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.2

Dataset License: CC BY 4.0.

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

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

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

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

<table>
<thead>
<tr>
<th>Source</th>
<th>Objectional Sites</th>
<th>Unobjectional Sites</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexa</td>
<td>0</td>
<td>1500</td>
<td>1500</td>
</tr>
<tr>
<td>DOMZ</td>
<td>1500</td>
<td>1000</td>
<td>2500</td>
</tr>
<tr>
<td>Google</td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>Yandex</td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>Yahoo</td>
<td>500</td>
<td>500</td>
<td>1000</td>
</tr>
<tr>
<td>Internal links</td>
<td>1000</td>
<td>0</td>
<td>1000</td>
</tr>
<tr>
<td>Total</td>
<td>4000</td>
<td>4000</td>
<td>8000</td>
</tr>
</tbody>
</table>

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 scraping 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:

\[
\begin{align*}
\text{observed agreement} & = A + D \\
\text{expected agreement} & = \frac{(A + B) \times (A + C) + (C + D) \times (B + D)}{n} \\
\text{kappa} & = \frac{\text{observed agreement} - \text{expected agreement}}{n - \text{expected agreement}}
\end{align*}
\]

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.

<table>
<thead>
<tr>
<th>Source (automatic classification)</th>
<th>Human (Manual) Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Objectionable</td>
</tr>
<tr>
<td>Objectionable</td>
<td>730</td>
</tr>
<tr>
<td>Unobjectionable</td>
<td>10</td>
</tr>
<tr>
<td>Subtotal</td>
<td>740</td>
</tr>
</tbody>
</table>

Observed agreement = 1520
Expected agreement = \((800 \times 740) + (800 \times 790)/1600 = 765\)
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](https://doi.org/10.17632/f239556fkr.2) (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


19. Singh, A. Malicious and benign webpages dataset. Data Brief 2020, 32, 106304. [CrossRef]

20. Vrbančić, G.; Fister, I., Jr.; Podgorelec, V. Datasets for phishing websites detection. Data Brief 2020, 33, 106438. [CrossRef]


