data

MDPI

Data Descriptor

Ground Truth Dataset: Objectionable Web Content

Hamza H. M. Altarturi and Nor Badrul Anuar *

Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
* Correspondence: badrul@um.edu.my

Abstract: Cyber parental control aims to filter objectionable web content and prevent children from being exposed to harmful content. Succeeding in detecting and blocking objectionable content depends heavily on the accuracy of the topic model. A reliable ground truth dataset is essential for building effective cyber parental control models and validation of new detection methods. The ground truth is the measurement for labeling objectionable and unobjectionable websites of the cyber parental control dataset. The lack of publicly accessible datasets with a reliable ground truth has prevented a fair and coherent comparison of different methods proposed in the field of cyber parental control. This paper presents a ground truth dataset that contains 8000 labelled websites with 4000 objectionable websites and 4000 unobjectionable websites. These websites consist of more than 2 million web pages. Creating a ground truth objectionable web content dataset involved a few phases, including data collection, extraction, and labeling. Finally, the presence of bias, using kappa coefficient measurement, is addressed. The ground truth dataset is available publicly in the Mendeley repository.

Dataset: 10.17632/f239556fkr.2; https://data.mendeley.com/datasets/f239556fkr.

Dataset License: CC BY 4.0.

Keywords: objectionable dataset; web content; objectionable content; ground truth dataset; website category; web filtering

check for updates

Citation: Altarturi, H.H.M.; Anuar, N.B. Ground Truth Dataset: Objectionable Web Content. *Data* **2022**, *7*, 153. https://doi.org/10.3390/data7110153

Academic Editor: Han Woo Park

Received: 9 August 2022
Accepted: 18 October 2022
Published: 7 November 2022

1. Introduction

Children utilize the Internet to learn, entertain, and socialize. Even though the Internet is useful for children, certain activities increase the danger of cyberbullying [1]. Fewer parents believe Internet benefits outweigh the risks for children [2]. These reasons highlight the need for cyber parental controls when parenting children online. Cyber parental control aims to filter objectionable web content and prevent children from being exposed to harmful content. Objectionable websites are any websites that contain textual or visual content that certain internet users oppose on the web, including, but not limited to, pornography, violence, drugs, hate, racism, sexual, homicidality, gambling, and weapons [3]. Unobjectionable websites are any websites that do not contain any of the abovementioned objectionable contents.

The literature covers different parts of cyber parental control, including the psychological and legal implications, parents' roles, cyber network risks, and the role of technology [4]. In terms of the technological role of cyber parental control, the literature proposes several frameworks and models. Comparing these frameworks reveals their strengths and weaknesses and provides creative alternatives. However, due to a lack of publicly accessible datasets that provide verifiable ground truth conducting an objective and consistent comparison between the various frameworks that have been presented in the field of cyber parental control has been nigh on impossible.

This paper presents a ground truth dataset that contains 8000 labelled websites in the English language, with 4000 objectionable websites and 4000 unobjectionable websites. This ground truth dataset uses the JSON format to describe website attributes, making it easy to use in analytics and programming tools. It contains over 2 million scraped and labelled web pages with objectionable and unobjectionable content. The dataset was collected manually from several sources. The ground truth dataset is available publicly in the Mendeley repository.

2. Related Work

Current studies involving filtering objectionable web content have evaluated their models and frameworks based on inconsistent datasets. To address this issue, this study synthesized the current datasets that have been used in the state-of-the-art solution for objectionable web content. After that, this study investigated the availability and suitability of these datasets in the field of cyber parental control. Table 1 enumerates the used datasets in the current literature and describes the dataset and its limitations.

Table 1. The description and limitations of the used datasets in the cyber parental control field.

Reference	Limitation	Dataset Description
[5]	• It does not contain other objectionable websites.	• 228,848 URLs • 2 categories: safe and malicious
[6]	• It does not contain all objectionable and unobjectionable categories. It focuses only on hate and violent contents. • Not publicly available	• 80,000 URLs • 2 categories: hate and violence
[7]	• It does not contain objectionable and unobjectionable categories. It focuses on phish and legitimate websites. • Not publicly available	• 101,098 URLs 2 categories: legitimate and phishing
[8]	• It does not contain objectionable and unobjectionable categories. It focuses on phish and legitimate websites. • Not publicly available	• 73,575 URLs 2 categories: legitimate and phishing
[9]	• It does not contain objectionable and unobjectionable categories. It focuses on phish and legitimate websites. • Not publicly available	• 126,077 websites • 2 categories: legitimate and phishing
[10]	• It does not contain an objectionable category. It focuses on unobjectionable content. • The number of collected websites is not addressed. • Not publicly available	• 12 categories: adult, alcohol, gambling, tobacco, dating, drugs, hate, violence, weapon, religion, occults, and unknown
[11]	• It does not contain an objectionable category. • Not publicly available	• 140 websites 5 categories: science, academics, fiction, sports, and news
[12]	• It is a collection of Chinese textual documents, not URL or website contents.	• 35,500 documents from different websites 2 categories: objectionable and non-objectionable
[13]	• It contains only text sentences, and it is a mix of English and Chinese languages,	• 4290 sentences from different websites • 2 categories: objectionable and non-objectionable
[14]	• Not publicly available	• 92,560 URLs • 2 categories: kids and non-kids
[15]	• Not publicly available	• 2000 URLs • 2 categories: objectionable and non-objectionable
[16]	• It is manually collected and labelled. • Not publicly available	• 11,121 websites • 2 categories: normal and objectionable-related
[17]	• Not publicly available	• 65,000 URLs • 2 categories: blacklist and whitelist
[18]	• Not publicly available	• 300 URLs • 3 categories: deviant, suspicious, clean

As Table 1 shows, there is a lack of a standard dataset in the current web content filtering studies. Most studies design and build their dataset to suit their model or framework.

Moreover, a few studies created interesting datasets, such as those in [5–9]. However, these datasets focus only on a partial topic of the objectionable topics. For this reason, these datasets are not applicable to the field of cyber parental control. Table 1 also shows that only [14–18] created applicable datasets for the field of cyber parental control; however, none of these is publicly available. Given these factors, there is a need to create a ground truth dataset that contains objectionable and unobjectionable web content data.

3. Data Description

The ground truth dataset contains raw data (in a JSON format) of objectionable and unobjectionable websites. The ground truth dataset contains two files, an objectionable dataset file and an unobjectionable dataset file. Each file contains the exact number of attributes. This research selected the attributes based on similar previous datasets [19,20]. Most of these attributes were extracted with the help of Selenium and BeautifulSoup libraries [21,22]. Table 2 addresses the dataset's attributes and the data type and description.

3.1. Domain Metadata File

The dataset contains metadata.json. This file gives an overview of the websites and their features. The details of each field of this file are as follows:

Table 2. Description of the attributes of all websites of the objectionable ground truth dataset.

Attribute	Data Type	Description
domain	String	A code (D#) replacing the domain name of the website
geo_locs	String	Names of the countries based on the 'domain's IP Address location using GeoIP Databases [23]
domain_length	Numeric	Number of domain's characters
tld	String	Top-Level Domain (TLD) of the webpage using Tld Library [24]
avg_time_response	Numeric	The response time of webpage request in milliseconds
start_scrapping_timestamp	Numeric	The timestamp in milliseconds of scrapping the webpage
domain_tls_ssl_certificate	Numeric	Value 0 if the webpage does not use a certificate and 1 if the webpage uses a certificate
internal_urls_no	String list	
internal_urls	Numeric	
source	String	The collected source of the website
label	String	A categorical string of the webpage, either objectionable or unobjectionable

3.2. Internal Web Pages Detailed File

The dataset contains webpages_detail.json. This file gives detailed information on each collected website's web pages (internal URLs) and features. The details of each field of this file are as shown in Table 3:

Table 3. Description of the attributes of all web pages (URLs) of the objectionable ground truth dataset.

Attribute	Data Type	Description
url	String	A code (D#_URL#) replacing the URL of the webpage
domain_name	String	The code (D#) of the domain that the webpage belongs to
created_time	String	Time created the record (format yyyy-MM-dd HH:mm:ss)
geo_loc	String	Name of the country based on the 'webpage's IP Address location using GeoIP Databases [23]
domain_length		
url_length	Numeric	Number of URL characters
time_response	Numeric	The response time of webpage request in milliseconds
html_char_length	Numeric	Number of characters in the full HTML

Table 3. *Cont.*

Attribute	Data Type	Description
text_char_length	Numeric	Number of characters in all visible texts
textual_tags_cnt	Numeric	Number of the list of all visible texts on the webpage
visual_content_no	Numeric	Number of the list of all visuals on the webpage
label	String	A categorical string of the webpage, either objectionable or unobjectionable
label_details	String	A sub-categorical string of the webpage, including but not limited to porn, gambling, erotica, sport, news, kids, etc.,
tld	String	Top-Level Domain (TLD) of the webpage using Tld Library [24]
protocol	String	Name of the protocol used by the webpage URL (http, https, ftp, etc.,)
tls_ssl_certificate	Numeric	False if the webpage does not use a certificate true if the webpage uses a certificate
source	String	The collected source of the website

4. Data and Methods

Researchers use two methods to create a website ground truth dataset. The first method is manual collection and inspection, which is time, cost, and resource consuming. This method suits a small amount of data but is impractical and might fail on large datasets. The second method is to label websites using blacklisting and whitelisting services, such as Alexa, DOMZ, and Google SafeBrowsing [25]. These services, however, limit their API, making it impossible to label a massive amount of data. Taken together, the methodology of creating the ground truth dataset in this paper adopted both methods and involved 3 phases. These phases were data collection, extraction, and labeling, in which many studies were used for creating web content datasets [19,26,27].

4.1. Web Pages Collection

This study collected websites from the Alexa dataset, search engines (Yandex, Google, Yahoo), and external webpages links. Each source categorized the websites into different topic categories. Based on the source categorization, this study classified the collected websites as either objectionable or unobjectionable. For the search engines, this study classified the collected websites from the search engine, based on the used keywords in the search query. For example, the collected websites using the keywords "porn", "erotic", "gambling", etc., were classified as objectionable. Table 4 shows the sources of the collected website.

Table 4. Sources of the categorized websites in the ground truth dataset.

Source	Objectionable Sites	Unobjectionable Sites	Total
Alexa	0	1500	1500
DOMZ	1500	1000	2500
Google	500	500	1000
Yandex	500	500	1000
Yahoo	500	500	1000
Internal links	1000	0	1000
Total	4000	4000	8000

4.2. Web Pages Content Extraction

Extracting website content required crawling each web page and then scraping it and parsing its content. Web crawling aimed to index the entire web pages contained in a specific website by systematically browsing the web. The scrapping of HTML code extracts relevant to web page contents, such as paragraphs, images, bold texts, web page titles, and metadata, was addressed.

Although there are several ways to crawl and scrape a website, Python offers a flexible and powerful way to do it. A few Python libraries support web crawling and scraping, such as BeautifulSoup, LXML, MechanicalSoup, Requests, Scrapy, and URLLib. Building an automatic and systematic website crawler and scrapper required using a combination of these libraries. The following pseudo-code illustrates the algorithm for web content extraction used in this paper.

The source code of this task is available publicly in the GitHub repository under a library called CrawlScrape [9]. CrawlScrape is an open-source Python library for the solution of efficient and easy web crawling and data scraping for dataset collection.

4.3. Labeling

This step aimed to label the collected websites based on their source categorization, features classification, and extracted topic classification. The extracted topic was classified as either objectionable or unobjectionable. There is a lack of agreement on the definition of "objectionable content" in the literature [3]. This study conceptualized objectionable web content terms as textual content that children users oppose on the web, including, but not limited to, pornography, violence, drugs, hate, racism, sexual, homicide, gambling, and weapons. The ground truth dataset labelled the content of web pages based on this definition as objectionable or unobjectionable.

5. Presence of Bias Results

In order to reduce the bias in the ground truth dataset to a specific source, this phase used several sources to collect the ground truth dataset. These resources were Alexa, DMOZ, Yandex, Google, and Yahoo. Furthermore, we randomly chose 1600 websites, representing 20% of the total number of websites in the dataset, and labelled them manually as objectionable and unobjectionable. Five people experienced in content classification and categorization were selected to do this task. In this way, we aimed to demonstrate the presence of selection bias in any of the sources. The Kappa coefficient was then applied to compare the manual labels of the randomly selected 1600 websites with the original labels from the source. The following equations were used to calculate the agreements of the manual and source labels:

$$\text{observed agreement} = \mathrm{A} + \mathrm{D} \tag{1}$$

$$\text{expected agreement} = \frac{((\mathrm{A}+\mathrm{B}) \times (\mathrm{A}+\mathrm{C})) + ((\mathrm{C}+\mathrm{D}) \times (\mathrm{B}+\mathrm{D}))}{\mathrm{n}} \tag{2}$$

$$\text{kappa} = \frac{\text{observed agreement} - \text{expected agreement}}{\mathrm{n} - \text{expected agreement}} \tag{3}$$

where

A: number of agreements on first label
B: number of no agreements on the first label
C: number of no agreements on the second label
D: number of agreements on the second label
n: number of dataset records

Kappa Coefficient Inspection

We calculated the agreements of the manual and source labels for the randomly selected websites by using the Kappa coefficient. Kappa Cohen's coefficient is "a statistical measure of inter-rater reliability or agreement used to assess qualitative documents and determine the agreement between two raters". Kappa coefficient comparing of the human (manual) and source (automatic) labeling of 20% of the websites in the ground truth dataset was 0.87 (calculations in Table 5), indicating very high agreement, and, thus, low selection bias.

Table 5. Kappa agreement table and equation results of the ground truth dataset.

		Human (Manual) Classification		
		Objectionable	**Unobjectionable**	**Subtotal**
Source (automatic) classification	Objectionable	730	70	800
	Unobjectionable	10	790	800
	Subtotal	740	860	1600

Observed agreement = 1520
Expected agreement = ((800 × 740) + (800 × 790))/1600 = 765
Kappa score = (1520 − 765)/(1600 − 765) = 0.904
Kappa score > 0.904 (almost perfect agreement between human classification and ground truth classification).

Author Contributions: Conceptualization, H.H.M.A.; methodology, H.H.M.A.; software, H.H.M.A.; validation, N.B.A.; formal analysis, H.H.M.A.; investigation, H.H.M.A.; resources, H.H.M.A.; data curation, H.H.M.A.; writing—original draft preparation, H.H.M.A.; writing—review and editing, N.B.A.; visualization, H.H.M.A.; supervision, N.B.A.; project administration, N.B.A.; funding acquisition, N.B.A. All authors have read and agreed to the published version of the manuscript.

Funding: The work of the authors was supported by the Impact-oriented Interdisciplinary Research Grant (IIRG), Universiti Malaya under grant IIRG001C-2020SAH.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Supplementary material associated with this article can be found in the online version at https://doi.org/10.17632/f239556fkr.2 (accessed on 8 September 2022) (https://data.mendeley.com/datasets/f239556fkr) (accessed on 8 September 2022).

Conflicts of Interest: The author declares that he has no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this paper.

References

1. Sasson, H.; Mesch, G. Parental mediation, peer norms and risky online behavior among adolescents. *Comput. Hum. Behav.* **2014**, *33*, 32–38. [CrossRef]
2. Ofcom. Children and Parents: Media Use and Attitudes Report 2018. Available online: https://www.ofcom.org.uk/__data/assets/pdf_file/0024/134907/children-and-parents-media-use-and-attitudes-2018.pdf (accessed on 24 November 2019).
3. Altarturi, H.; Saadoon, M.; Anuar, N.B. Cyber parental control: A bibliometric study. *Child. Youth Serv. Rev.* **2020**, *116*, 105134. [CrossRef]
4. Altarturi, H.H.; Anuar, N.B. A preliminary study of cyber parental control and its methods. In Proceedings of the 2020 IEEE Conference on Application, Information and Network Security (AINS), Kota Kinabalu, Malaysia, 17–19 November 2020; pp. 53–57.
5. Altay, B.; Dokeroglu, T.; Cosar, A. Context-sensitive and keyword density-based supervised machine learning techniques for malicious webpage detection. *Soft Comput.* **2019**, *23*, 4177–4191. [CrossRef]
6. Liu, S.; Forss, T. New classification models for detecting Hate and Violence web content. In Proceedings of the 2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Lisbon, Portugal, 2–14 November 2015; pp. 487–495.
7. Marchal, S.; François, J.; State, R.; Engel, T. PhishStorm: Detecting phishing with streaming analytics. *IEEE Trans. Netw. Serv. Manag.* **2014**, *11*, 458–471. [CrossRef]
8. Sahingoz, O.K.; Buber, E.; Demir, O.; Diri, B. Machine learning based phishing detection from URLs. *Expert Syst. Appl.* **2019**, *117*, 345–357. [CrossRef]
9. Rao, R.S.; Vaishnavi, T.; Pais, A.R. CatchPhish: Detection of phishing websites by inspecting URLs. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 813–825. [CrossRef]
10. Kotenko, I.; Chechulin, A.; Shorov, A.; Komashinsky, D. Analysis and evaluation of web pages classification techniques for inappropriate content blocking. In *Advances in Data Mining: Applications and Theoretical Aspects, Proceedings of The 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, 16–20 July 2014*; Springer: Cham, Switzerland, 2014; pp. 39–54.
11. Narwal, N. Web page filtering for kids. *Int. J. Inf. Technol.* **2021**, *13*, 19–25. [CrossRef]

12. Zeng, J.; Duan, J.; Wu, C. Adaptive Topic Modeling for Detection Objectionable Text. In Proceedings of the 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Atlanta, GA, USA, 17–20 November 2013; pp. 381–388.
13. Duan, J.; Zeng, J. Web objectionable text content detection using topic modeling technique. *Expert Syst. Appl.* **2013**, *40*, 6094–6104. [CrossRef]
14. Rajalakshmi, R.; Tiwari, H.; Patel, J.; Kumar, A.; Karthik, R. Design of Kids-specific URL Classifier using Recurrent Convolutional Neural Network. *Procedia Comput. Sci.* **2020**, *167*, 2124–2131. [CrossRef]
15. Patel, O.; Tiwari, A.; Patel, V.; Gupta, O. Quantum based neural network classifier and its application for firewall to detect malicious web request. In Proceedings of the 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, South Africa, 7–10 December 2015; pp. 67–74.
16. Zhao, C.; Zhang, Y.; Zang, T.; Liang, Z.; Wang, Y. A Stacking Approach to Objectionable-Related Domain Names Identification by Passive DNS Traffic (Short Paper). In Proceedings of the International Conference on Collaborative Computing: Networking, Applications and Worksharing, Shanghai, China, 1–3 December 2018; pp. 284–294.
17. Hussain, M.; Ahmed, M.; Khattak, H.A.; Imran, M.; Khan, A.; Din, S.; Ahmad, A.; Jeon, G.; Reddy, A.G. Towards ontology-based multilingual URL filtering: A big data problem. *J. Supercomput.* **2018**, *74*, 5003–5021. [CrossRef]
18. Zamry, N.M.; Maarof, M.A.; Zainal, A. Islamic Web Content Filtering and Categorization on Deviant Teaching. In *Recent Advances on Soft Computing and Data Mining, Proceedings of The First International Conference on Soft Computing and Data Mining (SCDM-2014), Johor, Malaysia, 16–18 June 2014*; Springer: Cham, Switzerland, 2014; pp. 667–678.
19. Singh, A. Malicious and benign webpages dataset. *Data Brief* **2020**, *32*, 106304. [CrossRef]
20. Vrbančič, G.; Fister, I., Jr.; Podgorelec, V. Datasets for phishing websites detection. *Data Brief* **2020**, *33*, 106438. [CrossRef]
21. Selenium for Python. Available online: https://pypi.org/project/selenium (accessed on 1 March 2022).
22. BeautifulSoup Library. Available online: https://pypi.org/project/beautifulsoup4 (accessed on 1 March 2022).
23. GeoIP Database. Available online: https://geolocation-db.com (accessed on 1 March 2022).
24. Tld Library. Available online: https://pypi.org/project/tld (accessed on 1 March 2022).
25. Chen, C.; Zhang, J.; Chen, X.; Xiang, Y.; Zhou, W. 6 million spam tweets: A large ground truth for timely Twitter spam detection. In Proceedings of the 2015 IEEE International Conference on Communications (ICC), London, UK, 08–12 June 2015; pp. 7065–7070.
26. Khalil, A.; Jarrah, M.; Aldwairi, M.; Jaradat, M. AFND: Arabic fake news dataset for the detection and classification of articles credibility. *Data Brief* **2022**, *42*, 108141. [CrossRef]
27. Ashouri, S.; Suominen, A.; Hajikhani, A.; Pukelis, L.; Schubert, T.; Türkeli, S.; Van Beers, C.; Cunningham, S. Indicators on firm level innovation activities from web scraped data. *Data Brief* **2022**, *42*, 108246. [CrossRef] [PubMed]