International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Jan-24)

http://dx.doi.org/10.12785/ijcds/150127

Use Word Cloud Image Of Web Page Text Content On Convolutional Neural Network (CNN) For Classification Of Web Pages

Siti Hawa Apandi[1], **Jamaludin Sallim**[1] **and Rozlina Mohamed**[1]

[1]*Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, Pekan, Pahang, Malaysia*

Received 3 May. 2023, Revised 2 Jan. 2024, Accepted 6 Jan. 2024, Published 15 Jan. 2024

Abstract: In the modern digital era, people easily access the internet to find information through website visits. Many individuals are attracted to online pages featuring games and video content. Prolonged exposure to such web pages can result in internet addiction, leading to negative consequences. To address this issue, it is crucial to impose restrictions on websites offering gaming and streaming content. To accomplish this, an essential tool is needed to classify web pages based on their content. In the categorization process, the text content of the web page is initially extracted. Since conventional matrix representations are not suitable for processing lengthy web page text, this study employs an innovative method involving the use of word cloud images to visually represent words extracted from the web page text after data pre-processing. Words that appear most frequently in the web page text are displayed in larger fonts and centered in the word cloud image, reflecting the subject matter of the web page. A Convolutional Neural Network (CNN) is then utilized to identify patterns in the central part of the word cloud image, facilitating the categorization of web pages based on their content. The proposed model for classifying web pages achieves an accuracy rate of 0.86, significantly improving the accuracy of web page categorization. By leveraging insights gained from web page classification, authorities can proactively monitor online user behavior, identifying individuals struggling with internet addiction and offering help if needed.

Keywords: Web page classification, document representation, word cloud image, deep learning, Convolutional Neural Network

1. INTRODUCTION

In the contemporary era, with the widespread use of mobile devices and computers, nearly everyone has internet access for purposes such as connecting with family and friends, seeking information, and entertainment, among others. The internet encompasses a platform for sharing information known as the World Wide Web or Web, consisting of various web pages. This online space allows people to efficiently retrieve information by navigating through web pages. The number of web pages continues to rise rapidly, with new additions every day. According to the WWW Size Project [1], Google has recorded at least 4.45 billion websites, highlighting the expansive growth of the internet.

The abundance of information available on the internet has raised concerns about information organization and management [2]. Information retrieval has become increasingly challenging [3]. For instance, users often face uncertainty about whether they have landed on the correct web page containing the desired data. This issue is addressed through the introduction of web page classification, a process that sorts and categorizes web pages based on their content [2]. This classification system simplifies the process for users to access accurate and timely information.

The Malaysian Communications and Multimedia Commission (MCMC) conducted a survey to gain insights into Malaysian internet users. The survey aims to monitor online behavior and identify user trends and tendencies in Malaysia. According to the survey's findings, internet users engage in various online activities. The majority of users access the web for leisure, participating in activities such as watching or downloading videos online and playing games [4]. The following sections will elaborate on these two types of websites.

An Online Video Streaming platform is a website that streams videos, TV shows, and films over the internet, offering users a convenient way to access content without the need to download and store large video files on their de-

E-mail address: sitihawa.apandi@gmail.com, jamal@umpsa.edu.my, rozlina@umpsa.edu.my

vices. This eliminates the burden of significant storage space consumption. According to Nielsen data, approximately 14 million users, including one million new streaming customers, make up about 78% of Malaysia's viewers aged 15 and older [5]. Some of the most well-known Online Video Streaming websites include YouTube, Viu, iFlix, and many others.

A game website, commonly known as a web browser game, is a type of computer game that users can play online through a web browser. Typically, these games are free to play. One of the key advantages for users is that web browser games don't require a high-end gaming PC for installation, as the necessary content is automatically downloaded from the game's website. These games come in various genres and can be enjoyed solo or with others [6]. Numerous game websites are accessible, offering both brand-new games and retro games that have been updated for web browsers. One well-known game currently accessible via a web browser is Pac-Man. Massively Multiplayer Online Role-Playing Games (MMORPGs) are a popular category of web browser games, with Runescape being a prominent example. These games often host thousands of players who can participate simultaneously in this online video game [7].

While some internet users responsibly harness the power of the internet to address study, work, and daily needs due to their self-discipline, others fall into negative internet habits. These individuals become entrenched in visiting unrelated websites for activities such as gaming, online video consumption, social networking, and blogging. Such behavior can have an adverse impact on their overall performance, contributing to problems like sleep deprivation, loss of appetite, physical inactivity, and difficulties in concentration [8].

Internet addiction can become a serious issue due to increased access to online activities, particularly the reckless browsing of gaming and online video websites. The term "internet addiction" refers to "individuals' inability to control their internet usage, leading to various problems in their lives, including psychological, social, school, and work difficulties" [9], [10]. According to an MCMC online survey, the majority of internet users are in their 20s, making college and university students more susceptible to internet addiction [4].

Researchers investigating internet addiction among college students found that a significant portion of these individuals faced challenges in completing their academic assignments, preparing for examinations, and staying attentive in class due to excessive time spent online, particularly on unproductive websites and gaming. Consequently, these students grapple with substantial academic difficulties, disrupting their daily routines [8]. The study also revealed that 50% of the students who faced expulsion from school due to poor academic performance attributed their problems to excessive internet use [11].

Preventing internet addiction can be achieved by restricting access to predetermined categories of websites. Authorities, such as administrators from colleges and universities, can play a pivotal role in implementing this measure. A common approach taken by these administrators is to establish guidelines and restrict internet usage to specific types of websites. Before implementing restrictions on internet access, it is essential to categorize the web page into its appropriate category using a web page classification model.

Deep learning has garnered substantial recent attention due to its ability to address challenges that conventional Artificial Intelligence methods might find daunting. It is a branch of machine learning centered on deep neural networks. Deep learning has seen increased exploration for its potential in handling classification tasks, particularly in the domain of web page classification, where it operates as a web page classifier [12], [13].

Text is a fundamental feature in web page classification [12], [14], but it cannot be directly input into a neural network. To bridge this gap, a technique known as matrix representation is employed to convert text into numerical data [15]. Neural networks require fixed-length matrices as input [16]. However, the feature vector's length can significantly increase when working with a large number of words in the text, requiring substantial processing power and potentially affecting performance [17], [18]. In such cases, text is often truncated to ensure it can be converted into a consistent-length vector representation [18].

Given these challenges, this study aims to investigate document representation techniques that can effectively capture the text data of a web page without necessitating text truncation, thus retaining potentially valuable information. Additionally, the study seeks to explore methods for transforming text data into image format. It is recognized that image data, with its pixel values, can be readily utilized as input for training deep learning models. Specifically, the Convolutional Neural Network (CNN), a potent deep learning algorithm extensively utilized in tasks involving image and text classification, has exhibited encouraging outcomes [19], [20].

The primary objective of this study is to employ Convolutional Neural Networks (CNN) to construct a model for web page classification. This model will be designed to determine whether a given web page pertains to the categories of Game or Online Video Streaming. By combining techniques to effectively represent text data and leveraging the capabilities of CNN, the study aims to enhance web page classification accuracy and efficiency in categorizing web content.

The data derived from web page classification can provide authorities with the opportunity to actively monitor the online browsing behavior of internet users. This monitoring can play a crucial role in identifying individuals who may be grappling with internet addiction. By recognizing such individuals, authorities can offer them the necessary support and assistance to address their addiction and promote healthy online habits.

This is a brief summary of the study's main contribution. In the realm of document representation techniques, this research introduces an innovative approach by utilizing word cloud images to visually represent words extracted from web pages after data pre-processing. Within these word cloud images, frequently occurring words are displayed in larger fonts and positioned at the center. These prominent terms signify words that are highly prevalent in the web page's text content, offering relevance to the content's subject matter. For web page categorization, the CNN-based model relies on patterns within the central section of the word cloud image, enabling precise judgments. This unique method contributes to advancing document representation techniques in the study's context.

The remainder of the paper follows this structure: Section 2 provides an overview of the literature review. Section 3 outlines the methodology. The findings concerning the performance of the web page classification model are deliberated in Section 4. Section 5 serves as the conclusion for this study.

2. LITERATURE REVIEW

A. *Web Page Classification*

As defined in [21], [22], [23], [24], web page classification involves the task of attributing one or more category labels to a web page. There are two primary methods for categorizing web pages: manual classification, where humans make the assignments, and automatic classification, which is performed by computer algorithms. The traditional approach to organize web page relied on human intervention to determine the appropriate categories for each page, but this method is time-consuming and impractical due to the sheer volume of web pages in existence today [15], [24].

In response to the challenges posed by the overwhelming number of web pages, automatic classification emerges as the most efficient solution. Automatic classification relies on a web page classifier to sort web pages into different categories. Machine learning and deep learning algorithms have been employed as web page classifiers in this context, and their performance is briefly summarized below.

A variety of machine learning algorithms, including k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes, Decision Trees, and Artificial Neural Networks (ANN), have been utilized as web page classifiers, often referred to as traditional machine learning methods. Safae, et al. [25] conducted a study to explore the use of these machine learning algorithms in web page classification. The algorithms they investigated encompassed KNN, SVM, Naïve Bayes, and ANN. Upon comparing the performance of these algorithms in web page classification, it became evident that the Artificial Neural Network (ANN), also known as the Neural Network (NN), consistently outperformed other machine learning methods, achieving the highest accuracy.

Furthermore, it was noted that these algorithms worked effectively with a limited number of web pages. However, their performance deteriorated significantly when confronted with a large volume of web pages to process [25].

In addition to conventional machine learning techniques, there is a growing adoption of deep learning methods for the classification of web pages. It demonstrates that deep learning is especially advantageous when handling large and complex datasets, in comparison to traditional machine learning algorithms [26]. A noteworthy benefit of deep learning lies in its capacity to automatically identify the pertinent features for classification, in contrast to traditional machine learning where these features must be manually defined [18]. Hence, this study will focus on the utilization of deep learning for web page classification.

Several researchers have proposed a multitude of systems for web page classification, employing various techniques to enhance classifier accuracy. The efficient performance of the majority of classification algorithms relies on the presence of high-quality training data. The representation of the document is closely linked to both the quality and quantity of training data [15].

Web pages encompass a range of data types, such as text, images, audio files, and videos [27], [28], [29]. Selecting the most suitable web page features that accurately reflect the content becomes challenging due to the increasing presence of extraneous information, such as advertisements and links. These extraneous elements, often referred to as noisy data, can significantly disrupt feature extraction and diminish classification accuracy [18]. To mitigate this issue, data pre-processing plays a pivotal role in web page classification. This stage involves refining document representation by eliminating irrelevant and noisy features, which has a positive impact on the accuracy and speed of the web page classifier while also reducing overfitting concerns [15].

Web page classification commonly employs two types of features: text and images [29], [30]. Document representation is necessary for these features since deep learning processes numerical data exclusively. In the case of images, deep learning takes input from the pixels, which are numerical values ranging from 0 (representing black) to 255 (representing white). Deep learning interprets patterns within these image pixels and categorizes the images accordingly.

Several studies have looked into the classification of web pages using features from images.

- López-Sánchez et al. [31] propose the use of a Deep Convolutional Neural Network (DCNN) and transfer learning to classify web pages based on existing multimedia content, achieving an accuracy of 90%.

- In a similar vein, López-Sánchez et al. [32] advocate the employment of a Deep Convolutional Neural Network (DCNN) and Case-Based Reasoning (CBR) for image-based web page classification, achieving an accuracy of 97.61%.
- Chechulin et al. [29] introduce image classification as a means to restrict internet users' access to inappropriate content, reporting an accuracy of 89%.
- López-Sánchez et al. [33] propose image-based web page classification by combining transfer learning methods with metric learning using a DCNN classifier, resulting in an accuracy of 98.97%.
- Hashemi et al. [34] suggest using a Convolutional Neural Network (CNN) to recognize images on web pages and categorize them as hard, soft, or symbolic propaganda for automated detection of dark web pages, achieving an accuracy of 86.08%.
- Nandanwar et al. [30] put forward a web page classification approach based on web page images, employing a Deep Convolutional Neural Network (DCNN) and the transfer learning method, with an accuracy of 86%.

While using images as input for deep learning is straightforward, the same cannot be said for text data. Textual data can be sourced from various elements, including the textual content, HTML tags, and the URL of a web page [29], [30]. Web page classification is not the same as text classification because web pages contain semi-structured data, utilizing tags to structure and organize the information displayed in web browsers [15], [30], [28], [35]. Pre-processing steps involved in text classification, such as data cleaning, word segmentation, and the removal of stop words, can be applied to web page text data [16], [36]. However, these pre-processed words can't be directly fed into a neural network. Each word must be converted into a word vector, which can be done in two ways: one-hot encoding and word embedding [16].

One-hot encoding replaces each word with a vector filled with zeros, except for the position corresponding to the word's index, which is assigned a value of 1. However, this method has a drawback in that it generates sparse vectors with numerous dimensions, leading to the curse of dimensionality and computational inefficiency [16]. An alternative approach is word embedding, where each word is represented by a real-valued vector in a specified vector space, and the vector values for each word are learned during neural network training. This allows words with similar meanings to have comparable representations. Two common methods for learning word embeddings are Word2Vec and GloVe. Nevertheless, these techniques struggle with out-of-vocabulary (OOV) words, which are words not present in the dictionary. To address this, a special token (Unknown - UNK) can be designated to represent OOV words [28].

Many approaches for classifying web pages using textual features have been explored in the literature.

- Du et al. [37] introduced a website classification approach using a combination of Convolutional Neural Network (CNN) and Gated Recurrent Unit (GRU). They extract information from website titles, keywords, and descriptions and represent it using a word embedding matrix, achieving an accuracy of 81.1%.
- Another method for web page classification, based on Recurrent Neural Networks (RNN), leverages meta tag information like titles, descriptions, and keywords. This approach employs word embeddings generated through the Glove method and achieves an accuracy of 85% [28].
- Li et al. [35] proposed a web page classification model that combines Convolutional Neural Networks (CNN) with Word2vec and the Skip-Gram model. They extract information from title tags, keyword meta tags, and the textual content of web pages, achieving an F-measure of over 94%.
- Zhao et al. [18] presented a web page categorization model using CNN and RNN, with word embeddings applied to features derived from web page titles, descriptions, and text content. Their approach achieves an accuracy of 90%.
- Maladkar [20] developed a URL classification method based on CNN models that utilize Glove-derived word embeddings, resulting in an accuracy of 85%.
- Deng et al. [36] introduced a web page categorization method that incorporates heterogeneous features, such as vector concatenation, structural features, and textual features. They use a combination of classifiers, including Long Short-Term Memory (LSTM) and Support Vector Machines (SVM), achieving an accuracy of 94.2%.
- He et al. [38] introduced an approach for classifying web news that involves noise detection, BERT-based semantic similarity noise filtering, and a Convolutional Neural Network (NF-CNN), achieving an F-measure of 95.61%.
- Another method for URL-based web page classification, aimed at categorizing web pages as Kids-specific or not, combines Convolutional Neural Network (CNN) and Bidirectional Gated Recurrent Unit (BGRU) and achieves an accuracy of 82.04% [39].

After reviewing the prior research on web page classification, this study has made several noteworthy observations. Firstly, it's evident that the predominant feature used for web page classification is text, rather than images. This aligns with the assertion that text features have become a pivotal element in web page classification, exerting the most significant influence on determining the type of a web page [12], [14]. Conversely, there is relatively limited research focused on web page classification based on URL features. This can be attributed to the fact that URLs offer incomplete representations of web page content [37].

Another key observation is that the majority of these research works make use of Convolutional Neural Networks (CNN). CNNs have demonstrated their adaptability for both image and text-based web page classification. However, when it comes to performance, it's worth noting that image-based web page classification exhibits higher performance levels in comparison to text-based web page classification.

This study summarizes the strengths and weaknesses of using text and image data in web page classification, as observed in the Table I.

TABLE I. Strengths and weaknesses of using text and image data

Data types	Strengths	Weaknesses
Text data	Key element that is utilised to categorise web pages.	Each word must be transformed into a word vector before inputting it into the classification algorithm. Text is frequently shortened to guarantee it can be transformed into a uniform-length vector representation.
Image data	Using images as input for a classification algorithm is simple and involves examining pixel patterns in the images. Higher performance levels are demonstrated via image-based web page classification.	Image features are not widely employed in web page classification, mainly due to the fact that not all web pages include images.

B. *Word Cloud Images*

The widely employed deep learning technique, Convolutional Neural Network (CNN), is renowned for its effectiveness in image analysis and has demonstrated promising results [20]. One illustrative example of CNN's application involves its use in classifying handwritten digits. In this case, a labeled dataset containing digits from one to ten was used to train the CNN. Through its layers and operations, the CNN automatically extracts valuable features from the input data. Essentially, the CNN learns the underlying patterns of these digits during the training process, enabling it to classify new input data with precision [40].

This study aims to apply this concept to classify web pages. Another technique for representing text in an image format is also explored.

Word cloud images, also referred to as text clouds or tag clouds, are visual representations that display a collection of words extracted from a document. These words are presented in a graphical format. In a word cloud image, the size and boldness of each word are indicative of its frequency of occurrence in the document. Essentially, the more often a word appears in the document, the larger and more prominently it is displayed in the word cloud image. This visual emphasis on frequently occurring words conveys their significance in the document, making it easier for viewers to identify the primary topics and themes that pertain to the content [41], [42].

3. METHODOLOGY

Figure 1 illustrates the procedural steps involved in developing the proposed web page classification model, encompassing five key stages: data collection, data pre-processing, obtaining the required features, building the web page classification model and evaluating the web page classification model. The description for each of these stages is detailed below.

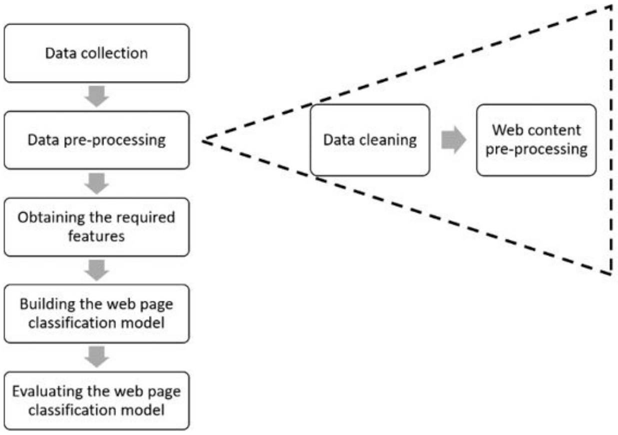

Data collection
Data pre-processing
Obtaining the required features
Building the web page classification model
Evaluating the web page classification model
Data cleaning
Web content pre-processing

Figure 1. Stages to develop the proposed web page classification model

Table II provides the details of the experimental environment employed in this study for conducting the experiment to develop the proposed web page classification model.

TABLE II. Details of the 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
Tools for deep learning	Deep Learning Toolbox

A. Data Collection

This study employs a self-collected dataset, utilizing real-world cases, specifically the website browsing records of Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) students. These records were made available by the Centre of Information Technology and Communication (PTMK), an entity within UMPSA responsible for tracking internet usage by maintaining records of accessed URL web pages. This dataset accurately reflects authentic user behaviors during internet browsing [37].

The website browsing records encompass data on URL web pages accessed by students during the week from March 17, 2019 (Sunday) 00:00:00 to March 23, 2019 (Saturday) 23:59:59. The dataset consists of 40 Microsoft Excel files, each dedicated to the web browsing history of an individual UMPSA student. In other words, there are records of website browsing for a total of 40 students. Furthermore, each Microsoft Excel file documenting web browsing activity contains over a thousand URL web pages, indicating that each student accessed thousands of web pages over the course of a week.

The URL web pages within the Excel file of website browsing records can be considered supervised since they have already been classified into specific URL categories. FortiGuard, a tool utilized by PTMK, has assigned a total of 53 URL categories to the web pages in the Excel file of the website browsing records.

B. Data Pre-processing

The raw dataset of URLs, obtained from website browsing records, often lacks the necessary structure for analysis, as it might lack completeness, consistency, or clarity. One of the most demanding tasks involves locating and extracting pertinent information from the dataset [43]. Consequently, data pre-processing plays a critical role in refining, rectifying, and priming the input data for mining purposes [44]. In this study, data pre-processing encompasses two main activities: data cleaning and the web content pre-processing. The first step in processing the raw dataset of URL web pages involves data cleaning, which encompasses several specific tasks. These tasks include eliminating null URLs, removing duplicate URLs to maintain distinct ones, and retrieving active URLs while discarding inactive ones. An inactive URL indicates that the hyperlink is no longer functional, rendering the web page inaccessible.

Table III displays an overview of the number of URLs after data cleaning. Observations have revealed a decrease in the quantity of URL records, aiming to obtain a set of appropriate URLs for analysis.

TABLE III. Number of URLs after 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 distinct URL	29,444
Records of inactive URL	14,716
Records of active URL	14,728

Typically, web pages are built using the default markup language, HTML (Hyper Text Markup Language) tags. In addition to HTML tags, there are XML tags. The primary distinction between HTML and XML tags is that HTML tags serve the purpose of data presentation, while XML tags are designed for data transfer. The majority of web pages on the internet are created using HTML tags rather than XML tags [45]. Therefore, this study selectively retains URLs that contain HTML tags within the web page source code by retrieving the HTML source code of each URL web page and saving it in a text file.

This study exclusively gathers URL datasets from the Game and Online Video Streaming categories, as the study's objective is to develop a model for categorizing web pages within these specific domains. Following the completion of data cleaning on the website browsing records, there are 216 web pages dedicated to Games and 234 web pages focused on Online Video Streaming. In an effort to expand the dataset size, additional web pages related to Games and Online Video content were also independently sourced. This resulted in a total of 640 Game web pages and 407 Online Video Streaming web pages being collected.

In this study, only web pages containing English content were gathered. This was accomplished by identifying the language of the text through the HTML lang attribute in the source code of the web page. The HTML lang attribute specifies the language utilized on the web page [46]. If the HTML lang attribute is absent in the web page's HTML source code, a manual examination of the web page's content is conducted to ensure that only English content is incorporated. The investigation found that there are a total of 475 web pages related to Game and 277 web pages dedicated to Online Video Streaming, all featuring English content.

Presently, the dataset comprises downloaded files of HTML source code from web pages, which are valuable for extracting web page content. It might be adequate to exclusively examine the web page information in order to determine its category [28]. Consequently, the textual content is extracted from the HTML source code of the web page, as this textual content serves as a crucial element in web page classification [12], [14].

Following that, the next step is to continue with web content pre-processing in order to eliminate noisy data from the HTML source code of the web page. An outline of the steps involved in web content pre-processing can be found below [43].

- Remove the HTML tags.
- Substitute numbers and symbols with empty spaces.
- Tokenization involves breaking down the text into smaller units, which can be individual words or tokens [15].
- Lemmatization is the process of reducing words to their dictionary forms by removing word affixes.
- Remove punctuation.
- Eliminate specific stop words that refer to technical terms used on the web page, such as 'com', 'www', 'HTTP,' 'form,' and other similar terms.
- Filter out words that have two or fewer characters as well as words with 15 or more characters.
- Exclude words that occur less than three times, as they are considered infrequent.

Once the web content pre-processing is complete, irrelevant elements like HTML tags, JavaScript, CSS code, and technical website terminology are removed, as they do not substantially contribute to the analysis. This practice is widely recognized in web-related research [15], [47]. Ultimately, what remains are tokenized words extracted from the HTML source code of the web page, specifically from the title and body content. These words are more valuable and meaningful for representing the web page.

C. Obtaining The Required Features

The tokenized words obtained from the data pre-processing of the web page are subjected to analysis using a bag-of-words model, also referred to as a "term-frequency counter." This model disregards the word order in a document and focuses solely on counting how often each word is used. It stands as the most commonly employed technique for document representation. Typically, the words within the web page are transformed into a matrix representation, which is then provided as input to the web page classifier [15]. However, a challenge arises when this matrix representation becomes excessively sparse, resulting in numerous zeros. To address this issue, word embedding is utilized to create denser textual features for the matrix representation. Nevertheless, this approach comes with a constraint: each web page's text content must be converted into a vector representation of equal length. If a web page's text content is excessively long, it is truncated to ensure uniform vector length after the conversion [18].

Instead of transforming the words from the web page into a matrix representation, this study opts for the utilization of a word cloud image to visualize the bag-of-words model. A word cloud image serves as a graphical representation of text. Within the word cloud image, the most frequently occurring or popular words are prominently featured in the center, distinguished by varying colors and larger text size compared to less frequent words.

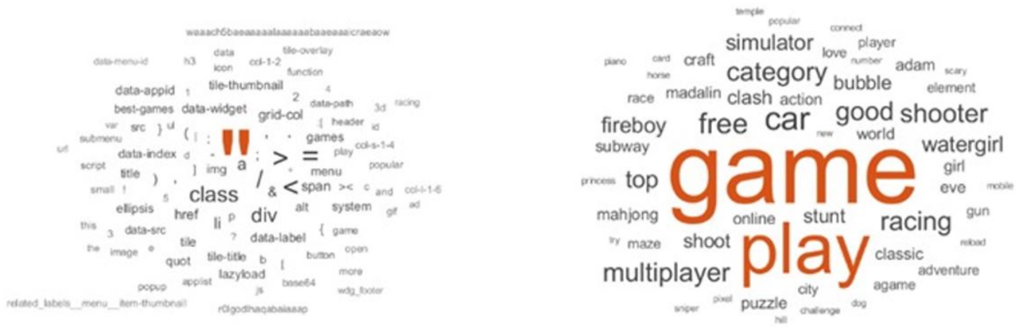
This study demonstrates the distinctions in word cloud images that are employed to represent web pages both before and after data pre-processing, as depicted in Figure 2 and Figure 3. It is clear that the word cloud images prior to data pre-processing contain a significant amount of noisy data, featuring unimportant information such as punctuation as the most frequently occurring element in the web pages. In contrast, the word cloud images following data pre-processing emphasize the most commonly used words associated with the category of the web pages. For instance, in Figure 2, the terms 'game' and 'play' emerge as the most prevalent words utilized on web pages related to Game, while in Figure 3, the terms 'movie' and 'watch' dominate the word cloud, indicating their popularity on web pages related to Online Video Streaming.

D. Building The Web Page Classification Model

This study employs a straightforward architecture for the Convolutional Neural Network (CNN), illustrated in Figure 4. The proposed model commences with the image input layer and subsequently includes three convolutional layers. Each of these convolutional layers is equipped with a batch normalization layer and a ReLU layer. The first and second convolutional layers are followed by a max-pooling layer with a pool size of 2. After the final convolutional layer, the model proceeds with a fully connected layer, a softmax layer, and a classification output layer.

It is worth noting that the convolutional layer, batch normalization layer, ReLU layer, and max-pooling layer are repetitively employed to enable the CNN to extract more comprehensive information regarding image features. The number of layers added is contingent on the complexity of the images under consideration, and thus, there are no strict rules dictating the exact number of layers to incorporate.

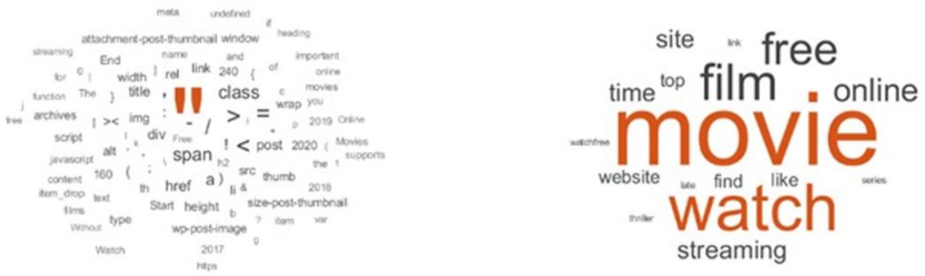
Table IV displays the number of word cloud images generated after conducting data pre-processing on the downloaded HTML source code web pages. Specifically, there



(a) Word cloud image for Game web page before data pre-processing

(b) Word cloud image for Game web page after data pre-processing

Figure 2. Examples of word cloud images for Game web pages before, and after data pre-processing



(a) Word cloud image for OVS web page before data pre-processing

(b) Word cloud image for OVS web page after data pre-processing

Figure 3. Examples of word cloud images for Online Video Streaming (OVS) web pages before, and after data pre-processing

are 308 word cloud images associated with Game web pages and 141 word cloud images associated with Online Video Streaming web pages. The dataset will be divided into two distinct subsets. The first subset is allocated for training and development purposes, while the second subset is reserved for validation and performance evaluation. The data is split in an 80% training and 20% validation ratio. Throughout the experiment, all datasets are randomized to ensure unbiased results.

TABLE IV. Number of dataset word cloud images

Dataset	Web Page Category		
	Game	Online video streaming	Total
Training (80%)	246	113	359
Validation (20%)	62	28	90
Total	308	141	449

Next, a range of training-related parameters for the Convolutional Neural Network (CNN) are specified. These parameters include the use of the Adam optimizer, 50 training epochs, and a batch size of 32. Once all the necessary elements, including the dataset, CNN architecture, and training parameters, are defined, the training process for the CNN is executed. This process culminates in the creation of a network, which represents the proposed model for classifying web pages.

Once the classification model is constructed, it becomes capable of automatically categorizing new web pages into the appropriate categories without manual intervention. In the subsequent step, this network will be employed in the validation process to assess and monitor its performance.

E. Evaluation Performance Of The Proposed Model For Classifying Web Page

The CNN-based model for classifying web pages is evaluated by examining its precision, recall, F-measure, and accuracy. These metrics are commonly used to assess the performance and effectiveness of classification models.

4. RESULT

This section will cover the findings pertaining to the web page classification model's performance. Figure 5 illustrates the confusion matrix outcomes. There are two classes: Game and OVS representing Online Video Streaming. Here

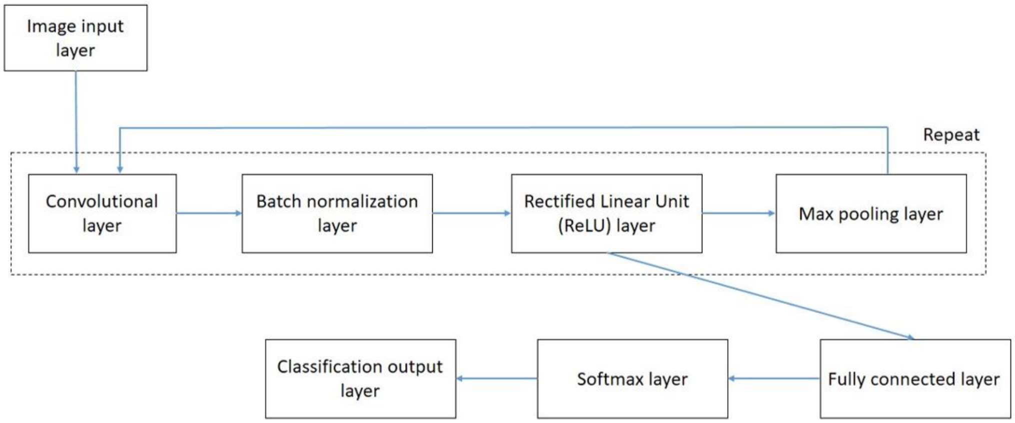


Figure 4. Structure layers of Convolutional Neural Network (CNN)

is the description of the results derived from the confusion matrix:

- A total of 56 datasets are accurately classified as Game, constituting 62.2% of the dataset.
- There are 21 datasets that have been correctly classified as OVS, representing 23.3% of the dataset.
- Out of the OVS datasets, 7 are mistakenly classified as Game, which accounts for 7.8% of the dataset.
- Among the Game datasets, 6 are inaccurately classified as OVS, making up 6.7% of the dataset.
- Among the 63 predictions made for Game, 88.9% are accurate, and 11.1% are incorrect.
- Of the 27 predictions made for OVS, 77.8% are accurate, while 22.2% are incorrect.
- Among the 62 Game datasets, 90.3% are correctly classified as Game, while 9.7% are erroneously predicted as OVS.
- Among the 28 OVS datasets, 75.0% are accurately classified as OVS, while 25.0% are classified as Game.
- In summary, the overall accuracy of the predictions is 85.6%, and 14.4% of the predictions are incorrect.

Based on the results obtained from the confusion matrix, various performance metrics can be calculated as follows:

1) Precision (P) as shown in (1).

$$P = \frac{TP}{TP + FP} = \frac{56}{56 + 7} = 0.889 \qquad (1)$$

2) Recall (R) as shown in (2).

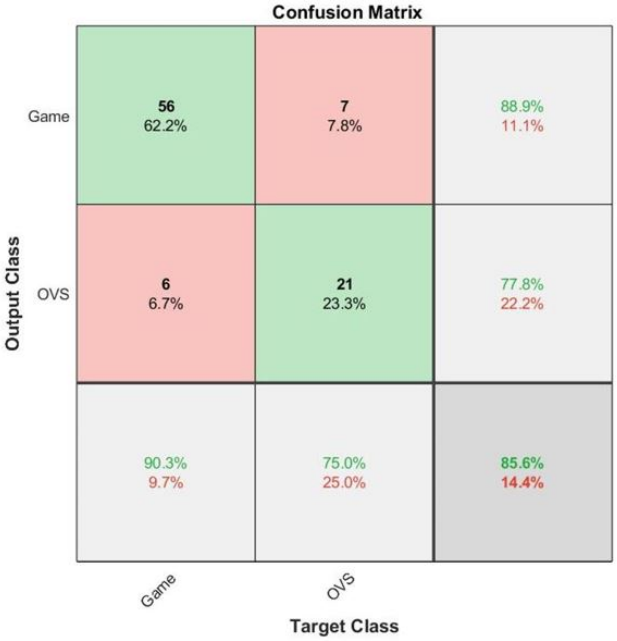
$$R = \frac{TP}{TP + FN} = \frac{56}{56 + 6} = 0.903 \qquad (2)$$

3) F-measure (F) as shown in (3).

$$F = \frac{2XPXR}{P + R} = \frac{2X0.889X0.903}{0.889 + 0.903} = 0.896 \qquad (3)$$

4) Accuracy (A) as shown in (4).

$$A = \frac{TP + TN}{TP + TN + FP + FN} = \frac{56 + 21}{56 + 21 + 7 + 6} = 0.856 \qquad (4)$$



	Game	OVS	
Game	56 62.2%	7 7.8%	88.9% 11.1%
OVS	6 6.7%	21 23.3%	77.8% 22.2%
	90.3% 9.7%	75.0% 25.0%	85.6% 14.4%

Figure 5. Outcomes of confusion matrix

5. CONCLUSION

This paper outlines the procedure for constructing a web page classification model based on Convolutional Neural Network (CNN). The model consists of five key stages: data collection, data pre-processing, obtaining the required features, building the web page classification model, and evaluation of the model's performance in classifying web pages. The implementation of this classification model facilitates the extraction of textual information from web pages, followed by data pre-processing and visualization of the most commonly used terms as a word cloud image. This visualization helps identify common words on the web page. To determine whether a web page falls within the Game or Online Video Streaming categories, the proposed web page classification model examines the word patterns in the word cloud image. The results of experiments demonstrate that the proposed web page classification model achieves an accuracy rate of 0.86. Such a model could be utilized by institutions, for instance, to establish regulations and restrict internet usage for users intending to visit Game and Online Video Streaming web pages, thereby serving as a preventative measure against internet addiction. Future endeavors in this field should focus on devising a methodology for selecting meaningful words from web page text content that more accurately reflect the page's purpose. This is essential because the most common words in a word cloud image may not always convey meaningful information and can potentially affect the performance of the web page classification model. Enhancements could also be made by incorporating a larger dataset of word cloud images related to Game and Online Video Streaming web pages to further develop the proposed CNN-based web page classification model.

Acknowledgment

The authors extend their gratitude for the invaluable guidance provided by Dr. Azlee Zabidi, Senior Lecturer at the Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA), who instructed them in the utilization of MATLAB for developing the proposed model. This model aims to categorize web pages based on patterns observed in word cloud images using Convolutional Neural Networks (CNN). The funding for this study was generously provided by the UMPSA Grant: PGRS2003104.

References

[1] N. Huss, "How many websites are there in the world? [2021]," https://siteefy.com/how-many-websites-are-there/, 2021, accessed: 2021-01-01.

[2] J. M. G. Costa, "Web page classification using text and visual features," Ph.D. dissertation, Coimbra University, Coimbra, Portugal, 2014.

[3] F. I. Khandwani and A. P. Kankale, "Preprocessing techniques for web usage mining," *International Journal of Scientific Development and Research (IJSDR)*, vol. 1, no. 4, pp. 330–334, 2016.

[4] *Internet Users Survey 2020*. Malaysian Communications and Multimedia Commission, 2020, https://www.mcmc.gov.my/skmmgovmy/media/General/pdf/IUS-2020-Report.pdf.

[5] R. Yunus, "Increasing streaming subscribers signals viewing preferences," https://themalaysianreserve.com/2020/08/03/increasing-streaming-subscribers-signals-viewing-preferences/, 2020, accessed: 2023-07-05.

[6] T. Barnett, "What is a browser-based game?" https://www.easytechjunkie.com/what-is-a-browser-based-game.htm, 2023, accessed: 2023-07-05.

[7] C. Hope, "Mmorpg," https://www.computerhope.com/jargon/m/mmorpg.htm, 2019, accessed: 2021-10-01.

[8] F. Cao and L. Su, "Internet addiction among chinese adolescents: prevalence and psychological features," *Child: care, health and development*, vol. 33, no. 3, pp. 275–281, 2007.

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

[10] R. A. Davis, "A cognitive-behavioral model of pathological internet use," *Computers in human behavior*, vol. 17, no. 2, pp. 187–195, 2001.

[11] G. M. University, "Internet addiction," https://shs.gmu.edu/healthed/internet-addiction/, accessed: 2021-04-01.

[12] M. Hashemi, "Web page classification: a survey of perspectives, gaps, and future directions," *Multimedia Tools and Applications*, vol. 79, no. 17-18, pp. 11 921–11 945, 2020.

[13] W. Luo, "Research and implementation of text topic classification based on text cnn," in *2022 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)*. IEEE, 2022, pp. 1152–1155.

[14] 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.

[15] 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.

[16] P. Song, C. Geng, and Z. Li, "Research on text classification based on convolutional neural network," in *2019 International conference on computer network, electronic and automation (ICCNEA)*. IEEE, 2019, pp. 229–232.

[17] A. R. Alharbi, S. D. Alharbi, A. Aljaedi, and O. Akanbi, "Neural networks based on latent dirichlet allocation for news web page classifications," in *2020 IEEE 2nd International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. IEEE, 2020, pp. 1–6.

[18] 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.

[19] F. De Fausti, F. Pugliese, and D. Zardetto, "Towards auto-

mated website classification by deep learning," *arXiv preprint arXiv:1910.09991*, pp. 9–50, 2019.

[20] K. Maladkar, "Content based hierarchical url classification with convolutional neural networks," in *2019 International Conference on Information Technology (ICIT)*. IEEE, 2019, pp. 263–266.

[21] J. Alamelu Mangai, V. Santhosh Kumar, and V. Sugumaran, "Recent research in web page classification–a review," *International Journal of Computer Engineering & Technology (IJCET)*, vol. 1, no. 1, pp. 112–122, 2010.

[22] P. V. Nainwani and P. Prajapati, "Comparative study of web page classification approaches," *International Journal of Computer Applications*, vol. 179, pp. 6–9, 2018.

[23] X. Qi, *Web page classification and hierarchy adaptation*. Lehigh University, 2012, http://wume.cse.lehigh.edu/pubs/qi-dissertation.pdf.

[24] E. Suganya and D. S. Vijayarani, "Web page classification in web mining research-a survey," *International Journal of Innovative Research in Science, Engineering and Technology*, vol. 6, pp. 17 472–17 479, 2017.

[25] 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.

[26] 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.

[27] S. Vijayarani and K. Geethanjali, "Web page noise removal-a survey," *International Journal of Scientific Research in Scienceand Technology*, vol. 3, no. 7, pp. 172–181, 2017.

[28] E. Buber and B. Diri, "Web page classification using rnn," *Procedia Computer Science*, vol. 154, pp. 62–72, 2019.

[29] 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.

[30] A. K. Nandanwar and J. Choudhary, "Web page categorization based on images as multimedia visual feature using deep convolution neural network," *Int. J. Emerg. Technol*, vol. 11, no. 3, pp. 619–625, 2020.

[31] D. López-Sánchez, A. G. Arrieta, and J. M. Corchado, "Deep neural networks and transfer learning applied to multimedia web mining," in *Distributed Computing and Artificial Intelligence, 14th International Conference*. Springer, 2018, pp. 124–131.

[32] D. López-Sánchez, J. M. Corchado, and A. G. Arrieta, "A cbr system for image-based webpage classification: case representation with convolutional neural networks," in *The Thirtieth International Flairs Conference*, 2017, pp. 483–488.

[33] D. López-Sánchez, A. G. Arrieta, and J. M. Corchado, "Visual content-based web page categorization with deep transfer learning and metric learning," *Neurocomputing*, vol. 338, pp. 418–431, 2019.

[34] M. Hashemi and M. Hall, "Detecting and classifying online dark visual propaganda," *Image and Vision Computing*, vol. 89, pp. 95–105, 2019.

[35] 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.

[36] L. Deng, X. Du, and J.-z. Shen, "Web page classification based on heterogeneous features and a combination of multiple classifiers," *Frontiers of Information Technology & Electronic Engineering*, vol. 21, no. 7, pp. 995–1004, 2020.

[37] 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.

[38] C. He, Y. Hu, A. Zhou, Z. Tan, C. Zhang, and B. Ge, "A web news classification method: fusion noise filtering and convolutional neural network," in *2020 2nd Symposium on Signal Processing Systems*, 2020, pp. 80–85.

[39] R. Rajalakshmi, H. Tiwari, J. Patel, A. Kumar, and R. Karthik, "Design of kids-specific url classifier using recurrent convolutional neural network," *Procedia Computer Science*, vol. 167, pp. 2124–2131, 2020.

[40] S. Rajwal, "Classification of handwritten digits using cnn," https://www.analyticsvidhya.com/blog/2021/07/classification-of-handwritten-digits-using-cnn/, 2021, accessed: 2023-07-05.

[41] V. Ibrahim, J. A. Bakar, N. H. Harun, and A. F. Abdulateef, "A word cloud model based on hate speech in an online social media environment," *Baghdad Science Journal*, vol. 18, no. 2, pp. 937–946, 2021.

[42] N. VM and D. A. Kumar R, "Implementation on text classification using bag of words model," in *Proceedings of the second international conference on emerging trends in science & technologies for engineering systems (ICETSE-2019)*, 2019, pp. 241–248.

[43] 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.

[44] 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.

[45] 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.

[46] I. Palii, "The comprehensive guide to the lang html attribute," https://sitechecker.pro/what-is-html-lang-attribute/, 2023, accessed: 2023-10-05.

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

Siti Hawa Apandi obtained her Master of Computer Science from Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA). Currently, she is pursuing a Ph.D. in the Faculty of Computing at UMPSA, with her primary research interests focusing on Soft Computing and Intelligent Systems.

Jamaludin Sallim is a Senior Lecturer in the Faculty of Computing at Universiti Malaysia Pahang Al-Sultan Abdullah. He earned his Ph.D. in Computer Science from Universiti Sains Malaysia. His areas of interest encompass intelligent process automation, software engineering, ICT Governance & Management and community informatics.

Rozlina Mohamed serves as a Senior Lecturer in the Faculty of Computing at Universiti Malaysia Pahang Al-Sultan Abdullah. She earned her Ph.D. in Computer Science from Aston University in Birmingham, UK. Her research interests encompass query processing, software requirements, software cost estimation, and information systems.