Vol 4 (2023) No 3

Jitsi Jurnal Ilmiah Teknologi Sistem Informasi

URL: http://jurnal-itsi.org ISSN 2722-4619 (Print) ISSN 2722-4600 (Online)

A Convolutional Neural Network (CNN) Classification Model for Web Page: A Tool for Improving Web Page Category Detection Accuracy

Siti Hawa Apandi[#], Jamaludin Sallim[#], Rozlina Mohamed[#]

[#] *Faculty of Computing, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia*
E-mail: sitihawa.apandi@gmail.com, jamal@ump.edu.my, rozlina@ump.edu.my

A B S T R A C T S

Game and Online Video Streaming are the most viewed web pages. Users who spend too much time on these types of web pages may suffer from internet addiction. Access to Game and Online Video Streaming web pages should be restricted to combat internet addiction. A tool is required to recognise the category of web pages based on the text content of the web pages. Due to the unavailability of a matrix representation that can handle long web page text content, this study employs a document representation known as word cloud image to visualise the words extracted from the text content web page after data pre-processing. The most popular words are shown in large size and appear in the centre of the word cloud image. The most common words are the words that appear frequently in the text content web page and are related to describing what the web page content is about. The Convolutional Neural Network (CNN) recognises the pattern of words presented in the core portions of the word cloud image to categorise the category to which the web page belongs. The proposed model for web page classification has been compared with the other web page classification models. It shows the good result that achieved an accuracy of 85.6%. It can be used as a tool that helps to make identifying the category of web pages more accurate.

Manuscript received August 21, 2023; revised August 25, 2023. accepted September 15, 2023 Date of publication September 30, 2023 International Journal, JITSI : Jurnal Ilmiah Teknologi Sistem Informasi licensed under a Creative Commons Attribution-Share Alike 4.0 International License

Keywords / Kata Kunci — ***Web page classification; document representation; word cloud image; deep learning; Convolutional Neural Network***

1. INTRODUCTION

Nowadays, access to the internet has become a basic necessity. Almost everyone is now connected to the internet. The number of internet users is expanding at a 5.7% annual rate, with more than 700,000 new users joining every day [1]. The majority of internet users, around 4.32 billion, accessed the internet through mobile devices, however desktops were also utilised to go online [1, 2]. With the increased utilization of mobile devices and computers, people use the internet more frequently for the purpose to interact with family and friends, search information, for entertainment and many more.

In the internet, there is an information space used to share information called as the World Wide Web (WWW) or simply the Web, where it contains billions of actives web pages. The internet has made the people easily obtain information by surfing the web pages. This situation has made the number of web pages serving people has grown at an alarming rate and continues to expand. With so much information available in the internet, it arises an issue on how to organize and manage the information effectively. For example, how do people know they have accessed the correct webpage that contain the information that they searched. In order to solve the issue, the concept of web page classification is established in order to organise and categorise web pages based on information content [3]. So that it is convenient for the people to obtain desired information accurately and quickly.

The Malaysian Communications and Multimedia Commission, popularly known as MCMC, has undertaken an internet user survey in Malaysia. The purpose of this survey is to monitor internet activities and understand the trends and tendencies among users in Malaysia. Based from the survey result, majority of internet users go online for entertainment purposes which are watch or download videos online and also play online games. The increase of internet access towards these online activities especially surfing the web page of game and also online video could lead to a serious problem which is internet addiction. The internet addiction is defined as excessive internet use that interferes with daily life.

As a way to detect the category of web pages, a web page classification model can be implemented to categorize Game and Online Video Streaming web pages. The web pages are classified based on its web page text content. In the existing work, the matrix representation has been used to convert the text into number to fed it to the web page classifier [5]. When there is a huge set of words in the document, the length of the feature vector will also grow in size which can be resource-intensive and expensive on the performance side [6, 7]. This study would like to explore the other document representation technique to represent the textual content of web page feature and how to represent text feature into image format as it can be directly fed to deep learning algorithm, named Convolutional Neural Network (CNN) **for the development of proposed** web page classification model.

The rest of the paper is organized as follows. Section 2 discuss the literature review of this study. Dataset website browsing records used in this study is explained in Section 3. Section 4 explains the process to develop a model of classifying web pages using CNN. Then, in Section 5, the comparison of the proposed classification model for web page with the other state-of-the-arts is discussed. Lastly, Section 6 concludes this study.

2. RESEARCH METHODOLOGY

2.1. Definition of Web page and Website

The Web keeps growing all the time. The January 2021 Web Server Survey by Netcraft found that there are more than one billion websites on the Web. However, more than 85% of all websites are inactive. This means that just about 10-15% of all websites on the Web are live and operational. According to the World Wide Web Size Project, at least 4.45 billion web pages are indexed in Google. Because the web page is distinct from the website, the overall number of web pages exceeds the total number of websites [8]. Web pages are website components that include the login page, about page, and contact page. The web page is linked to a specific URL under the respective domain. While the websites typically consist of one or more web pages connected together by hyperlinks to allow users to browse from one web page to another [9]. The website is identified by domain name. For example are wikipedia.org, google.com and amazon.com [10]. The webpage and website are viewed in a web browser, such as Chrome, Firefox and others [11]. The web browser can also display other documents such as a PDF document, but only a HTML document is termed as web page [9]. HTML is an abbreviation for Hyper Text Markup Language. It is the most often used markup language for constructing web pages. This study is going to focus on two categories of web pages which are Online Video Streaming and Game.

2.1.1. Online Video Streaming Website

Online Video Streaming website is an online platform that let the internet users to watch video clips, TV shows and movies streamed from the internet. The Online Video Streaming web page can save users time and less hassle. The internet users can watch whatever video that they want without having to download and store large video files, which would occupy a lot of storage space on user device. It can be seen based on the statistics that last year, one million additional streaming subscribers were added to 14 million users, accounting for approximately 78% of Malaysia's viewing population aged 15 and above [12]. Some of the most popular Online Video Streaming websites are YouTube, Viu, iFlix and many more.

2.1.2. Game Website

Game website, often known as a web browser game, is a computer game that is accessed via the internet using a web browser [13]. It is usually free-to-play. The game website has the advantage of not requiring a high-powered gaming PC to install the game because the web browser automatically gets the necessary content from the game's website. These games cover all genres and can be played by one or more players [14, 15]. There are thousands of the game website available, including new games and classic games re-made as the web browser game. Pac-Man is an example of a classic game and now available as the web browser game. Runescape is an example of popular MMORPG (Massively Multiplayer Online Role-Playing Game) web browser game [13]. It is a video game in which thousands of people compete in an online environment [16].

2.2. Internet Addiction

Some internet users have sufficient self-control, allowing them to use the internet appropriately and to deal with various concerns linked to their studies, work, or general life requirements. On the contrary, others have a negative usage of the internet when they get accustomed surfing irrelevant websites such as gaming, online video

social networking, blogging and others. This might negatively impact their overall performance as it is associated with many problems such as fewer sleeping hours, not feeling hungry, not being very active or not being able to focus properly [17].

The increase of internet use towards online activities especially surfing the Online Video Streaming and Game website could lead to a serious problem which is internet addiction. The term "internet addiction" can be defined as "individuals' incapability of controlling their internet usage, which in due course causes psychological, social, school and/or work difficulties in their life" [18, 19]. The students in colleges and universities are more exposed to the internet addiction because their age which is 20's, contribute to the majority of internet users based on the internet survey by the MCMC. Those online activities can be a distraction to the students learning environment. Researchers studying internet addiction in college students discovered that these students faced difficulties in finishing their homework, preparing for tests, and getting adequate sleep for their morning classes. These challenges were a result of spending too much time on the internet, specifically in browsing unrelated websites and playing interactive online games. As a result, the students are faced with severe academic problems and it disturb their daily routines [17]. In a study, half of the students who were interviewed after being dismissed for academic failure said that using the internet too much was a reason for their problems [20].

There is a confusion between internet addicts and workaholic behaviours, as the users spending too much time in front of the computer and always stay connected to the internet. Souligna [21] states that the user that can be categorized as an internet addict is regularly accessed the web page with the purpose of recreational internet use rather than for business or school purposes. Surfing the Online Video Streaming and Game website can be categorized as the recreational internet use. Thus, the user that frequently access those websites and spend too much time on it can be classified as the internet addicts.

The colleges and universities can take part to protect the students from getting the internet addiction. The colleges and universities need to monitor the online browsing behaviour of students who use the internet connection provided by the colleges and universities. The most common step that the colleges and universities can take is to set rules and limits the usage of the internet towards those specific category of web pages. The web page classification is important to determine the category of web pages.

2.3. Web Page Classification

The classification of a web page, which is also called categorization, means assigning the web page to one or more categories. The main issue of categorizing web pages can be broken down into smaller, more detailed problems which are subject classification, functional classification, sentiment classification and other types of classification. This study is going to focus on specific problem of web page classification which is subject classification, also known as topic classification, that classifies web pages based on their contents or topic of the web page [11, 22-24].

There are two ways to classify web pages: one is done manually and the other is done automatically. The regular way to classify web pages is by having a human expert assign them to different categories manually. This way of classifying web pages requires a significant amount of human labour and time. It cannot be utilised to classify web pages because the quantity of web pages on the internet is rising [5, 25]. Thus, the automatic method web page classification is the best way to handle the large number of web pages to be categorised.

The web page classifier is required in automatic web page classification in order to categorise the web page to one or more category labels.

In automatic web page classification, the web page classifier is needed in order to classify the web page to one or more category labels. The existing web page classification works have employed machine learning and deep learning as the web page classifiers. Their performance is briefly described below.

As the web page classifier, several machine learning algorithms were applied such as k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes, Decision Trees and Artificial Neural Network (ANN). It is also known as traditional machine learning algorithms. Safae, et al. [26] conducted a survey of the various machine learning algorithms used to categorise web pages. The machine learning algorithms that are reviewed includes k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes and Artificial Neural Network (ANN). After analysing the machine learning algorithms used in web page classification, it is discovered that the Artificial Neural Network (ANN), also known as the Neural Network (NN), performs better than the other machine learning algorithms. It is also noted that the Neural Network (NN) achieved the highest accuracy when compared to the other machine learning algorithms. It has been discovered that the existing algorithms function well with a small number of web pages. However, they get slower and less effective when they have to handle a lot of web pages [26].

Besides the traditional machine learning algorithms, deep learning techniques have started being used to classify web pages. Deep learning is a type of machine learning that makes use of a deep neural network. Deep learning can be a good option instead of traditional machine learning when dealing with big and complicated datasets [27]. Deep learning has the advantage of automatically discovering the features that will be used for categorization,

whereas machine learning requires these features to be provided manually [6]. Fig. 1 shows a comparison of deep learning and machine learning performance. From the figure, it can that deep learning shows a good performance compared to machine learning when deal with huge amount of data [28]. Thus, this study would focus about the deep learning.

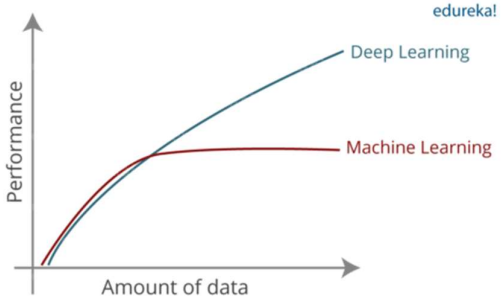

FIG 1. Comparison Performance Of Deep Learning And Machine Learning Based On Amount Of Data

FIG 2. Example Of Word Cloud Image

2.4. Word Cloud Images

There is an example that shows how a Convolutional Neural Network (CNN) has been used to classify handwritten digits. By providing label datasets of digits from one to 10, it used to train the CNN. Then, the CNN essentially extracts useful features from the given input automatically. The CNN is going to learn the pattern of digit in order to classify the new input given.

This study tries to implement this concept to the web page classification. There is other technique to represent text into image format.

Word cloud images are also called text clouds or tag clouds. In word cloud images, it displays a set of words retrieved from the document by displaying the words in the form of graphical representation. Words that are used often in a document will be shown as larger and bolder in the word cloud images. The frequent word depicts its important in the document to convey information. By looking at the big and bold words in the word cloud images, it makes easy to identify what is topic that belong to the document.

Fig. 2 shows an example of word cloud image [29]. Based on the figure, words like “movie” and “film” are more stand out by appear in big size and bold since those words are more frequently used in the document. By looking at the word cloud image, the document relates to movie topic

The word cloud image can be a powerful tool if it uses in the right setting. For example, when collect user feedback, there is a pile of information that need to be read. The user feedback needs to analyse to see what the user like most and what they are least. By applying the word cloud image, it can help identify key points from the user feedback [30].

The advantage of the word cloud image are: [31]

- It helps make it easier to see trends and patterns that might be hard to see if they are arranged in a table.
- It is easier to understand when information is presented in a way that looks interesting and uses pictures instead of just words. When the words are bigger in size, it means that the headlines are reported by more people or widely known.
- The word cloud image is more likely to be shared because it shows a visual representation of the information

3. DATASET WEBSITE BROWSING RECORDS

Due to there is no universally acknowledged datasets, this study utilizes self-gathered datasets to develop the proposed model for classifying web pages. In this study, the dataset used a real-world case which is the website browsing record of students in Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA). It is a good representation of realistic user behaviour when surfing the Internet [32]. The website browsing record of students in UMPSA is provided by the Information & Communication Technology Centre (PTMK) which is an organization in UMPSA that monitors internet use among UMPSA users by keeping records of URL web pages that are accessed.

The record of website browsing contains the records of URL web pages that are accessed by the students that collected in a week from 2019-03-17 (Sunday) until 2019-03-23 (Saturday). There are 40 Microsoft Excel files provided by the PTMK. Each Microsoft Excel file contains a record of website browsing for each UMPSA student. That means there are records of website browsing for 40 students. Each Microsoft Excel file of website browsing

contains more than thousands of URL web pages. This shows that each student accessed almost more than thousands of URL web pages in a week.

The description of types of information contained in the Microsoft Excel file of website browsing is shown in Table 1.

TABLE 1. Types of information contained in the file of website browsing record

	Column Name	Description of Column
1.	Rank	Id number of record. The latest record of website browsing has small id number while the older record of website browsing has big id number
2.	Username	Username of student
3.	Group name	Group of users which is student
4.	Source IP	Source IP of student
5.	Endpoint device	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 information about domain such as DNS, endpoint details, src port, port, protocol and mac

4. A CONVOLUTIONAL NEURAL NETWORK (CNN) CLASSIFICATION MODEL FOR WEB PAGES

After collect the dataset website browsing records, it needs to go through several activities to develop the CNN classification model for web pages which are data pre-processing, obtaining the required features, building the web page classification model and evaluating the web page classification model. The description of each activity is as follows.

4.1. Data Pre-processing

In general, raw dataset URL web pages are not arranged well for analysis because they might be not complete, not consistent, and difficult to comprehend. The most difficult step is to find and extract the relevant data from the dataset [33]. Thus, data pre-processing is a crucial step in cleaning, correcting, and preparing input data for mining [34]. The data pre-processing involves several steps to prepare data for analysis. These steps include cleaning the data, data normalization, transformation, feature extraction, and selection [33].

In this study, the data pre-processing consists of two activities which are data cleaning and web content pre-processing.

4.1.1. Data Cleaning

The initial stage of data pre-processing is to clean the data. The purpose of data cleaning is to remove redundant, useless, error, incomplete and inconsistent data [34]. The data cleaning is also performed to fetch cleaned data of active web pages [35].

In this study, the data cleaning is performed to remove duplicate URL, fetch active URL, keep the URL that contains HTML tags in the web page source code and include the web pages that has English content. The data cleaning steps is shown in Fig. 3.

The steps of data cleaning are described below.

- Remove records that contain null URLs from the excel file of website browsing. The total number of URL for raw dataset is 541,702. After remove the records with null URL, the total number of URL is 538,983.
- Keep the unique URL by removing the duplication of the URL. The result on 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. From the unique URL which is 29,444 records, the inactive URL has 14,716 records while the active URL has 14,728 records.

  Usually, web page is developed by using HTML tags. Besides the HTML tags, there is also XML tags. The most salient difference between HTML and XML tags is that the HTML tags used for the presentation of data while the XML tags used for transfer of data.

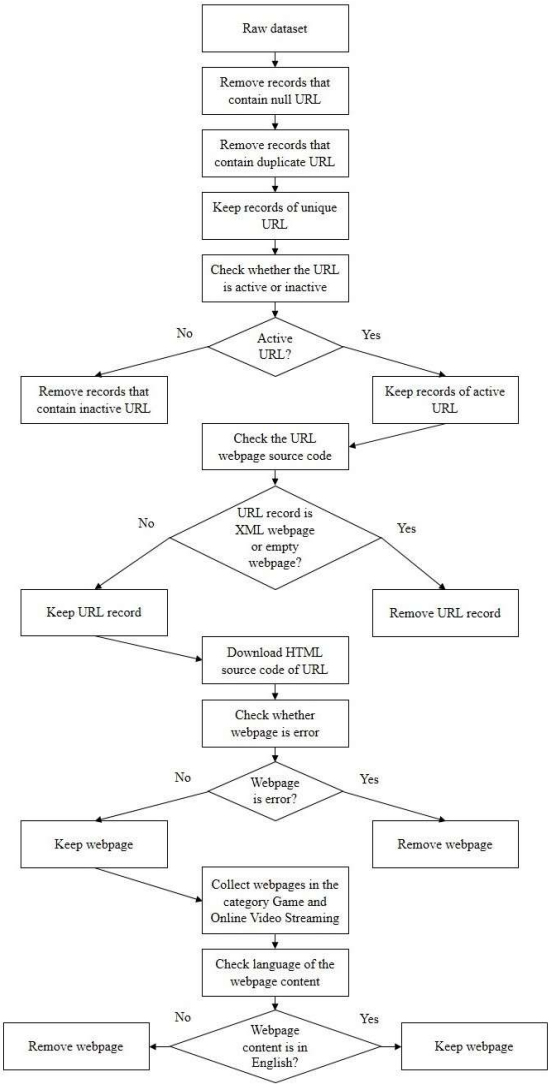

FIG 3. Steps Of Data Cleaning

Most web pages use HTML tags instead of XML tags to create their web pages [36]. Thus, this study keeps the URL that contains HTML tags in the web page source code. The data cleaning steps are as follows.

- Check each URL web page source code. Remove URLs that do not contain HTML tags in the web page source code including remove XML web page and empty web page.
- Download the HTML source code of each URL's web page.
- Check the web page's HTML source code whether it contains sentences related to the error page such as '404 not found' in the title of the web page. If it has occurred, the URL is removed.

  This study only collects dataset of the URL in the category Game and Online Video Streaming. After performing the data cleaning on the website browsing record provided by the PTMK, there are 216 Game web pages and 234 Online Video Streaming web pages. In order to expand the number of datasets in this study, the web pages on Games and Online Video are also search by ourselves. The total number of URL collected are 640 Game web pages and 407 Online Video Streaming web pages. The data cleaning step is doing again to only include the web pages that has English content, and the procedure is described below.
- Identify the language of text content based on the HTML lang attribute in the web page's HTML source code to ensure that the selected web pages' content is in English. If the HTML lang attribute is not defined in the web page's HTML source code, the web page content is identified manually to make sure only web page content in English is included. There are 475 Game web pages and 277 Online Video Streaming web pages

4.1.2. Web Content Pre-processing

Now, the dataset has the downloaded files of the web page's HTML source code. These files can be used to extract the content of the web page. It may be sufficient to analyse merely the information on the web page to determine what category a web page belongs to [37]. Thus, the textual content from the web page's HTML source code is extracted because the textual content is an integral feature used in the classification of web pages [38, 39]. Cleaning the noise of the web page's HTML source code is called web content pre-processing. The steps of the web content pre-processing are shown in Fig. 4. Once the web content pre-processing is complete, it gets rid of unimportant information like HTML tags, JavaScript, CSS code, and technical terms used on websites. These things do not contribute much to the analysis and are commonly used on websites [5, 40]. Finally, there will be left with words derived from the web page's HTML source code in the part title and body content which are more valuable and meaningful to represent the web page.

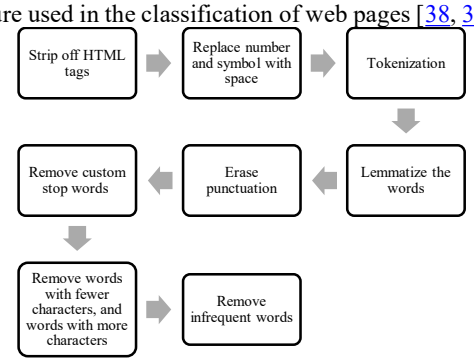

FIG 4. Steps Of Web Content Pre-processing

4.2. Obtaining The Required Features

The tokenized words on the web page produced from data pre-processing are analysed using a bag-of-words model. This technique, sometimes known as a "term-frequency counter," ignores the sequence of words in a document and only counts how many times each word shows up. It is the most used technique for document representation. Normally, the words in the web page are converted to matrix representation to fed it to the web page classifier [5]. However, there is a problem where it is too sparse which means there are too many zero in the matrix representation. Then, the word embedding appears to dense text features for the matrix representation. But it has a constraint which is each web page text content needs to be converted into an equal-length vector representation. If the web page text content is too long, it will be truncated to ensure that each text is equal in length after vectorization [6].

Instead of converting the words on the web page to matrix representation, this study uses a word cloud image to visualise the bag-of-words model. The word cloud image is a graphical representation of text data. It is sometimes called a text cloud or tag cloud. In the word cloud image, it can be seen that the most frequent or popular words more stand out by appearing in the centre, a different colour is used to represent the words and the size of the words is big compared to the other words. In Fig. 5, the words 'game' and 'play' are the most frequent words used on the Game web pages. Meanwhile, the words 'movie' and 'watch' are the most frequent words in the Online Video Streaming web page as shown in Fig. 6

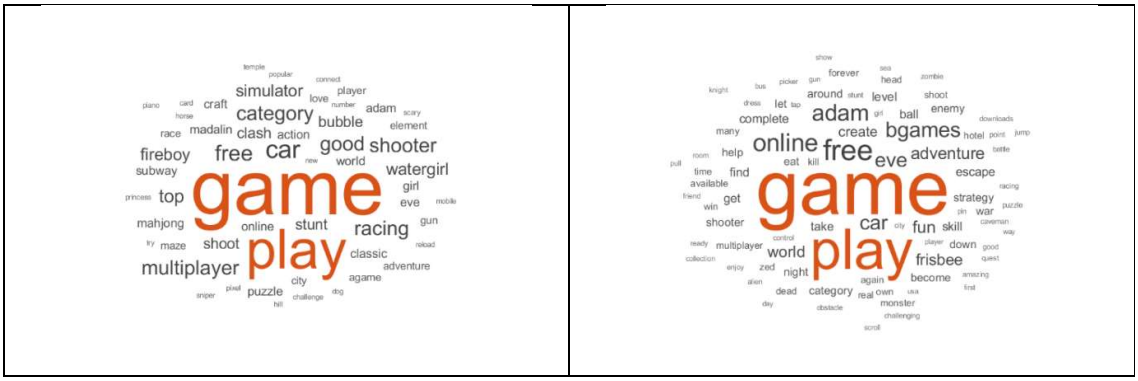

FIG 5. Sample Of Word Cloud Images For Game Web Page

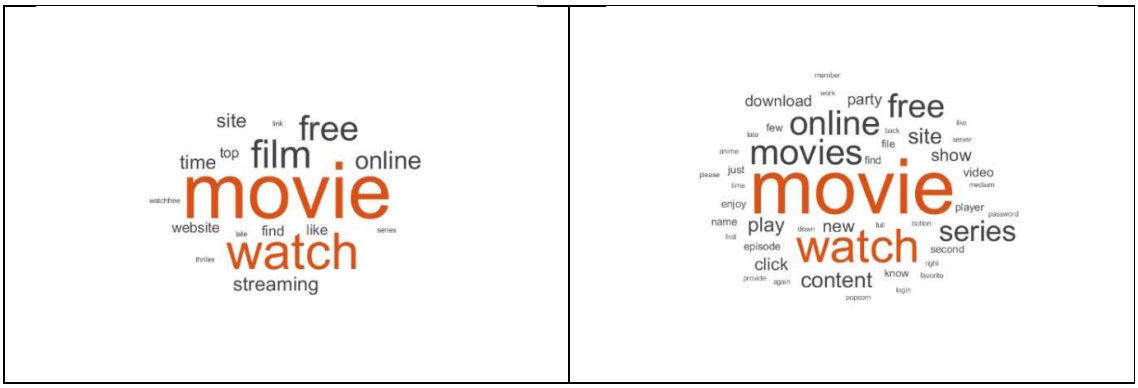

FIG 6. Sample Of Word Cloud Images For Online Video Streaming Web Page

4.3. Building The Web Page Classification Model

The word cloud images used to represent the web page will be utilised as the dataset to develop a CNN classification model for web pages. After performing the data pre-processing on the downloaded HTML source code web pages, the number of dataset word cloud images produced are 308 word cloud images to represent Game web pages and 141 word cloud images to represent Online Video Streaming web pages. The dataset will be split into two subsets. The first subset is used for training (or development). The second subset is used for validation (or performance evaluation) with the portion 80% dataset training and 20% dataset validation. In the experiment, all dataset are randomized.

Before the word cloud images are fed to the classifier, they need to be processed into a more easily processed format which are:

- The images are to be changed over into a grayscale structure. This is because coloured images have various colour channels, resulting in massive amounts of data to work with, making the process computationally intensive.
- The images are to be resized to make the small shape of the image so that there is a small number of pixels in the data to be processed.

Nowadays, CNN is among the dominant approach in the deep learning research community and it shows good performance [41]. In this study, the classification model for web pages is built using CNN. Usually, the CNN use image as an input dataset for training the model. Thus, the web page representation using the word cloud image will be fed as input to the CNN. For the development of a CNN-based classification model for web pages, it has used the MATLAB software and Deep Learning Toolbox.

This study constructs a simple architecture of the CNN as shown in Fig. 7. The proposed model starts with the image input layer. Then it continues with three convolutional layers, each of it have batch normalization layer and ReLU layer. The first and second convolutional layer has a max pooling layer with a pool size of 2. After the last convolutional layer, it follows with fully connected layer, softmax layer and classification output layer. It is noticed that the convolutional layer, batch normalization layer, ReLU layer and max-pooling layer are repeated to make the CNN capture more information about the image features. The number of layers added is relying on the image's complexity; thus, there are no rules on the number of layers to add.

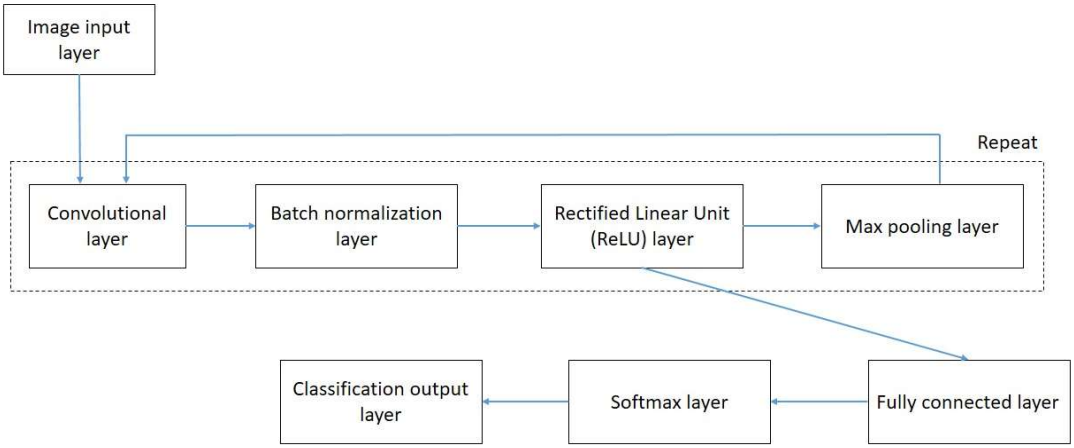

FIG 7. Structure Layers Of The Convolutional Neural Network (CNN)

There are different types of CNN layers, and their role is briefly described as follows.

1) Image input layer
   It specifies the input image's size which is 227 x 227 pixels with the channel size is one which means there is just one colour channel, which is grayscale in the input image.
2) Convolutional layer
   It extracts different features from the input images. In this study, a 3x3 sized filter is being sliding across the input image. The number of filters used on the convolutional layer is 8, 16 and 32. Between the input image and the filter, a mathematical procedure called convolution is applied. The result of this mathematical procedure is called as 'feature map' that gives information on the images' features. Later, this feature map is supplied to other layers, allowing them to learn about the input image's other features.
3) Batch normalization layer
   Over a mini-batch, each input channel is normalised. Batch normalisation is a technique for speeding up CNN training and reducing network initialization sensitivity.
4) Rectified Linear Unit (ReLU) layer
   It is the most commonly used activation function, and it performs a threshold procedure to each input element, setting any value less than zero to zero.
5) Max pooling layer
   It is a form of pooling layer that extracts the most value from the Pooling Kernel-covered region of the image.
6) Fully connected layer
   This layer aggregates all of the features learnt by the preceding layers throughout the image in order to discover larger patterns and classify images. The number of various classes that need to be classified

must be determined in this layer. For this study, the total amount of classes is two which are Game and Online Video Streaming.

7) Softmax layer
In the output layer, it serves as an activation function that predicts a multi-class classification probability distribution.

8) Classification output layer
Utilizing the probabilities supplied by the softmax activation function, this layer can assign one of the exclusive classes to each of the inputs.

Various training related parameters in CNN are defined to be used in the experiment which are Adam optimizer, 50 MaxEpochs and 32 MiniBatchSize. Once all the dataset, CNN architecture and training parameters have been defined; then the training process of the CNN is executed. Then, it will produce a network which is the proposed classification model for web pages. For the next step, the network will be used in the validation process to monitor the performance.

4.4. Evaluating The Web Page Classification Model

Fig. 8 shows the result of the confusion matrix. There are two classes which are Game and OVS stand for Online Video Streaming. The description of the result of the confusion matrix is as follows:

- 56 datasets are correctly classified as Game. This corresponds to 62.2% dataset.
- 21 datasets are correctly classified as OVS. This corresponds to 23.3% dataset.
- 7 of the OVS datasets are incorrectly classified as Game and this corresponds to 7.8% dataset.
- 6 of the Game datasets are incorrectly classified as OVS and this corresponds to 6.7% dataset.
- Out of 63 Game predictions, 88.9% are correct and 11.1% are wrong.
- Out of 27 OVS predictions, 77.8% are correct and 22.2% are wrong.
- Out of 62 Game datasets, 90.3% are correctly predicted as Game and 9.7% are predicted as OVS.
- Out of 28 OVS datasets, 75.0% are correctly classified as OVS and 25.0% are classified as Game.
- Overall, 85.6% of the predictions are correct and 14.4% are wrong.

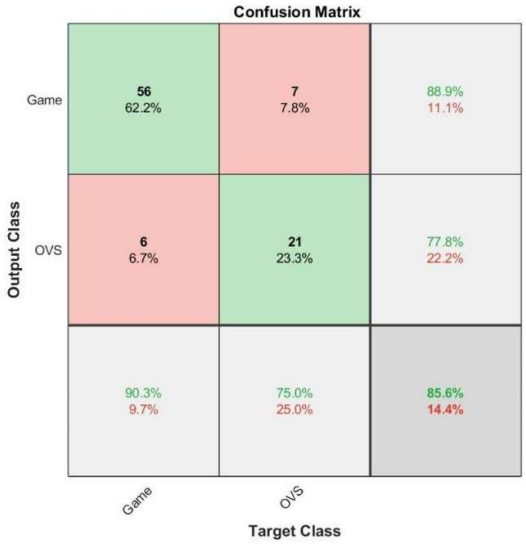

FIG 8. Result Of Confusion Matrix

The accuracy of the web page classification model determines how well it performs. The proposed model for classifying web pages attained an accuracy of 85.6%. This model can be used to automatically decide if a new web page is about Game or Online Video Streaming.

5. COMPARISON BETWEEN PROPOSED CLASSIFICATION MODEL FOR WEB PAGE AND EXISTING CLASSIFICATION MODEL

Table 2 shows a comparison of the proposed classification model for web page compared to the previous models. As can be seen, the accuracy of the proposed model: 0.86 is higher than the previous works [32] and [37] but lower than the previous work [6] since it combined two deep learning algorithms which are CNN and RNN.

TABLE 2. Comparison accuracy between the proposed model for classifying web page and the existing classification model

Classification Model For Web Page	Accuracy
Web page classification using Gated Recurrent Unit and Text Convolution [32]	0.81
Web page classification using RNN [37]	0.85
Web page classification using CNN and RNN [6]	0.90
Proposed web page classification model using CNN	**0.86**

All the model for classifying web page in the table utilized the text features contained in the web page. The previous model for classifying web page use word embedding to represent the web page for the model training

process. While in the proposed model for classifying web page use a different technique to represent the web page which is the word cloud image. The words contained in the web page after pre-processing are analysed using the bag-of-words. Then it is visualize using the word cloud image where the popular words are displayed in bigger size and placed in centre position of the word cloud image. The CNN is employed to analyse the word cloud image's word distribution to determine if a web page falls within the category of a Game or Online Video Streaming. By propose the different document representation technique for the text features, it shows a little bit of increase in accuracy performance of the model for classifying web page compared to the existing model. However, it does not compete with the accuracy performance of the existing model for classifying web page that use combination of deep learning algorithms.

6. CONCLUSIONS

This paper addressed about how a classification model for web pages was developed using a Convolutional Neural Network (CNN). The goal was to classify whether a web page is for playing games or for watching online videos. A word cloud image is used to visualise the tokenized words taken from the text content of a web page. The model for classifying web pages will examine the pattern of words in the word cloud image to predict the category of the web page. The experimental results show that the proposed model for classifying web pages has performed a good result that achieved an accuracy of 0.86. The performance of the proposed classification model for web pages is compared to the other web page classification model. Even if the proposed web page classification model's accuracy is not higher than that of the other web page classification model, the proposed classification model can be used as a tool that helps determine the category of web pages more accurately. For future work, there is a need for a method to select significant words from the text content of the web page that can provide more meaning on the purpose of the web page because the most popular words in the word cloud image sometimes do not offer any meaning and may affect the performance of the web page classification model. Besides that, this work can be improved by incorporating more dataset word cloud images related to Game and Online Video Streaming web pages into the experiment to develop the proposed model for classifying web pages using CNN.

Discussion is the basic explanation, relationship and generalization shown by the results. The description answers the research question. If there are doubtful results then display it objectively.

3.1. Specification

Use Times New Roman font type throughout the text, with the font size as exemplified in this writing guide. Spacing is single and the contents of the text or text using the left-right alignment (justified).

3.2. Page Size

The page size is A4 (210 mm x 297 mm). Page margins are 25 mm top-bottom, left and right.

3.3. Script Layout

An easy way to make layouts is to use this guide directly.

3.4 Headings

Use the style headings in this template directly. The style has been formatted in such a way as to provide appropriate heading spacing.

ACKNOWLEDGMENT

Thank you to the Information & Communication Technology Centre (PTMK) at Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) for granting permission to use UMPSA students' website browsing records as data collection in this study. The authors appreciate Dr. Azlee Zabidi who is a Senior Lecturer in the Faculty of Computing, at UMPSA, for his helpful advice on how to use MATLAB to develop a classification model for web pages using Convolutional Neural Network (CNN). The research described in this paper was made possible by UMPSA Grant: PGRS2003104**.**

REFERENSI

[1] Datareportal. (n.d., 1 October 2021). Digital Around The World. Available: https://datareportal.com/global-digital-overview

[2] J. Johnson. (2021, 1 October 2021). Worldwide digital population as of January 2021. Available: https://www.statista.com/statistics/617136/digital-population-worldwide/

[3] J. M. G. d. Costa, "Web Page Classification using Text and Visual Features," Master, Universidade de Coimbra, 2014.

[4] 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), 2019, pp. 238-243.

[5] A. Osanyin, O. Oladipupo, and I. Afolabi, "A Review on Web Page Classification," Covenant Journal of Informatics and Communication Technology, vol. 6, pp. 11-32, 2018.

[6] 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), 2019, pp. 1531-1535.

[7] 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), 2020, pp. 1-6.

[8] N. Huss. (2021, 1 January 2021). How Many Websites Are There Around the World? [2021]. Available: https://siteefy.com/how-many-websites-are-there/

[9] whatisblogger.com. (n.d., 1 October 2021). Web Page vs Website: What is the difference between a Web Page and Website (with Examples) | Website versus Web Page. Available: https://whatisblogger.com/website-versus-web-page/

[10] Wikipedia. (n.d., 1 October 2021). Website. Available: https://en.wikipedia.org/wiki/Website

[11] 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. 17472-17479, 2017.

[12] S. Birruntha. (2021, 1 October 2021). Top 5 streaming platforms in Malaysia. Available: https://themalaysianreserve.com/2021/01/05/top-5-streaming-platforms-in-malaysia/

[13] C. Hope. (2017, 1 October 2021). Browser-based game. Available: https://www.computerhope.com/jargon/b/browserbased-game.htm

[14] Wikipedia. (n.d., 1 October 2021). Browser game. Available: https://en.wikipedia.org/wiki/Browser_game

[15] J. Hadley and L. Morton. (2021, 1 October 2021). The best browser games to play right now. Available: https://www.pcgamer.com/best-browser-games/

[16] C. Hope. (2019, 1 October 2021). MMORPG. Available: https://www.computerhope.com/jargon/m/mmorpg.htm

[17] F. Cao and L. Su, "Internet addiction among Chinese adolescents: prevalence and psychological features," Child: care, health and development, vol. 33, pp. 275-281, 2007.

[18] K. S. Young and R. C. Rogers, "The relationship between depression and Internet addiction," Cyberpsychology & behavior, vol. 1, pp. 25-28, 1998.

[19] R. A. Davis, "A cognitive-behavioral model of pathological Internet use," Computers in human behavior, vol. 17, pp. 187-195, 2001.

[20] G. M. U. S. H. Services. (n.d., 1 April 2021). Internet Addiction. Available: https://shs.gmu.edu/healthed/internet-addiction/

[21] S. Souligna, "A Browser Based Intervention Approach Towards Managing Internet Addiction Disorder," Master thesis, Auckland University of Technology, 2017.

[22] X. Qi, "Web page classification and hierarchy adaptation," Doctor of Philosophy in Computer Science, Lehigh University, 2012.

[23] P. V. Nainwani and P. Prajapati, "Comparative Study of Web Page Classification Approaches," International Journal of Computer Applications, vol. 179, pp. 6-9, 2018.

[24] [24] 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, pp. 112-122, 2010.

[25] E. Suganya and D. Vijayarani, "Web Page Classification in Web Mining Research-A Survey," International Journal of Innovative Research in Science, Engineering and Technology, vol. 6, pp. 17472-17479, 2017.

[26] 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), 2018, pp. 220-226.

[27] 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, pp. 2797-2819, 2017.

[28] A. Bakshi. (2021, 20 December 2021). What is Deep Learning? Getting Started With Deep Learning. Available: https://www.edureka.co/blog/what-is-deep-learning

[29] T. T. Nguyen, K. Chang, and S. C. Hui, "Word cloud model for text categorization," in 2011 IEEE 11th International Conference on Data Mining, 2011, pp. 487-496.

[30] B. Labs. (2014, 24 March 2022). Word Clouds & the Value of Simple Visualizations. Available: https://boostlabs.com/blog/what-are-word-clouds-value-simple-visualizations/

[31] R. Kusumaningrum and S. Adhy, "WCLOUDVIZ: Word Cloud Visualization of Indonesian News Articles Classification Based on Latent Dirichlet Allocation," Telkomnika, vol. 16, pp. 1752-1759, 2018.

[32] 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), 2018, pp. 134-141.

[33] H. Jamshed, M. 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, pp. 206-221, 2019.

[34] M. J. H. Mughal, "Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview," Information Retrieval, vol. 9, 2018.

[35] 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), 2016, pp. 312-317.

[36] 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.

[37] E. Buber and B. Diri, "Web Page Classification Using RNN," Procedia Computer Science, vol. 154, pp. 62-72, 2019.

[38] M. Hashemi, "Web page classification: a survey of perspectives, gaps, and future directions," Multimedia Tools and Applications, vol. 79, pp. 11921-11945, 2020.

[39] 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), 2017, pp. 0320-0323.

[40] B. A. Alahmadi, P. A. Legg, and J. R. Nurse, "Using internet activity profiling for insider-threat detection," in Proceedings of the 17th International Conference on Enterprise Information Systems (WOSIS-2015), 2015, pp. 709-720.

[41] F. De Fausti, F. Pugliese, and D. Zardetto, "Toward Automated Website Classification by Deep Learning," Rivista di Statistica Ufficiale, vol. 3, pp. 9-50, 2020.