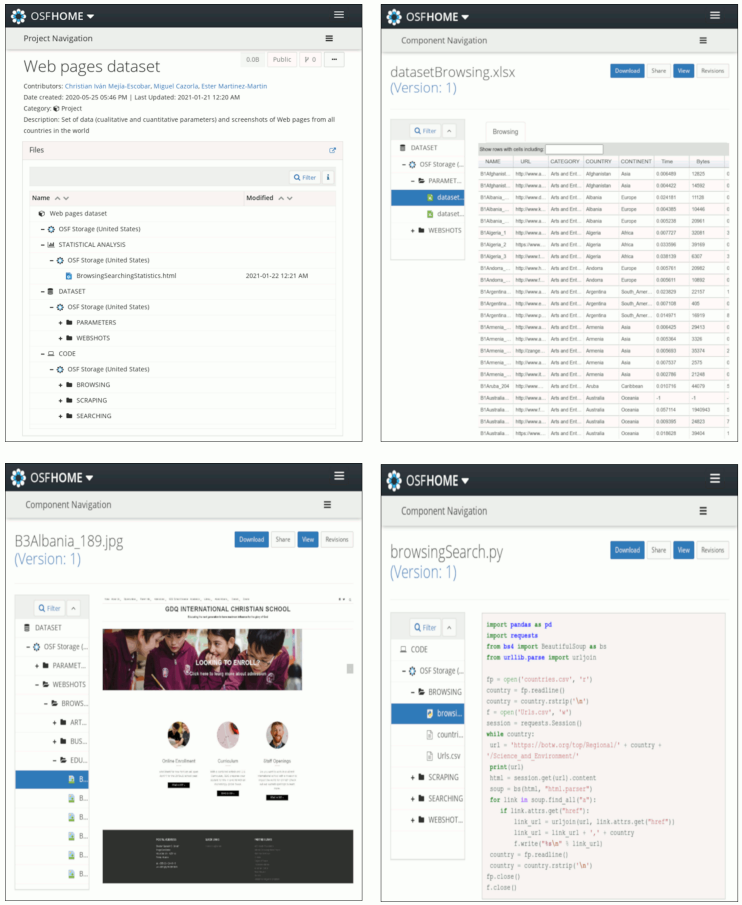
Parameter	Browsing				Searching			
	Min.	Max.	Mean	Std. Dev.	Min.	Max.	Mean	Std. Dev.
URL length	14	161	31.73	12.59	21	200	69.84	30.99
Time (ms)	1.25	72.9	18.72	17.39	1.9	73.3	22.33	15.83
Size (KB)	0	200.89	50.49	44.33	0	176.87	46.82	39.58
Images	0	15	2.05	3.48	0	17	2.51	4.14
Scripts	0	20	3.35	5.1	0	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



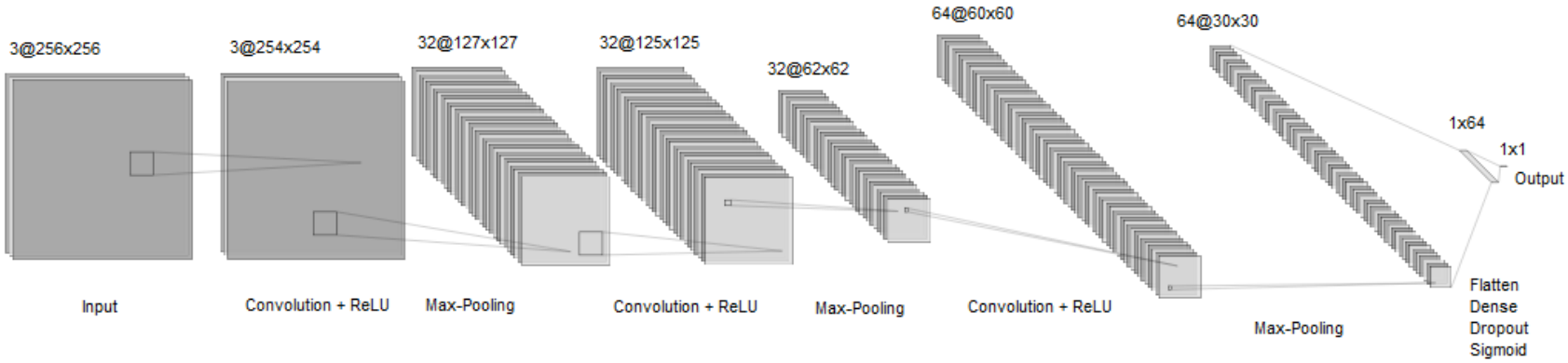
Web dataset collection

All products generated in this work are **public** and **available** through OSF (*Open Science Foundation*), a free and open platform to support, disseminate and enable collaboration of scientific research.

`https://osf.io/7ghd2/`

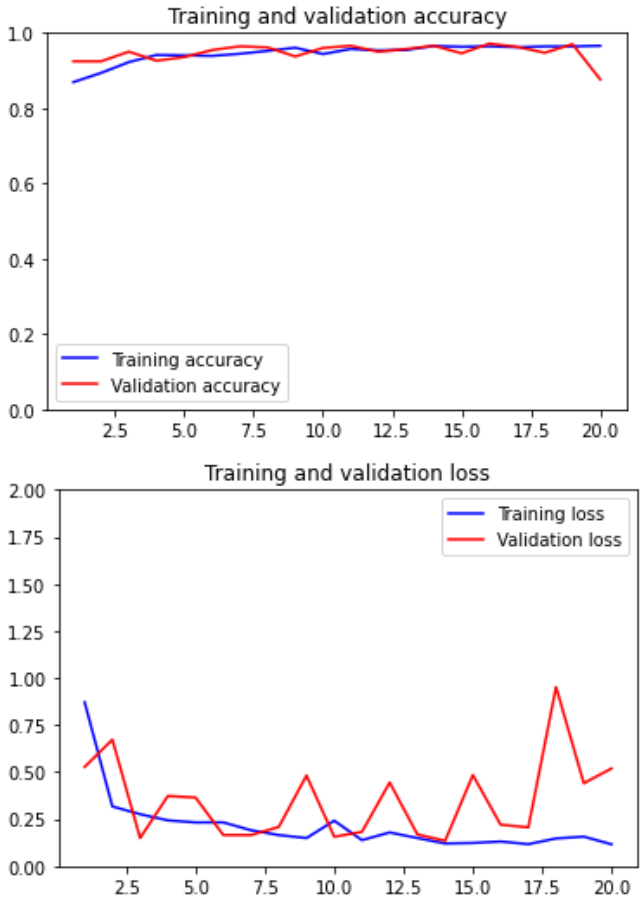


Detection of error Web pages (binary categorization)



The model's architecture is based on the convolutional neural network proposed by Liu et al.[16] to detect malicious Web sites.

Detection of error Web pages (binary categorization)



Detection of error Web pages (binary categorization)

Table: Dataset for binary classification (Browsing webshots).

	Browsing		
Category	**Webshots**	**Valid**	**Error**
Arts & Entertainment	447	397	50
Business & Economy	1058	892	166
Education	419	368	51
Government	730	669	61
News & Media	458	394	64
Science & Environment	497	462	35
Total	**3609**	**3182**	**427**

Table: Results of the binary Web categorization.

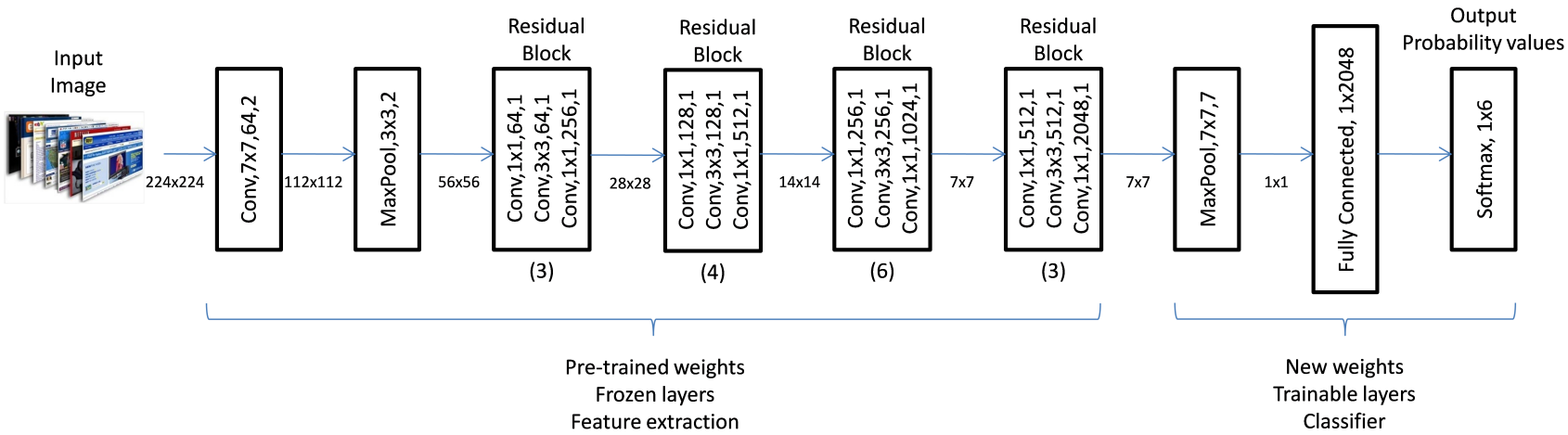
	Searching					
Category	Webshots	Valid (Prediction)	Error (Prediction)	Valid (Real)	Error (Real)	Accuracy
Arts & Entertainment	8569	7747	822	7355	1214	94.68%
Business & Economy	8699	8004	695	7546	1153	93.79%
Education	8742	8083	659	7524	1218	92.80%
Government	8088	7363	725	6685	1403	90.85%
News & Media	11574	10597	977	9650	1924	90.80%
Science & Environment	8893	8137	756	7496	1397	92.34%
Total	54565	49931	4634	46256	8309	92.47%

Detection of error Web pages (binary categorization)

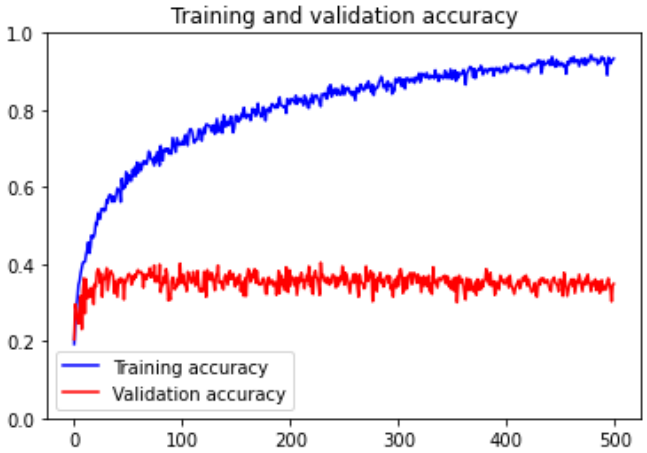
Table: Composition and size of the final dataset.

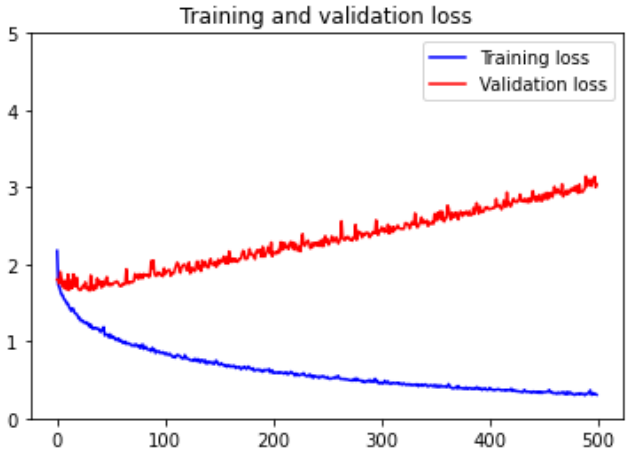
Category	**Browsing** Webshots	**Searching** Webshots	**Total**
Arts & Entertainment	397 (147 MB)	7355 (2.58 GB)	7752
Business & Economy	892 (300 MB)	7546 (2.48 GB)	8438
Education	368 (126 MB)	7524 (2.64 GB)	7892
Government	669 (253 MB)	6685 (2.47 GB)	7354
News & Media	394 (237 MB)	9650 (3.19 GB)	10044
Science & Environment	462 (193 MB)	7496 (2.63 GB)	7958
Total	**3182** (1.22 GB)	**46256** (15.99 GB)	**49438**

Multiclass categorization of Web pages



Multiclass categorization of Web pages





Multiclass categorization of Web pages

The model is correct in most cases for the categories of arts and entertainment, government, and news and media; however, the number of successes is low. This test takes 444 images, achieving an accuracy of 38.29%. For remaining categories, the model gets significantly confused. The classification of these categories is a **hard problem**. Nowadays, the composition of Web pages is becoming more complex, and the content has a high variability of visual features, even within the same category.

Confusion matrix

	Arts & Entertainment	Business & Economy	Education	Government	News & Media	Science & Environment
Arts & Entertainment	38	6	7	17	3	3
Business & Economy	16	20	1	25	6	6
Education	11	8	16	27	5	7
Government	6	4	9	46	5	4
News & Media	26	2	3	10	31	2
Science & Environment	17	7	4	21	6	19

Web dataset

- We were able to collect 3609 webshots from Browsing and 54565 from Searching, a total of 58174 webshots; however, the final dataset was reduced to 49438 due to the elimination of Web pages with error messages. This type of non-valid pages is equivalent to 15%, a significant value that reflects a problem on the Internet that affects webmasters, search engines, and users in general. To address this problem, more efficient debugging processes are needed. As an approach, we implemented an automatic detection of error Web pages achieving acceptable accuracy.
- The statistical analysis of the quantitative parameters: download time, size, number of images, scripts, CSS files, number of tables, iFrames, and style tags, show similar behavior. These variables have a very heterogeneous distribution, high variability, and tend to be strongly concentrated in the low values. This suggests that Web design follows an implicit rule of optimization of all these parameters, since the higher their values, the longer the download and display time of the page would increase, causing the consequent user's discomfort.

Web categorization

- According to the results, the automatic categorization of Web pages based exclusively on visual appearance (webshot) is a highly complex problem. Our dataset has proven to be difficult to classify; the difficulty increases when the categories cover a wide range of topics (like the case presented here). In addition, within each topic there is also a lot of variability in the visual aspect of the images.
- Although it is not possible to reliably distinguish between the categories in the dataset using only the webshots, the successes of the Deep Learning model for Web categorization, especially for government and, arts and entertainment categories, allow us to presume that it is feasible to identify distinctive visual patterns, which may be part of a next work. A high level of accuracy has not been achieved; however, it can serve as a baseline for future research. We suppose that increasing the dataset, preprocessing the images to have the same size and resolution, cropping and scaling the Web pages, could improve the results.

Future work

Our work can motivate the following developments:

- Extending the dataset provided, increasing the categories, URLs, webshots, and other qualitative and quantitative parameters.
- Achieving a higher level of automation in the process of collecting, organizing, and capturing webshots.
- Considering alternative URLs sources, that is, a search engine other than Google for Searching, and a Web directory other than BOTW for Browsing.
- Improving the accuracy achieved in multi-class categorization of Web pages with deeper convolutional neural networks.

References


V. De Boer, M. Van Someren, and T. Lupascu, "Classifying web pages with visual features," *WEBIST 2010 - Proceedings of the 6th International Conference on Web Information Systems and Technology*, vol. 1, pp. 245–252, 2010, doi: 10.5220/0002804102450252.


M. Du, Y. Han, and L. Zhao, "A Heuristic Approach for Website Classification with Mixed Feature Extractors," *IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, 2018, doi: 10.1109/ICPADS.2018.00028.


S. Lassri, H. Benlahmar, and A. Tragha, "Machine Learning for Web Page Classification: A Survey," *International Journal of Information Science & Technology (iJIST)*, vol. 3, no. 5, pp. 38-50, 2019.


X. Qi and B. D. Davison, "Web page classification: Features and algorithms," *ACM Computing Surveys*, vol. 41, no. 2, 2009, doi: 10.1145/1459352.1459357.


J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," *IEEE conference on computer vision and pattern recognition CVPR2009*, pp. 248-255, 2009. [Online]. Available: http://www.image-net.org/


K. Reinecke and K. Z. Gajos, "Quantifying visual preferences around the world," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI-14*, pp. 11–20, 2014, doi: 10.1145/2556288.2557052.


D. López-Sánchez, J. M. Corchado, and A. G. Arrieta, "A CBR system for image-based webpage classification: Case representation with convolutional neural networks," *FLAIRS 2017 -Proceedings of the 30th International Florida Artificial Intelligence Research Society Conference*, pp. 483–488, 2017.