IV. Conclusion

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

Acknowledgment

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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