INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

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

Automatic Topic-Based Web Page Classification Using Deep Learning

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

[a] *Faculty of Computing, Universiti Malaysia Pahang, 26600 Pekan, Pahang, Malaysia*
[b] *Student Development Section, Universiti Kuala Lumpur, Branch Campus Malaysia France Institute, Selangor, Malaysia*
Corresponding author: [*]sitihawa.apandi@gmail.com

***Abstract*—The internet is frequently surfed by people using smartphones, laptops, or computers in order to search for information online on the web. The increase in information on the web has made the web pages grow daily. The automatic topic-based web page classification is used to manage excessive web pages by classifying them into different categories based on the web page content. Different machine learning algorithms have been employed as web page classifiers to categorize the web pages. However, there is a lack of studies that review the classification of web pages using deep learning. This study reviewed the automatic topic-based classification of web pages utilizing deep learning that many key researchers have proposed. The relevant research papers are selected from reputable research databases. The review looked at the dataset, features, algorithm, pre-processing used in the classification of web pages, document representation technique, and performance of the web page classification model. The document representation technique used to represent the web page features is an important aspect in the classification of web pages as it affects the performance of the web page classification model. The integral web page feature is the textual content. Based on the review, it was found that the image-based web page classification showed higher performance compared to the text-based web page classification. Due to the lack of matrix representation that can effectively handle long web page text content, a new document representation technique, word cloud image, can be used to visualize the words extracted from the text content web page.**

***Keywords*— Deep learning; document representation technique; machine learning; web page classification; web page classifier.**

Manuscript received 31 Jan. 2023; revised 17 Apr. 2023; accepted 21 May 2023. Date of publication 30 Nov. 2023.

I. Introduction

With the advent of the internet, information has become more widely available and accessible online. The internet has enabled people to easily obtain information by surfing the web pages. In order to supply information online to internet users, many web pages were created on the web. This situation has made the number of web pages growing very quickly. With so much information available on the internet, the issue of organizing and managing the information effectively arises. For example, how people know they have accessed the correct webpage containing the information they searched. As a solution, web page classification is introduced to categorize web pages according to information content [1] so that it is convenient for people to obtain the desired information accurately and quickly.

Traditional manual and automatic approaches are both available for the classification of web pages. The traditional manual approach of classifying web pages put in human labor to manually place the web pages into different categories by the experts. It takes a lot of time and quite labor-intensive. Due to too many web pages day, the traditional manual approach cannot be utilized to classify web pages [2], [3]. Therefore, utilizing the automatic approach is the best way to handle the rising number of web pages that need to be categorized.

In automatic web page classification, the web page classifier is required to classify the category label of the webpages. As the web page classifier, machine learning has been utilized. Machine learning has shown good performance with small quantity of webpages. However, when dealing with large quantity of webpages, it became non effective [4]. In order to overcome the limitation of machine learning, the deep learning is introduced [5]. Thus, this study would focus on the classification of web pages using deep learning.

Researchers have proposed a variety of automatic web page classification systems, in which different techniques have been formulated to tackle the problem of the lack of accuracy in the classifiers' performance. Most classification algorithms' accuracy is based on the quality and quantity of training data, which rely on the document representation technique [2]. A few things need to be considered when developing a web page classification model. Four factors affect web page classification: features, algorithms, web page content pre-processing, dataset selection, and generation [6].

The contribution of this study is to review the previous works of automatic topic-based classification of web pages utilizing deep learning that the other researchers have proposed. From that, this study can analyze the previous works in terms of the dataset, features, algorithm and pre-processing used to classify the web pages and also the performance of the web page classification model.

A. Web Page Classification

The process of allocating a web page to one or more category labels is called web page classification, or sometimes referred to as web page categorization. The basic problem of classifying web pages can be subdivided into more specialized problems such as subject classification, functional classification, sentiment classification, and other sorts of classification. This study will concentrate on a specific aspect of web page classification known as subject classification, sometimes known as topic classification, in which web pages are classified based on their contents or the topic of the web page [3], [6], [7].

Classification of web pages plays a vital role in effective internet use, spam filtering, and a variety of other applications. Finding relevant results quickly from millions of websites is a critical issue that search engines must address. As a result, certain search engines needed to do topic-based classification on web pages to provide better user results. Furthermore, web pages must be classified to establish internet usage policies for institutions or individuals. Cyber security can also employ web page classification to prevent dangerous websites from being presented to the user [8].

The classification of web pages differs from the classification of text because web pages contain semi-structured data, which are HTML tags that are used to structure and organize the information that will be displayed through the web browser [2], [8]–[10]. Furthermore, hypertext links web pages together [2]. The web page contains noises, which are progressively irrelevant pieces of information. These noises will greatly interfere with feature extraction and reduce classification accuracy [11]. This issue can be solved by data pre-processing. It is an important stage in the web page classification process to make a better document representation by removing irrelevant and noisy features, which would negatively impact the web page classifier's accuracy, and speed and reduce overfitting issues [2]. Due to this, the document representation technique is important to represent the selected web page features since it may affect the accuracy performance of the model for classifying web pages.

II. Material and Method

The relevant research papers that discuss automatic topic-based web page classification using deep learning need to be searched via Scopus, Web of Science, and Google Scholar. The keywords are listed in the research papers' title, keyword, and abstract. Table I shows the keywords used to search in the research databases, and various search techniques like Boolean operator, phrase searching, and truncation are applied. A brief explanation of those search techniques is as follows:

- Boolean operators are logical search operators such as OR, AND that allow users to narrow or broaden their search.
- Phrase searching means double quotes are used with the keywords to get more relevant results compared to searching without double quotes.
- Truncation is used in the research databases, meaning that the truncation symbol, the asterisk (*), is placed at the word ending. As seen in the table, for example, web* will search for results containing website, webpage, and webpages. The truncation makes it possible to search for multiple word forms simultaneously, enhancing the number of search results discovered.

TABLE I
Keywords Used in the Research Databases

Research Database	Keywords Used
Scopus	TITLE-ABS-KEY (("web* classification*" OR "web page* classification*" OR "web* categorization*" OR "web page* categorization*" OR "URL* classification*") AND ("deep learning" OR "deep neural network*" OR "* neural network*"))
Web of Science	TS=(("web* classification*" OR "web page* classification*" OR "web* categorization*" OR "web page* categorization*" OR "URL* classification*") AND ("deep learning" OR "deep neural network*" OR "* neural network*"))
Google Scholar	allintitle: (("web* classification*" OR "web page* classification*" OR "web* categorization*" OR "web page* categorization*" OR "URL* classification*") AND ("deep learning" OR "deep neural network*" OR "* neural network*"))

TABLE II
The Criteria of Inclusion and Exclusion to Select Research Papers

Criteria	Inclusion	Exclusion
Year of publications	2017 – 2022	2016 and before
Language	English	Non-English
The study's nature	Focus on the proposed topic-based web page classification model using deep learning algorithm	- Do not focus on the proposed topic-based web page classification model using deep learning algorithm - Focus on the URL classification in terms of phishing or not phishing and malicious or benign

The criteria utilized to determine the research papers' inclusion and exclusion are shown in Table 2. First, the research papers selected must be published within a period of five years, which is between 2017 and 2022. Second, the search focuses on papers published in English, and non-English publications are excluded to avoid additional translation work. Lastly, the research papers must discuss topic-based web page classification models using deep learning algorithms. The research papers that do not focus on this are excluded. Besides that, the research papers on URL classification, which are phishing, non-phishing, malicious, or benign, are also excluded to avoid confusion as users could not know what category the web page URL has.

Fig. 1 below shows the process flow for finding and choosing the research papers to be analyzed in this study.

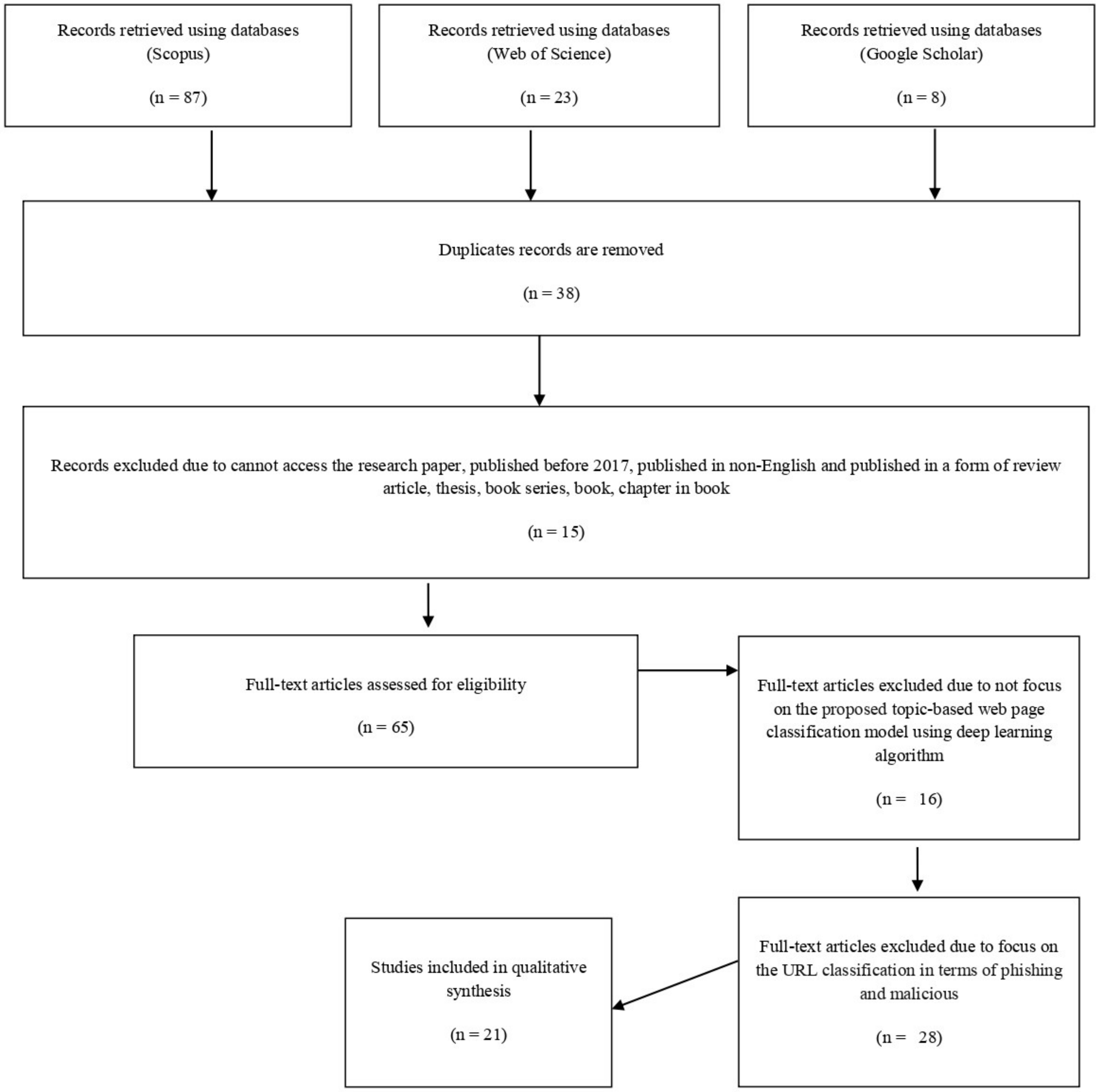
Records retrieved using databases (Scopus) (n = 87)
Records retrieved using databases (Web of Science) (n = 23)
Records retrieved using databases (Google Scholar) (n = 8)
Duplicates records are removed (n = 38)
Records excluded due to cannot access the research paper, published before 2017, published in non-English and published in a form of review article, thesis, book series, book, chapter in book (n = 15)
Full-text articles assessed for eligibility (n = 65)
Full-text articles excluded due to not focus on the proposed topic-based web page classification model using deep learning algorithm (n = 16)
Full-text articles excluded due to focus on the URL classification in terms of phishing and malicious (n = 28)
Studies included in qualitative synthesis (n = 21)

Fig. 1 Flow for finding and choosing the research papers

The total number of specific papers found in the three research databases is 118. There are 38 duplicate research papers found, which were removed. Then, the research papers were screened with the inclusion and exclusion criteria. There were 15 research papers that were further excluded due to the reasons of accessibility and published before 2017 in non-English and published in the form of a review article, a thesis, a book series, a book, and a chapter in a book. The remaining 65 research papers are analyzed for their eligibility for this study by reading the abstracts. This study would focus on the research that discusses the proposed topic-based web page classification model using a deep learning algorithm that other researchers have introduced. After analyzing the research papers by reading the abstracts, some of the research papers were excluded for not focusing on and discussing the proposed topic-based web page classification model using a deep learning algorithm. Besides that, research papers discussing URL classification in terms of phishing or not phishing and malicious or benign are also excluded as they only classify the URL as phishing or not phishing and malicious or benign without knowing what actual category the URL belongs to. Based on the exclusion criteria, 44 research papers were excluded. Finally, 21 research papers are suitable to be analyzed in this study.

III. Results and Discussion

Table 3 presents comparative works on the topic-based web page classification model using a deep learning algorithm organized by year in increasing order. The comparison takes a look at the dataset, features, algorithm, pre-processing used in the process to classify the web pages, document representation technique, and also the performance of the web page classification model.

TABLE III
COMPARATIVE EXISTING WORKS OF THE TOPIC-BASED WEB PAGE CLASSIFICATION MODEL USING DEEP LEARNING ALGORITHM

Reference (Year)	Dataset	Feature	Algorithm	Document representation technique	Performance
[12] (2017)	Artificially generated datasets	Image	Convolutional Neural Network (CNN)	Vectors of image pixels	Accuracy: 90%
[13] (2017)	Artificially generated datasets	Image	Convolutional Neural Network (CNN)	Vectors of image pixels	Accuracy: 97.61%
[14] (2018)	Publicly available web directories	Image	Convolutional Neural Network (CNN)	Vectors of image pixels	Accuracy: 89%
[15] (2018)	Artificially generated datasets	Title tag Description meta tag Keyword meta tag	Convolutional Neural Network (CNN) Gated Recurrent Unit (GRU)	Word embedding using Word2Vec	Accuracy: 81.1%
[8] (2019)	Publicly available web directories	Title tag Description meta tag Keyword meta tag	Recurrent Neural Network (RNN) Long Short Term Memory (LSTM)	Word embedding using Glove	Accuracy: 85%
[16] (2019)	Artificially generated datasets	Image	Convolutional Neural Network (CNN)	Vectors of image pixels	Accuracy: 98.97%
[10] (2019)	Publicly available web directories	Title tag Keyword meta tag Text content	Convolutional Neural Network (CNN)	Word2Vec combined with the Skip-Gram model	F-measure: 0.94
[17] (2019)	Artificially generated datasets	Image	Convolutional Neural Network (CNN)	Vectors of image pixels	Accuracy: 86.08%
[11] (2019)	Artificially generated datasets	Title tag Description meta tag Text content	Convolutional Neural Network (CNN) Recurrent Neural Network (RNN)	Word embedding	Accuracy: 90%
[18] (2019)	Artificially generated datasets	Title tag Description meta tag Keyword meta tag URL Text content	Convolutional Neural Network (CNN)	Word embedding	Accuracy: 85%
[19] (2020)	Publicly available web directories	HTML tag structure Text content	Long Short Term Memory (LSTM)	Word embedding using Word2Vec	Accuracy: 94.2%
[20] (2020)	Publicly available web directories	Title tag Text content	Convolutional Neural Network (CNN)	Word embedding	F-measure: 0.9561
[21] (2020)	Publicly available web directories	URL	Convolutional Neural Network (CNN) Gated Recurrent Unit (GRU)	Word embedding using Glove	Accuracy: 82.04%
[9] (2020)	Artificially generated datasets	Image	Convolutional Neural Network (CNN)	Vectors of image pixels	Accuracy: 86%
[22] (2021)	Artificially generated datasets	Text content	Convolutional Neural Network (CNN)	BERT word embedding	Accuracy: 86.71%
[23] (2021)	Publicly available web directories	Text content	Convolutional Neural Network (CNN) Long Short Term Memory (LSTM)	Vectors	Accuracy: 95%

Reference (Year)	Dataset	Feature	Algorithm	Document representation technique	Performance
[24] (2021)	Artificially generated datasets	Title tag Description meta tag Text content	Convolutional Neural Network (CNN)	Word embedding	F-measure: 0.85
[25] (2021)	Publicly available web directories	Title tag Text content	Long Short Term Memory (LSTM)	Word embedding using Glove	Accuracy: 85.32%
[26] (2022)	Artificially generated datasets	URL	Convolutional Neural Network (CNN) Long Short Term Memory (LSTM)	Vectors	F-measure: 0.9343
[27] (2022)	Artificially generated datasets	Title tag Description meta tag Text content	Convolutional Neural Network (CNN)	Word embedding vectors	F-measure: 0.81
[28] (2022)	Publicly available web directories	Title tag Description meta tag	Convolutional Neural Network (CNN)	Word embedding	Accuracy: 79.51%

For the dataset used in the web page classification, most existing works used artificially generated datasets compared to publicly available web directories, due to the lack of a unified and recognized web page dataset [11], [16]. This relates to the observation by Hashemi [29] that most studies reported on using artificially generated datasets to get datasets for web page classification. They crawl websites according to predetermined categories using an algorithm and filter out noise. This strategy has the advantage of generating the dataset automatically without the need for human labor and giving users the flexibility to modify the generated dataset's properties [6].

The web page contains various data types such as text content, HTML tags, images, audio files, and videos [8], [14]. Two types of features are used in the web page classification: image and text. The text features can be derived from many data sources: textual content, HTML tags, and web page URL[9], [14]. The list of features can be classified as on-page features directly located on the page needing classification. It might be enough to analyze merely the information on the web page to determine the type of web page [8].

For the features used for the classification of web pages, there are:

- seven studies that used a combination of HTML tag structure and text content, as shown in [10], [11], [19], [20], [24], [25], [27]
- six studies that used images as shown in [9], [12]–[14], [16], [17]
- three studies that used the feature of HTML tags structure as shown in [8], [15], [28]
- two studies that used each feature of text content as shown in [22], [23]
- two studies that used URL features as shown in [21], [26]
- one study used a combination of HTML tags structure, text content and URL as shown in [18].

From here, it can see that the most popular web page feature used for classification is text compared to image. This related to the statement that text features have become an integral component of web page classification and have the greatest impact on a web page type[29], [30]. A little work was found for web page classification based on URL feature because URLs lack information to represent web pages' content[15].

For the web page classification that uses text features, it needs to do the pre-processing, including removing punctuation, special characters, numbers, and stop words, converting alphabetical characters to lowercase, tokenizing, and lemmatizing[19], [31]. After the data pre-processing, every word needs to be converted into a word vector to feed to the classification algorithm. The data representation for text features used in the existing works is by using word embedding techniques such as Word2Vec, Glove, and BERT. There is a limitation in using the existing data representation technique whereby the feature vector's length will increase if the document has a large number of words, which might be expensive in terms of performance and use of resources [11], [32]. As a result, the text needs to be converted into an equal-length vector representation, and the remaining text will be truncated if too long [11]. From this, it can be seen that other data representation techniques need to be explored to represent the text.

For the web page classification that uses image features, the data pre-processing steps are:

- remove images with discriminatory content, such as navigation icons, banners, and advertisements [9], [12]–[14], [16]
- reject images whose dimensions are outside a specific range [13]
- resize the image to a fixed size [17]

Then, the image features use pixels of images as input and are fed to the deep learning algorithm.

Most of the existing works use supervised deep learning as the web page classifier for the algorithm. The majority of the existing works use Convolutional Neural Network (CNN) for image classification, as shown in [9], [12]–[14], [16], [17] and also text classification as shown in [10], [18], [20], [22], [24], [27], [28]. The other supervised deep learning being used in the existing works is Recurrent Neural Network (RNN). The RNNs come in two varieties: Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU). Usually, the RNN includes the LSTM and GRU, used for text classification, as shown in [8], [19], [25]. Some studies on the existing works use a combination of CNN and RNN, including the LSTM or GRU for text classification, as shown in [11], [15], [21], [23], [26]. From this, it can be seen that the

CNN is flexible to use for image and text-based web page classification.

The key performance indicators for the web page classification model consist of precision, recall, F-measure and accuracy. Majority of the existing works use accuracy to report the performance of web page classification model as shown in [8], [9], [11]–[19], [21]–[23], [25], [28]. The other metric used by the existing works to report the performance is F-measure as shown in [10], [20], [24], [26], [27]. In terms of performance value, the existing works use image classification shows high performance compared to text classification.

One popular technique currently employed in deep learning research is the Convolutional Neural Network (CNN), which excels in certain applications, such as image recognition [33]. For example, CNN has been used for classifying image of handwritten digits. By providing label datasets of digits image from one to 10, it is used to train the CNN. Then, the CNN essentially extracts useful features from the given input automatically. The CNN will learn the pattern of digit image to classify the new input given. There is a specialty of deep learning where it excels in automatically discovering the features that will be used for categorization, compared to machine learning, which needs these features to be provided manually [11].

The review shows that the web page classification model achieved good performance when it uses image as input to be fed to the web page classifier. This study tries to apply this concept to web page classification by exploring new techniques to represent texts in image format. It is called word cloud images or also known as text clouds or tag clouds. By presenting the words as a graphical representation, word cloud images show a group of words that have been extracted from the document. The size and boldness of the word display in the word cloud images will increase with the frequency with which a word appears in a document. The frequently used words depict its importance in the document to convey information. By looking at the big and bold words in the word cloud images, it is easier to identify what topic belongs to the document. Fig. 2 shows an example of word cloud image. Based on the figure, words like "movie" and "film" stand out more by appearing in big size and bold since those words are more frequently used in the document. The document relates better to the movie topic by looking at the word cloud image.

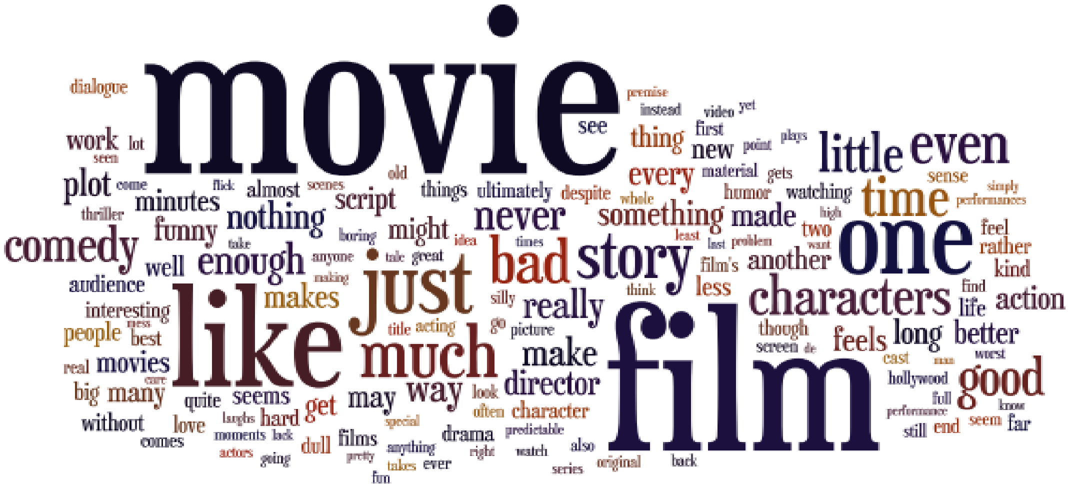

Fig. 2 Example of word cloud image

IV. Conclusions

This study has reviewed the automatic topic-based classification of web pages utilizing deep learning done by researchers in the field. In conclusion, this study will summarize the findings from previous research and outline potential future directions for web classification. Many factors need to be considered when developing a model for classifying web pages in terms of dataset, features, algorithm and document representation technique as it will affect the performance. It faced a challenge to choose the selected web page features for classification. This is due to the fact that web pages contain an increasing amount of noise. The data pre-processing can help to fix this problem. The review found that when an image is utilized as input for the web page classifier during the classification process, the web page classification model's performance improves. The word cloud image can be used as the new document representation technique for the text to overcome the limitation of the matrix representation. It is noticed that the phase of data pre-processing and document representation technique are significant factors in classifying web pages because it influence how well the web page classification model performs. Future work will focus on the implementation of the word cloud image to represent the text feature of the web pages. A Convolutional Neural Network (CNN) can develop a model for classifying web pages. Following that, the web page's category will be determined using the word cloud image's pattern of displayed words. The performance of the proposed technique can be compared with the existing techniques to determine if it can produce good results in classifying web pages.

Acknowledgment

This paper's research was funded by a Universiti Malaysia Pahang: PGRS2003104 grant.

References

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

[2] 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.

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

[4] 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.

[5] 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.

[6] X. Qi, "Web page classification and hierarchy adaptation," Ph.D dissertation, Lehigh Univ., Bethlehem, 2012. [Online]. Available: http://wume.cse.lehigh.edu/pubs/qi-dissertation.pdf

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

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

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] K. Maladkar, "Content based hierarchical URL classification with Convolutional Neural Networks," in *2019 International Conference on Information Technology (ICIT)*, IEEE, 2019, pp. 263–266.

[19] L. Deng, X. Du, and J. 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.

[20] 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.

[21] R. Rajalakshmi, H. Tiwari, J. Patel, A. Kumar, and R. Karthik, "Design of kids-specific URL classifier using Recurrent Convolutional Neural Network," *Procedia Comput Sci*, vol. 167, pp. 2124–2131, 2020.

[22] S. Alqaraleh, H. M. N. Sirin, and F. Ozkan, "Performance comparison of Turkish web pages classification," in *2021 Innovations in Intelligent Systems and Applications Conference (ASYU)*, IEEE, 2021, pp. 1–5.

[23] S. Suleymanzade and F. Abdullayeva, "Full content-based web page classification methods by using deep neural networks," *Statistics, Optimization & Information Computing*, vol. 9, no. 4, pp. 963–973, 2021.

[24] C.-G. Artene, M. N. Tibeică, and F. Leon, "Using BERT for multi-label multi-language web page classification," in *2021 IEEE 17th International Conference on Intelligent Computer Communication and Processing (ICCP)*, IEEE, 2021, pp. 307–312.

[25] A. K. Nandanwar and J. Choudhary, "Semantic features with contextual knowledge-based web page categorization using the GloVe model and stacked BiLSTM," *Symmetry (Basel)*, vol. 13, no. 10, p. 1772, 2021.

[26] Z. Li, S. Zhang, J. Yin, M. Du, Z. Zhang, and Q. Liu, "Fighting against piracy: An approach to detect pirated video websites enhanced by third-party services," in *2022 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, 2022, pp. 1–7.

[27] C.-G. Artene, D.-D. Vecliuc, M. N. Tibeică, and F. Leon, "An experimental study of Convolutional Neural Networks for functional and subject classification of web pages," *Vietnam Journal of Computer Science*, vol. 9, no. 04, pp. 435–453, 2022.

[28] A. W. Murdiyanto and M. Habibi, "Analysis of deep learning approach based on Convolution Neural Network (CNN) for classification of web page title and description text," *Compiler*, vol. 11, no. 2, pp. 51–58, 2022.

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

[30] 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.

[31] 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.

[32] 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.

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