INTERNATIONAL JOURNAL
ON INFORMATICS VISUALIZATION

journal homepage : www.joiv.org/index.php/joiv

Data Pre-processing of Website Browsing Records: To Prepare Quality Dataset for Web Page Classification

Siti Hawa Apandi [a,*], Jamaludin Sallim [a], Rozlina Mohamed [a], Norkhairi Ahmad [b]

[a] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, 26600, Malaysia*
[b] *Student Development Section, Universiti Kuala Lumpur Malaysia France Institute, Bandar Baru Bangi, Selangor, 43650, Malaysia*
*Corresponding author: *sitihawa.apandi@gmail.com*

Abstract*— The increased usage of the internet worldwide has led to an abundance of web pages designed to supply information to internet users. The use of web page classification is becoming increasingly necessary to organize the growing number of web pages. This classification model serves as a tool to restrict internet usage to specific categories of web pages. To develop the classification model, it's crucial to check the quality of the dataset, as it determines the performance of the web page classification model. Raw datasets are typically unreliable and subject to noise, which complicates data analysis. This is why data pre-processing is necessary to prepare the dataset properly. In this study, website browsing records serve as the dataset. The primary goal of this paper is to investigate data pre-processing techniques for website browsing records, focusing on Game and Online Video Streaming web pages. Data pre-processing involves two main steps: data cleaning and web content pre-processing. After completing the data cleaning process, the datasets are reduced from the original. This demonstrates that many datasets can be eliminated due to their inactivity or unsuitability as the datasets for Game and Online Video Streaming web pages. Meanwhile, web content pre-processing removes noise from an HTML document, retaining only relevant words that can represent the web page by creating a word cloud image. Convolutional Neural Networks (CNN) will be used to construct a model for categorizing web pages to determine whether they fall under Game or Online Video Streaming. The pre-processed data will be used as the input for this model.

***Keywords*— Data cleaning; data pre-processing; web content pre-processing; web page classification; website browsing records.**

Manuscript received 31 Jan. 2023; revised 22 Aug. 2023; accepted 20 Oct. 2023. Date of publication 31 Mar. 2024.

I. INTRODUCTION

The internet has made it easy for people to access information through web pages. As a result, the number of web pages providing services to users has grown significantly and continues to increase. The abundance of information available on the internet has presented the challenge of efficiently organizing and managing this information [1]. Consequently, information retrieval has become a difficult task [2]. This problem can be addressed through web mining, which involves using techniques to locate and gather valuable information from the internet [2]–[4]. Web mining is a form of data mining that focuses explicitly on data from the internet, aiming to discover and identify patterns within this web data [5]. The initial step in web mining involves categorizing web pages into different classes using one data mining technique: classification [6].

Web page categorization offers valuable data for efficient internet browsing, spam filtering, and numerous other applications. Search engines face the significant challenge of quickly locating relevant results among many websites. As a result, many search engines have resorted to categorizing web pages by topic to enhance the quality of search results they deliver to users. Moreover, web page categorization is essential for establishing internet usage guidelines for businesses and private users. Cybersecurity software can also leverage web page classification to prevent users from accessing harmful websites by blocking them before they load [7].

During the COVID-19 outbreak, the Malaysian Communications and Multimedia Commission (MCMC) conducted the Internet User Survey 2020 (IUS 2020) in collaboration with an independent survey house. The survey aimed to track online activity and identify user trends and behaviors related to internet use in Malaysia. The IUS 2020 study gathered information from 2,401 internet users and 384 non-users nationwide. Additionally, 384 internet users were included from each state for state-level surveys. Respondents were randomly selected, and interviews were conducted over the phone [8].

According to the survey, there was an 88.7% increase in internet users in 2020 compared to 2018, representing a 1.3% rise. The most prevalent age groups among internet users were adults in their 20s (46.0%) and 30s (21.2%). In 2020, internet users spent more time online, with a 13% increase from the 37% reported in 2018. Specifically, 50% of users went online for five to 12 hours daily. Internet users engage in a variety of online activities. They commonly use the internet for entertainment, such as surfing online video streaming and gaming websites. This trend is evident in the statistics for these two activities. Firstly, the percentage of internet users watching or downloading videos online increased from 77.6% in 2018 to 87.3% in 2020. Secondly, in 2020, 42.8% of internet users played online games, an increase from 35.2% in 2018 [8].

The increase in internet usage for online activities, especially surfing online video streaming and gaming websites, has the potential to lead to a severe problem: internet addiction. “Internet addiction” can be defined as “an individual’s inability to control their internet usage, which eventually leads to psychological, social, academic, and work difficulties in their life” [9], [10]. College and university students, often in their 20s, are particularly vulnerable to internet addiction due to the attractions found on the web, contributing to a significant portion of addiction cases among internet users. Engaging in online activities can distract students from the learning environment. When examining internet addiction among college students, researchers have found that excessive internet use often leads to difficulties in completing homework and assignments, preparing for exams, or getting enough rest to attend early morning classes. These academic issues disrupt the students’ daily routines [11]. In one study, 50% of students asked about their academic failures leading to expulsion attributed their problems to excessive internet usage [12].

Colleges and universities can play a crucial role in protecting students from internet addiction. They should monitor the online browsing behaviors of students who use the internet connection provided by their institutions. One common approach for college and university administrators is establishing guidelines and restricting internet access to specific categories of web pages. Before implementing these restrictions, web pages must be categorized into their respective categories using a web page classification model.

There are two approaches to classifying web pages: the traditional manual and the automatic. In the traditional manual approach, an expert must manually assign each web page to a category, which is labor-intensive and time-consuming. Given the ever-growing number of web pages today, this manual method is impractical for classification [13], [14]. Therefore, the automatic approach is the most suitable for handling the vast quantity of web pages requiring classification.

In the automatic approach to classifying web pages, the web page classifier is responsible for assigning category labels to web pages. Machine learning has been used as the web page classifier and has demonstrated exemplary performance when dealing with several web pages. However, its effectiveness diminishes when processing large web pages [15]. Deep learning has been introduced as the web page classifier to overcome the limitations of machine learning, especially when handling complex and extensive datasets [16]. Deep learning offers the advantage of automatically discovering the features for categorization, whereas machine learning typically requires manual feature selection [17].

Researchers have proposed various web page classification methods, each utilizing techniques to improve classifier accuracy. Most classification algorithms' accuracy depends on the quantity and quality of training data, which relies on the document representation technique employed [13]. Web pages contain various data types, including text, images, audio files, and videos [4], [7], [18]. Additionally, web pages increasingly incorporate irrelevant data, such as advertising posters, navigation bars, hyperlinks, and headers/footers. This irrelevant data, often referred to as noisy data, can significantly disrupt feature extraction and reduce classification accuracy [4], [17], [19]. Gathering the necessary data from web pages to select the most representative features becomes challenging. However, this issue can be addressed by cleaning up the data and performing data pre-processing to remove irrelevant information [4].

Raw datasets cannot be directly provided to classification algorithms and are expected to be trained effectively. Data pre-processing is essential to convert raw datasets into a format that classification algorithms can understand and use to analyze data features. Data pre-processing encompasses various stages, including data cleaning, integration, transformation, and reduction. Table I provides descriptions of these data pre-processing stages [20], [21].

TABLE I
DATA PRE-PROCESSING STAGES

	Data Pre-processing Stages	Description
1.	Data cleaning	Identifying erroneous or noisy data and rectifying or removing them from the dataset.
2.	Data integration	The merging or integrating data from multiple sources creates a unified and coherent dataset.
3.	Data transformation	Raw data is typically transformed into a format that can be used for analysis. To ensure that the raw data is on the same scale, a common step is normalizing it.
4.	Data reduction	A method for reducing the size of the initial data volume and representing it in a more compact form. This process enhances the representation of input data without compromising its integrity.

Web pages contain semi-structured data, which includes tags used to structure and organize the information presented through a web browser. This semi-structured nature makes the categorization of web pages different from the categorization of plain text [7], [13], [22], [23]. Web content pre-processing is essential to extract useful information from web pages. The pre-processing steps typically involve tokenization, normalization, and noise removal. Refer to Table II for a detailed description of these steps in web content pre-processing [21].

TABLE II
WEB CONTENT PRE-PROCESSING

	Steps in Web Pre-processing	Description
1.	Tokenization	The practice of dividing lengthy texts into smaller units, such as breaking sentences into words. It is also known as text segmentation or lexical analysis.
2.	Normalization	It involves a series of connected tasks aimed at standardizing all words. These tasks include stemming, lemmatization, capitalizing or lowercasing letters, removing punctuation and unnecessary words, and replacing numbers with words.
3.	Noise removal	The process of stripping away HTML, XML, and other markup and metadata.

Data pre-processing may be time-consuming, but the resulting dataset is expected to be accurate and ready for analysis by classification algorithms, which are used to build models for classifying web pages. Data pre-processing plays a crucial role in web page classification by enhancing dataset quality by removing unnecessary and noisy data. This, in turn, positively impacts the accuracy, speed, and ability of the web page classifier to avoid overfitting issues [13], [20], [21].

This paper aims to investigate the process of data pre-processing for website browsing records, with a specific focus on collecting data from Game and Online Video Streaming web pages. This focus is driven by our primary objective of developing a model capable of classifying these types of web pages. The initial step involves executing data pre-processing on the raw dataset, which includes data cleaning to retrieve active web pages and download the HTML source code of web pages with English content. Subsequently, the next stage in data pre-processing is web content pre-processing, aimed at extracting meaningful words from the HTML source code of web pages by eliminating noisy data.

The remaining portions of this paper are arranged as follows: The dataset of website browsing records is described in Section II, along with the study's methodology. The results and discussion on pre-processing website browsing records in Section III follow this. Lastly, the study's conclusion and future work are presented in Section IV.

II. MATERIALS AND METHOD

Due to the unavailability of universally acknowledged datasets, this study utilizes self-gathered datasets that reflect real-world scenarios. Specifically, the dataset consists of the website browsing records of students at Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA). These records provide a good representation of realistic user behaviors when surfing the internet [24]. The website browsing records of UMPSA students are made available by the Center of Information Technology and Communication (PTMK), an organization within UMPSA that monitors internet usage among UMPSA users. They keep records of the URL web pages that are accessed.

The record of website browsing contains information about URL web pages accessed by students for one week, from 2019-03-17 (Sunday) 00:00:00 to 2019-03-23 (Saturday) 23:59:59. PTMK has provided 40 Microsoft Excel files, each containing the browsing records for one UMPSA student. In total, there are records for 40 students. Each Microsoft Excel file includes data on the URLs of web pages visited, and the number of URLs in each file exceeds one thousand. This indicates that each student accessed thousands of web pages within a week. The description of the information contained on the website browsing files is presented in Table III. This study only utilized information related to the URL category and the URLs.

TABLE III
TYPES OF INFORMATION IN THE WEBSITE BROWSING RECORD FILE

	Column Names	Column Descriptions
1.	Rank	The record ID numbers within the website browsing data follow a pattern where the latest record has a smaller ID number while the older records have more significant ID numbers. This numbering scheme helps distinguish the chronological order of website browsing records.
2.	Username	Username of student.
3.	Group name	Group of users: student.
4.	Source IP	Source IP of student.
5.	Endpoint device	The device used by the student.
6.	Location	Location of student
7.	Dst IP	Destination IP of web page.
8.	URL category	Category of URL.
9.	Title	Title of web page.
10.	Domain	Domain of URL.
11.	URL	URL of web page.
12.	Action	Log.
13.	Time	Date and time of URL being accessed.
14.	Details	Details about the domain include DNS, endpoint details, src port, port, protocol, and MAC.

The URL web pages in the Excel file of the website browsing record can be categorized as supervised because they have been classified into URL categories. There are 53 URL categories for the URL web pages in the Excel file of the website browsing record that FortiGuard has assigned. It is a tool that the PTMK employs to categorize the URL category for a URL web page.

Fig. 1 shows the process of developing the proposed web page classification model, which consists of five stages: data collection, data pre-processing, obtaining the required features, building the web page classification model, and evaluating the web page classification model. This paper will focus only on the two early stages: data collection and data pre-processing.

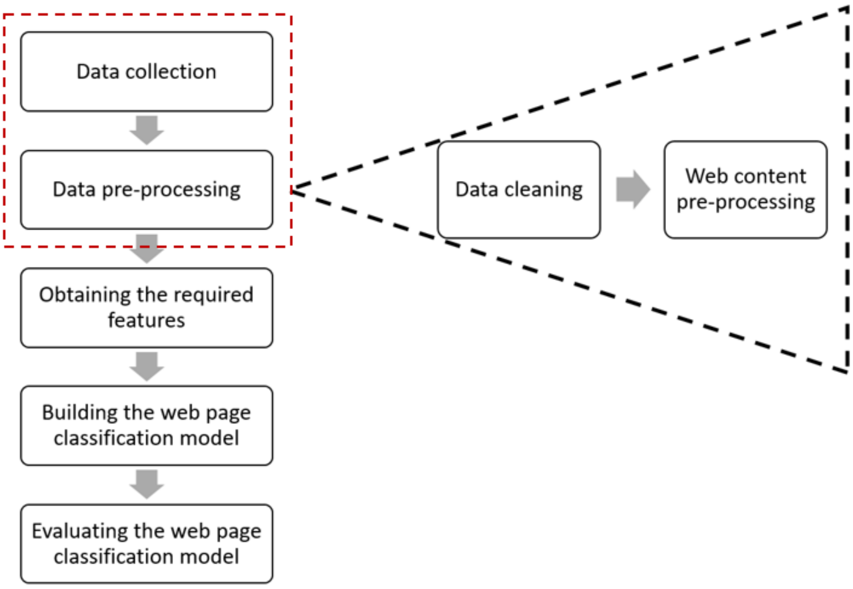

Data collection
Data pre-processing
Obtaining the required features
Building the web page classification model
Evaluating the web page classification model
Data cleaning
Web content pre-processing

Fig. 1 Process to develop the proposed web page classification model

Table IV shows the specification of the experimental environment used in this study to run the experiment for the data pre-processing website browsing records.

TABLE IV
SPECIFICATION OF EXPERIMENTAL ENVIRONMENT

Specification	Description
Operating System (OS)	Windows 10
Central Processing Unit (CPU)	Intel(R) Core (TM) i7-8750H 2.20GHz
Graphics Processing Unit (GPU)	Nvidia GeForce GTX 1060
Memory	16GB
Programming language	MATLAB

III. Results and Discussion

The raw dataset URLs, derived from website browsing records, are typically not structured for analysis, as they might not be complete, consistent, or easy to comprehend. The most challenging step is to find and extract relevant data from the dataset [21]. Thus, data pre-processing is a crucial step for cleaning, correcting, and preparing input data for mining [25]. This study's data pre-processing consists of data cleaning and web content pre-processing.

A. Data Cleaning

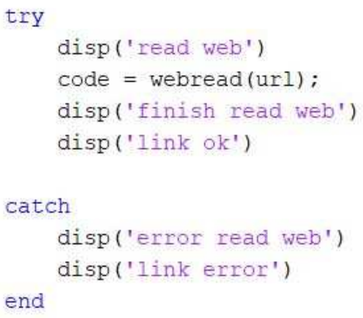
The raw dataset's URL web pages must undergo data cleaning as the initial step. The primary goal of data cleaning is to eliminate redundant, useless, erroneous, incomplete, and inconsistent data [25]. It is also employed to retrieve cleaned data of active web pages [26]. In this study, data cleaning is executed to achieve several specific tasks, including removing duplicate URLs, fetching active URLs, retaining URLs containing HTML tags in the web page source code, and including web pages with English content. Refer to Fig. 3 for an illustration of the data-cleaning steps. The steps of data cleaning are described below.

- Remove records that contain null URLs from the Excel file of website browsing, as shown in Table V.
- Keep the unique URL by removing the duplication of the URL.

There are 541,702 URLs in the raw dataset overall. After eliminating the records containing null URLs, there are 538,983 URLs. Then, the records with duplicate URLs were removed, and the total number of URLs became 79,661. The data cleaning procedure by removing duplicate URLs is performed again on those records. This is because the URL records in each Microsoft Excel file of the website browsing record may have duplicate URLs. Thus, the total number of unique URL records is 29,444.

- Check whether the URL is active or inactive.

The inactive URL means the link has been broken, and the web page cannot be accessed anymore. Thus, the inactive URL records are removed, and only active URL records are kept. A MATLAB code has been built to identify whether the URL records are active or inactive, as shown in Fig. 2.



```
try
    disp('read web')
    code = webread(url);
    disp('finish read web')
    disp('link ok')

catch
    disp('error read web')
    disp('link error')
end
```

Fig. 3 Code to identify whether the URL is an active or inactive link

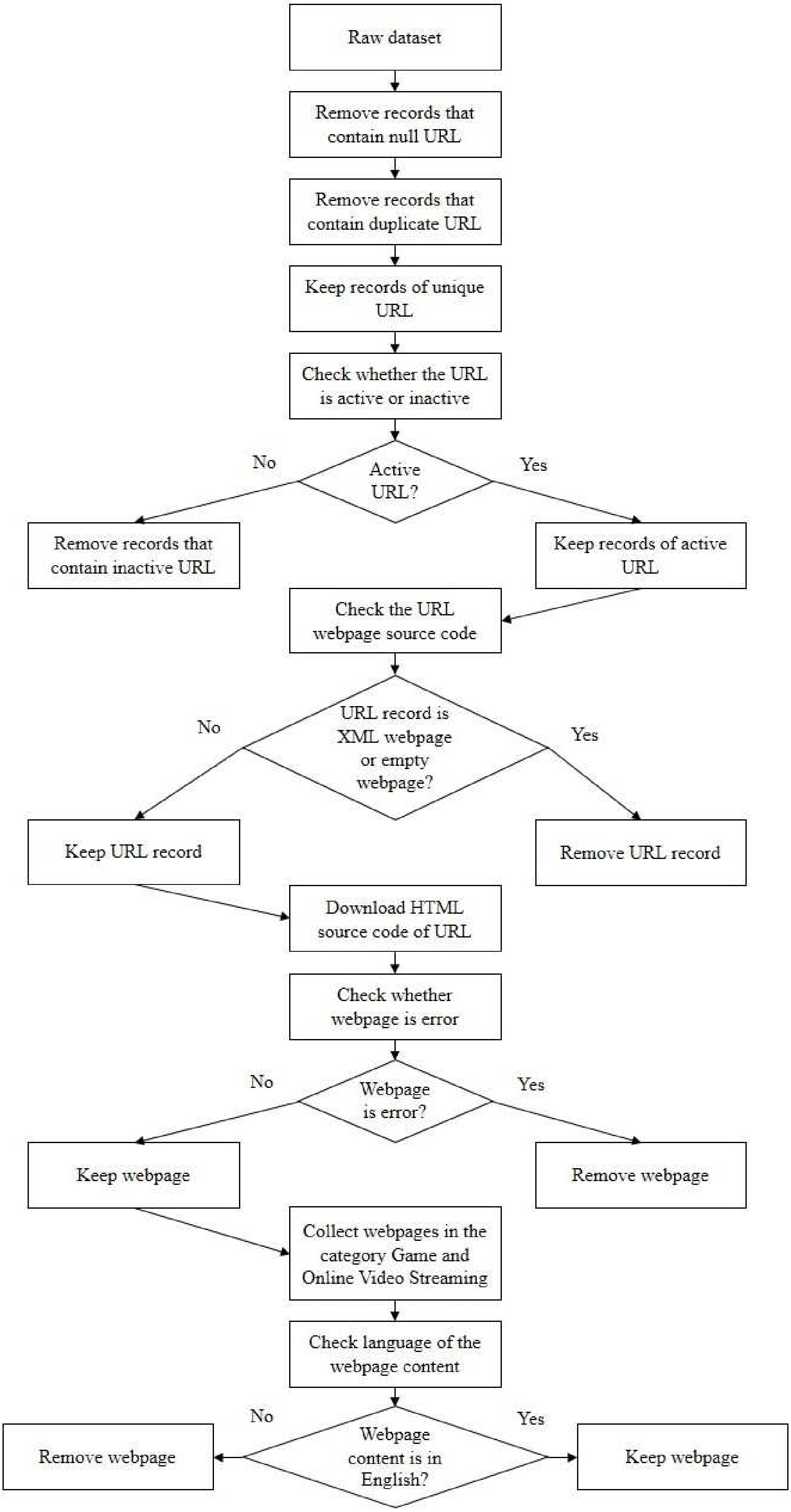

Raw dataset
Remove records that contain null URL
Remove records that contain duplicate URL
Keep records of unique URL
Check whether the URL is active or inactive
Active URL?
No
Yes
Remove records that contain inactive URL
Keep records of active URL
Check the URL webpage source code
URL record is XML webpage or empty webpage?
No
Yes
Keep URL record
Remove URL record
Download HTML source code of URL
Check whether webpage is error
Webpage is error?
No
Yes
Keep webpage
Remove webpage
Collect webpages in the category Game and Online Video Streaming
Check language of the webpage content
Webpage content is in English?
No
Yes
Remove webpage
Keep webpage

Fig. 2 Steps of data cleaning

TABLE V
EXAMPLE OF A RECORD WITH A NULL URL

URL Category	Title	Domain	URL
Business Opportunity			

From the unique URL, which has 29,444 records, the inactive URL has 14,716 records, while the active URL has 14,728 records. It is identified that there is a reduction in URL records. The summary of the number of URLs after performing data cleaning is shown in Table VI.

TABLE VI
NUMBER OF URLS AFTER PERFORMING DATA CLEANING

Type of Records	Number of URLs
Records of raw URL	541,702
Records of null URL	2,719
Records without null URL	538,983
Records of duplicate URL	509,539
Records of unique URL	29,444
Records of inactive URL	14,716
Records of active URL	14,728

Most web pages are developed by using HTML tags. Hypertext Markup Language is referred to as HTML. It serves as the default markup language when constructing web pages. Besides the HTML tags, there are also XML tags. The most salient difference between HTML and XML tags is that the HTML tags are used for the presentation of data while the XML tags are used for the transfer of data [27]. Thus, this study only keeps the URL that contains HTML tags in the web page source code. The steps of data cleaning are described below.

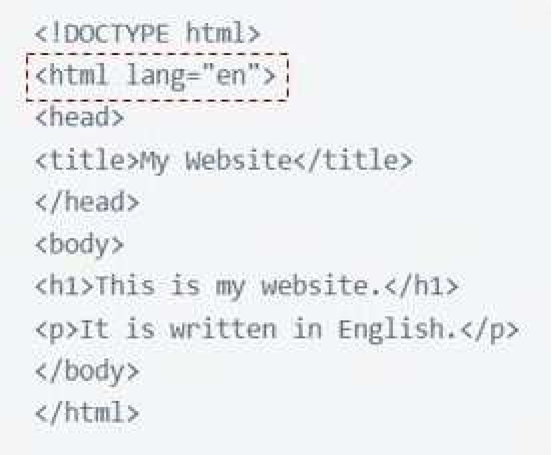
- Check each URL web page source code. Remove URLs not containing HTML tags in the web page source code, including XML and empty web pages.
- Download each URL's HTML source code web page and save it in a text file.
- Check the HTML source code of the web page to see whether it contains sentences related to the error page, such as ‘404 not found in the web page's title. If it has occurred, the URL is removed. This step is performed to ensure only active URL records are kept.

This study only collects the Games and Online Video Streaming category URLs. After data cleaning on the website browsing record, there are 216 Game web pages and 234 Online Video Streaming web pages. To expand the dataset for this research, the web pages on Games and Online Video have also been searched for us. Table VII shows the number of URLs collected, which are 640 Game web pages and 407 Online Video Streaming web pages.

TABLE VII
NUMBER OF DATASET URLS IN THE CATEGORY GAME AND ONLINE VIDEO STREAMING

Dataset URLs	Web Page Category		Total
	Game	Online video streaming	
From the website browsing record	216	234	450
Search by ourselves	424	173	597
Total	640	407	1,047

The next step is to include only web pages with English content by identifying the language of the text based on the HTML lang attribute in the web page’s HTML source code. The HTML lang attribute specifies the language used on the web page. For example, as shown in Fig. 4, the HTML lang attribute is set to “en” on the <html> element to inform browsers and search engines that the web page is in English [28]. If the HTML lang attribute is not defined in the web page’s HTML source code, the web page’s content is manually checked to ensure that only English content is included. It has been determined that there are 475 Game web pages and 277 Online Video Streaming web pages with English content.



```
<!DOCTYPE html>
<html lang="en">
<head>
<title>My Website</title>
</head>
<body>
<h1>This is my website.</h1>
<p>It is written in English.</p>
</body>
</html>
```

Fig. 4 Example of the use of the HTML lang attribute

B. *Web Content Pre-processing*

Now, the dataset contains the downloaded files of the HTML source code web page for each URL record. It could be sufficient to determine what a web page category is solely by studying the content on the web page [7]. As a result, the textual content is retrieved from the web page's HTML source code because it is a crucial component used to classify web pages and has the biggest influence on the type of web page [29], [30].

Cleaning up noisy data from the HTML source code of a web page is known as web content pre-processing. Below is a description of the procedures involved in web content pre-processing [21].

- Strip off HTML tags.
- Replace the number and symbol with space.
- Tokenization.
- Lemmatization involves reducing words to their dictionary forms by removing word affixes.
- Erase punctuation.
- Remove custom stop words, and technical terms used on the web page, such as ‘com,’ ‘www,’ ‘HTTP,’ ‘form,’ and others.
- Remove words with two or fewer characters and words with 15 or more characters.

- Remove infrequent words that do not appear more than two times.

After completing web content pre-processing, irrelevant information is removed, including HTML tags, JavaScript, CSS code, and technical terms commonly used in websites. As noted in [13] and [31], this practice is essential because such elements do not significantly contribute to the analysis. In the end, only tokenized words from the web page's HTML source code within the title and body content remain, as they hold greater value and meaning for representing the web page.

Using the bag-of-words approach, tokenized words are analyzed. This method, often called a term-frequency counter, counts the frequency of each word without considering its order in the document. It's a widely used technique for document representation [13]. Typically, words from web pages are converted into a matrix representation to be fed into a web page classifier. However, a common issue arises: the matrix representation can be excessively sparse, resulting in many zeros. To address this, word embedding creates dense text features for the matrix representation. Nevertheless, this approach has a constraint: each web page's text content must be transformed into an equal-length vector representation. If the web page's text content is too lengthy, it will be truncated to ensure equal vector lengths after the transformation [17].

Rather than converting the words on the web page into a matrix representation, this study visualizes the bag-of-words model using a word cloud image. A word cloud image (a text cloud or a tag cloud) is a visual representation of text data. In the word cloud image, the most frequent or widespread words stand out because they appear in the center, are represented using different colors, and are larger than the other words.

This study illustrates the differences between word cloud images used to represent web pages before and after web content pre-processing, as shown in Fig. 5 and Fig. 6. It is evident that the word cloud images before web content pre-processing contain a significant amount of noisy data, displaying irrelevant information such as punctuation as the most frequent token in the web pages. In contrast, the word cloud images after web content pre-processing highlight the most popular words related to the category of the web pages. For example, in Fig. 5, the words 'game' and 'play' emerge as the most popular words used on Game-related web pages, while in Fig. 6, the words 'movie' and 'watch' dominate the word cloud, indicating their popularity on Online Video Streaming web pages. These popular words are represented in red and appear in a larger font size.

Nowadays, the Convolutional Neural Network (CNN) is among the dominant approaches in the deep learning research community [32]. Typically, CNNs use images as input datasets to train models. In this study, the representation of web pages using word cloud images will be used as input for CNN to develop a model for classifying web pages.

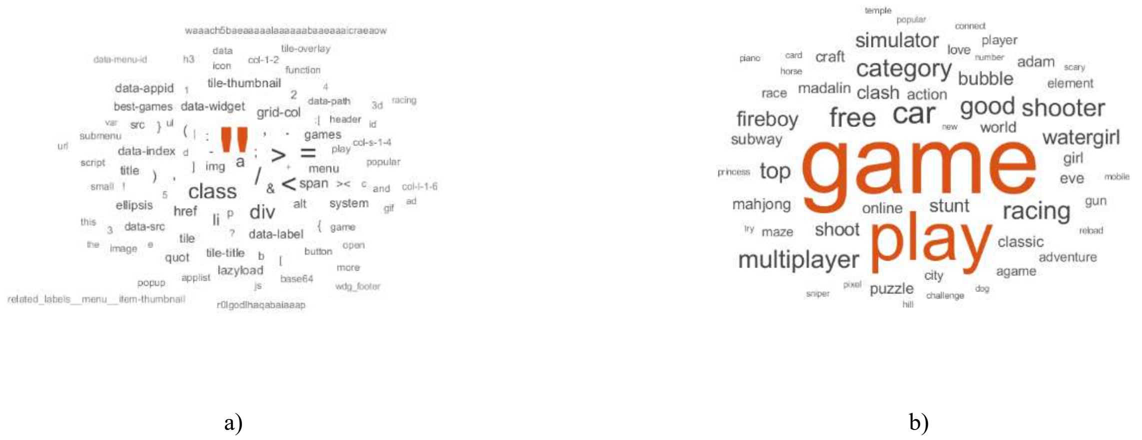


Fig. 5 Examples of word cloud images for Game web pages a) before web content pre-processing, and b) after web content pre-processing

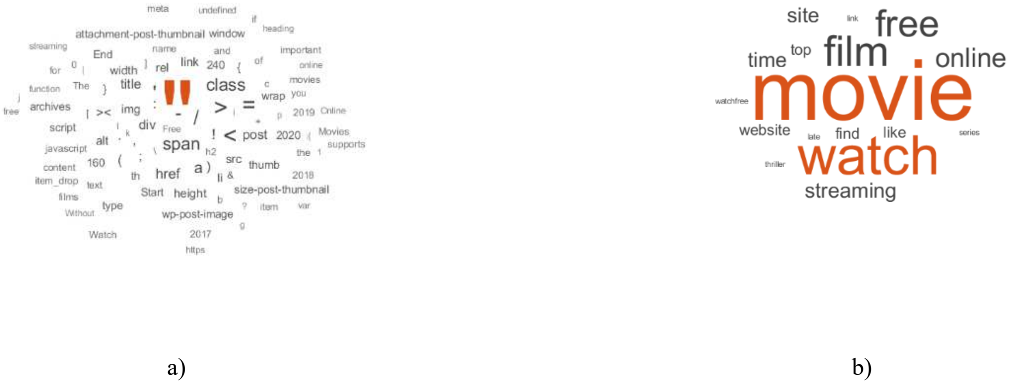


Fig. 6 Examples of word cloud images for Online Video Streaming web pages a) before web content pre-processing, and b) after web content pre-processing

IV. Conclusion

The raw dataset of website browsing records needs to undergo data pre-processing before it can be used for data analysis. Data pre-processing comprises two key activities: data cleaning and web content pre-processing. Data cleaning, the initial step in this process, involves retrieving the active web page and downloading its HTML source code in English content for web pages categorized under Game and Online Video Streaming. The subsequent activity in data pre-processing and web content pre-processing is aimed at removing noisy data from HTML documents, leaving only meaningful words that can represent the web page. These words are then presented as word cloud images, showcasing the most popular words at the center of the image. In our forthcoming work, the CNN-based web page classifier will provide this word cloud images to determine whether a given web page belongs to the Game or Online Video Streaming category.

Acknowledgment

The authors thank Dr. Azlee Zabidi, Senior Lecturer in the Faculty of Computing at Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), for his invaluable guidance on utilizing MATLAB for the data pre-processing of website browsing records. The research discussed in this article was made possible through funding from UMPSA, specifically Grant PGRS2003104.

References

[1] J. M. G. Costa, "Web page classification using text and visual features," M.S. thesis, Coimbra Univ., Coimbra, 2014.

[2] Faizan I Khandwani and Ashok P Kankale, "Preprocessing Techniques for Web Usage Mining," *International Journal of Scientific Development and Research (IJSDR)* , vol. 1, no. 4, pp. 330–334, 2016.

[3] S. Sharma and A. Bhagat, "Data preprocessing algorithm for web structure mining," in *2016 Fifth International Conference on Eco-friendly Computing and Communication Systems (ICECCS)*, IEEE, 2016, pp. 94–98.

[4] S. Vijayarani and K. Geethanjali, "Web Page Noise Removal-A Survey," *Int J Sci Res Sci Technol*, vol. 3, no. 7, pp. 172–181, 2017.

[5] S. S. Kumar and M. K. Singh, "Web Pattern Analysis Using Web Structure Mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 7, no. 5, 2017.

[6] P. V. Nainwani and P. Prajapati, "Comparative study of web page classification approaches," *Int J Comput Appl*, vol. 179, pp. 6–9, 2018.

[7] E. Buber and B. Diri, "Web page classification using RNN," *Procedia Comput Sci*, vol. 154, pp. 62–72, 2019.

[8] *Internet Users Survey 2020*. Malaysian Communications and Multimedia Commission, 2020. Accessed: Apr. 01, 2021. [Online]. Available: https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/IUS-2020-Report.pdf

[9] R. A. Davis, "A cognitive-behavioral model of pathological Internet use," *Comput Human Behav*, vol. 17, no. 2, pp. 187–195, 2001.

[10] K. S. Young and R. C. Rogers, "The relationship between depression and Internet addiction," *Cyberpsychology & behavior*, vol. 1, no. 1, pp. 25–28, 1998.

[11] F. Cao and L. Su, "Internet addiction among Chinese adolescents: prevalence and psychological features," *Child Care Health Dev*, vol. 33, no. 3, pp. 275–281, 2007.

[12] G. M. University, "Internet Addiction." Accessed: Apr. 01, 2021. [Online]. Available: https://shs.gmu.edu/healthed/internet-addiction/

[13] A. Osanyin, O. Oladipupo, and I. Afolabi, "A review on web page classification," *Covenant Journal of Informatics and Communication Technology*, vol. 6, no. 2, pp. 11–28, 2018.

[14] E. Suganya and D. S. Vijayarani, "Web page classification in web mining research-A survey," *Int J Innov Res Sci Eng Technol*, vol. 6, pp. 17472–17479, 2017.

[15] L. Safae, B. El Habib, and T. Abderrahim, "A review of machine learning algorithms for web page classification," in *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, IEEE, 2018, pp. 220–226.

[16] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, "Systematization of knowledge (sok): A systematic review of software-based web phishing detection," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2797–2819, 2017.

[17] Q. Zhao, W. Yang, and R. Hua, "Design and research of composite web page classification network based on deep learning," in *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, IEEE, 2019, pp. 1531–1535.

[18] A. Chechulin and I. Kotenko, "Application of image classification methods for protection against inappropriate information in the internet," in *2018 IEEE International Conference on Internet of Things and Intelligence System (IOTAIS)*, IEEE, 2018, pp. 167–173.

[19] C. Patel and H. Diwanji, "A Survey on Web Content Extraction and Noise Reduction from Webpage," *Int J Sci Res Sci Eng Technol*, vol. 1, no. 6, pp. 127–130, 2015.

[20] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings* , vol. 3, no. 1, pp. 91–99, 2022.

[21] H. Jamshed, S. A. Khan, M. Khurram, S. Inayatullah, and S. Athar, "Data Preprocessing: A preliminary step for web data mining," *3c Tecnología: glosas de innovación aplicadas a la pyme*, vol. 8, no. 1, pp. 206–221, 2019.

[22] H. Li, Z. Zhang, and Y. Xu, "Web page classification method based on semantics and structure," in *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, 2019, pp. 238–243.

[23] A. K. Nandanwar and J. Choudhary, "Web page categorization based on images as multimedia visual feature using Deep Convolution Neural Network," *International Journal on Emerging Technologies*, vol. 11, no. 3, pp. 619–625, 2020.

[24] M. Du, Y. Han, and L. Zhao, "A heuristic approach for website classification with mixed feature extractors," in *2018 IEEE 24th International Conference on Parallel and Distributed Systems (ICPADS)*, IEEE, 2018, pp. 134–141.

[25] M. J. H. Mughal, "Data mining: Web data mining techniques, tools and algorithms: An overview," *International Journal of Advanced Computer Science and Applications*, vol. 9, no. 6, 2018.

[26] N. Sharma, R. Agarwal, and N. Kohli, "Review of features and machine learning techniques for web searching," in *2016 11th International Conference on Industrial and Information Systems (ICIIS)*, IEEE, 2016, pp. 312–317.

[27] L. Yi, B. Liu, and X. Li, "Eliminating noisy information in web pages for data mining," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 296–305.

[28] I. Palii, "The Comprehensive Guide to the Lang HTML Attribute." Accessed: Oct. 05, 2023. [Online]. Available: https://sitechecker.pro/what-is-html-lang-attribute/

[29] S. M. Babapour and M. Roostaee, "Web pages classification: An effective approach based on text mining techniques," in *2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, IEEE, 2017, pp. 320–323.

[30] M. Hashemi, "Web page classification: A survey of perspectives, gaps, and future directions," *Multimed Tools Appl*, vol. 79, no. 17–18, pp. 11921–11945, 2020.

[31] B. A. Alahmadi, P. A. Legg, and J. R. Nurse, "Using internet activity profiling for insider-threat detection," *Special Session on Security in Information Systems*, vol. 2, pp. 709–720, 2015.

[32] F. De Fausti, F. Pugliese, and D. Zardetto, "Towards automated website classification by deep learning," *Rivista di Statistica Ufficiale*, pp. 9–50, 2019.