A Multilevel N-gram Model with Naïve Bayes Classification of Personal Web History Datasets

Lee Chin Kho*, Ji Liang Hau, Sze Song Ngu, Annie Joseph, Nordiana Rajaee, and Siti Kudnie Sahari
Department of Electrical and Electronic Engineering, University Sarawak Malaysia, Kota Samarahan, Sarawak, Malaysia

Abstract

The rapid expansion of 4G and 5G networks has accelerated the proliferation of Internet-connected devices, leading to a massive increase in Internet of Things (IoT) traffic and the generation of diverse big data. Big data analytics has been widely adopted across healthcare, gaming, cybersecurity, business, and other sectors to extract actionable insights, uncover hidden patterns, and support informed decision-making on their targeted datasets. However, big data analytics of personal history datasets is a passive and underrated field. This paper addressed the gap by proposing a novel multilevel N-gram model combined with a Naïve Bayes classifier to classify the personal website history datasets. In the first stage, URL strings are decomposed into multiple N-gram levels (unigrams, bigrams, trigrams) to capture both simple lexical features and contextual patterns. In the second stage, the extracted features are classified using the Naïve Bayes algorithm, which applies Bayes' theorem under the assumption of conditional independence to compute category probabilities. Empirical validation on a standardised dataset demonstrates that the proposed approach achieves an average F1-score of 88%, outperforming existing baseline methods documented in prior literature. These findings highlight the effectiveness of the proposed method for big data analysis of web usage, particularly for personal history datasets.

Keywords: *Big Data Analytics, Naïve Bayes Classification, N-gram modeling, and personal datasets*

1. Introduction

Technology is moving forward at a speed that is difficult for many to follow. In some countries, 5G networks are still being tested and expanded, while at the same time, early work on 6G has already begun. Wireless communication is no longer a luxury but part of everyday life, changing the way people interact and work. Devices such as smartphones, laptops, and even cashiers' machines in supermarkets are constantly linked to the Internet, producing large amounts of data. Most of this data is stored in cloud systems, where it grows into what we now call big data. This expression was first used by Michael Cox and David Ellsworth in 1990 when dealing with supercomputers' challenges [1].

The rapid advancement in networking technology has led to better data storage capabilities, overcoming limitations caused by the influx of messy data in IoT traffic. As a result, the amount of data has increased in volume and complexity. Researchers have introduced a method called big data analysis to extract useful information from this enormous, high-velocity, and diverse data. This involves collecting, organising, and analysing large datasets to identify patterns and other valuable information

* Corresponding author. Tel.: +60-128880163
E-mail address: lckho@unimas.my

Manuscript History:
Received 22 September, 2025, Revised 27 March 2026, Accepted 22 April, 2026, Published 30 April, 2026.

https://doi.org/10.33736/jese.10735.2026

e-ISSN:

hidden within them. Big data analysis is a new technique that requires an intellectual system to sort out the crucial features of big data. It can reduce complexity and handle the cognitive burden in a knowledge-based society, benefiting many people. The most important aspect of big data analysis is identifying the significant features in the data. The rapid evolution of big data analysis has led to substantial growth in e-business and the deepening of global connections. Governments also use big data analysis to better serve their citizens, for instance, by developing smart cities.

In addition, the URL website links served as the Internet's core. Every Internet connection needs a URL link to route from one place to another. Tons of URLs and website links exist in the current digital world. The people with the Internet will indeed be in contact with websites or URL links. Therefore, numerous parts could be taken into observation and analysis from the datasets primarily stored in either cloud storage, engine storage, or databank. Analysing the big data related to URL websites will help get more data about browsing. Information such as popular surfing websites, studies of behaviour when browsing the Internet, and others can be explored using big data analysis. Big data analysis can be applied in any field and serves as a new way of retrieving data for the diversity of big data. This paper focuses on the personal web history of datasets in big data analysis. The history website datasets included characteristics such as visit time, click count, last visit, and, most importantly, the URL link. Lastly, a dedicated multilevel N-gram feature extraction and an optimised Naïve Bayes classification of URL links are proposed and implemented in the big data analysis of the historical website datasets.

The subsequent sections are structured as follows. Section 2 reviews the relevant literature and identifies the main research gaps. Section 3 introduces the proposed methodology, covering data preprocessing, model development, and analysis procedures. Section 4 presents and discusses the results, and Section 5 concludes the study and outlines potential avenues for future work.

2. Background and Motivation

Big data analytics involves analysing large data sets to extract valuable insights that can be presented visually to aid decision-making. All big data analysis tools follow this process, which includes categorising, summarising, matching, and performing advanced functions and algorithms to create graphical visualisations of the data.

Big data analysis has become increasingly popular in various fields, such as healthcare, business, cybersecurity, and entertainment. With the widespread use of smartphones, a large volume of data is being generated and flowing into the network. Wearable devices, for instance, are used in the healthcare industry to study the relationship between respiration rate and emotions, as outlined in [2]. In business, social media platforms have become popular for people to share their lifestyle choices. Companies can analyse this data by integrating buying and selling platforms into social media to generate more revenue. Using big data analysis in economic data provides better management in workflow systems, decision-making, and overall management concepts, as emphasised in [3]. By predicting potential issues, big data analysis can help avoid significant losses. However, there is limited research using big data analysis for historical website datasets. Some related work on browsing datasets has been explored, such as analysing user browsing behaviour by tracking mouse clicks to create massive time log sessions [4]. The paper [5] outlines various analysis tasks, including providing recommendation advertisements to users based on features such as previous searches, page views, and rankings. The authors in [6] proposed TiMR, which hybridises with the M-R framework of time-oriented data processing systems to present advertisements through real-time streams naturally.

To further study big data analysis of historical websites, classification served as the core of the historical website datasets. This is because the history website datasets were a list of URLs without any relevant grouping. So, it was vital to have a classification model to measure and categorise them. A paper studied the Content-Based Hierarchical URL Classification with Convolutional Neural Networks (CNN) [7]. The result was generally positive, at 85.9%. In addition, another paper implemented classification by Support Vector Machines (SVM) and Maximum Entropy Classifiers along with

character n-gram-based features [8], [9]. Besides, a paper proposed latent semantic analysis in the classification of the URLs to extract the features within them [10]. The result showed an accuracy of 82% based on standard features such as the website's content, the title of the website, and metadata. In another paper, the study implemented the n-gram keyword-based language model for the feature extraction, followed by the Naïve Bayes classifier on the DMOZ open-source datasets of 15 categories, with an accuracy of 82.72%. Furthermore, this paper selected and used the N-gram language model with the SVM classifier. This paper also used another method of classification called the N-gram language model. A comparison was made, and it showed that both of them have a similar result in terms of F1-score measurement [11].

Besides, a study evaluating the classification eligibility of blood donors using Decision Tree and Naïve Bayes Classifier revealed that the Naïve Bayes Classifier significantly outperformed the Decision Tree in terms of accuracy. According to the experimental phase, the Decision Tree was vulnerable to overlap when the number of classes and criteria was high [12]. Data preprocessing plays an important role in the accuracy of the study, and the results indicate that integrating Naïve Bayes classification with appropriate data preprocessing can improve prediction accuracy [13]. Additionally, another study demonstrated that the hardware implementation of the Naïve-Bayes classifier on a very low-cost platform yields comparable and competitive results [14].

The literature review indicates that the study of personal website history datasets is often overlooked by many parties. It also found that the big data collected from different IoT traffic sources requires classification to better handle the hidden information. Based on the reviewed paper, the existing classification methods for URLs have limitations in their performance. For example, content-based hierarchical URL classification with convolutional neural networks (CNN) is only capable of identifying whether the features of the website content are legitimate or not [7]. Furthermore, the F1-score of performance measurement did not exceed 86%, which is still considered insufficient accuracy. To address these issues, this paper proposes a new Multilevel N-gram model with a Naïve Bayes classifier, specifically targeting personal website history datasets. Table 1 presents a comparison of URL classification methods along with their performance metrics.

Table 1. URL classification methods and their performance

Classification Method	Key Features	F1-Score
Content-based Hierarchical CNN [7]	Uses CNN on URL content hierarchy	85.9%
N-gram + SVM + Max Entropy [15]	Uses character n-gram-based features	78.0%
N-gram + Naïve Bayes [16]	N-gram keyword model for feature extraction	82.0%
N-gram + SVM [11]	N-gram language model with SVM classifier	82.7%
Multilevel N-gram + Naïve Bayes classifier [Proposed]	Included removal of duplication, stop-words, and punctuation before applying the classifier	88%

3. Methodology

A multilevel N-gram model integrated with a Naïve Bayes classifier has been developed for the classification of URL datasets. Initially, datasets sourced from the DMOZ open-source provider are loaded into the model to categorise and group URLs according to their respective classifications. Then, the outcomes will be subjected to big data analysis to extract deeper insights from the personal website history datasets. To validate the outcomes, metrics such as precision, recall, and F1-score are computed for comparison with related studies. The overall study framework is illustrated in Figure 1.

The simulations are conducted using an Intel® Core™ i7 Processor, RTX3060 GPU, and 16 GB RAM, with Python 3. The classification system is implemented within the Jupyter Notebook environment of Python for machine learning and big data analysis.

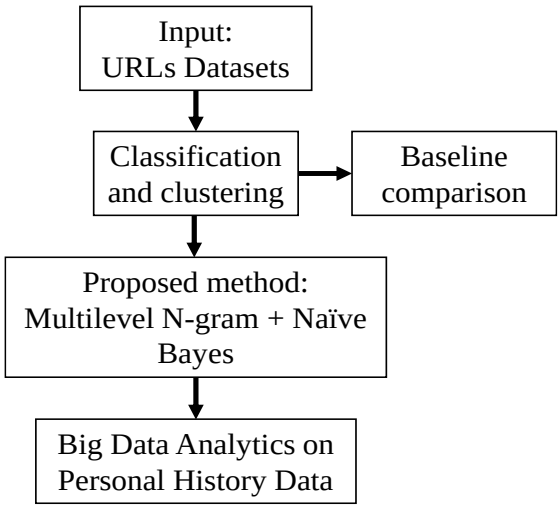


Figure 1. Overall study framework

3.1. Classification URLs

The proposed method uses a keyword-based Multilevel N-gram model with a Naïve Bayes classifier to classify the URLs. The term "multilevel" here refers to three distinct stages: Prepossessing, URL features extraction, and Classifier, as illustrated in Figure 2.

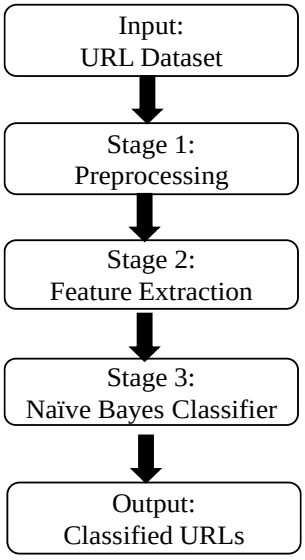


Figure 2. Stages of Classification URLs

The primary focus of this method is on the Naïve Bayes classification, which derives from the Bayes Theorem and is particularly suited for text classification. According to Equation (1), P(c|x) is the posterior probability of class (c, the target) given the predictor (x, the attributes), while P(x|c) denotes the likelihood, which is the probability of the predictor given the class and other factors, each probability corresponding to its respective context. In order to increase the accuracy, an N-gram model with values of N set to 1, 2, and 3 is used, as described in Equation (2). This approach facilitates the extraction of

features in a more effective manner, as detailed in section 3.1.2. These extracted N-gram features are subsequently used to estimate the probabilities of the Naïve Bayes classifier.

However, due to the discrete nature of N-gram features, it is possible that certain N-grams appearing in the testing phase are not observed during training. To address this issue, smoothing is incorporated to mitigate zero-probability estimates in the Naïve Bayes classifier. This ensures that unseen N-gram features continue to contribute to the probability computation, resulting in more stable, robust classification performance.

$$\mathbf{P(c|x)} = \frac{\mathbf{P(x|c)P(c)}}{\mathbf{P(x)}} \quad (1)$$

$$\mathbf{Ngrams_K} = \mathbf{X} - (\mathbf{N} - \mathbf{1}) \quad (2)$$

3.1.1. Prepossessing

The URL link can be found in various places, as long as there is an internet connection. Typically, the format of a URL link is " http://www.????.com….". To reduce the wording in the URL links, Python is used here to eliminate the stop words and punctuation. This action mainly simplifies the URL links, thereby enhancing the efficiency of simulations and calculations. Additionally, the duplicate URL links within the datasets will be filtered out to prevent any adverse effect on the outcome. When the duplicated URLs are removed, fewer URLs are used in training and testing, which minimises their impact on the classification process.

These preprocessing steps play a crucial role in improving classification performance by reducing noise and redundancy in the dataset. Removing irrelevant wording and duplicate URL links produces more meaningful N-gram features and reduces the dimensionality of the feature space. Without such preprocessing, noisy and repetitive patterns may degrade the classifier's performance, leading to lower precision and F1 Scores.

3.1.2. URL features extraction

All URLs may exhibit similar wording or characteristics; however, it is essential to identify the differences that distinguish them. Recognising these features is crucial for effective classification. The N-gram model is used here to extract relevant features from the URLs. The N-gram model can be segmented into various components, including unigrams, bigrams, trigrams, and beyond. For example, the sentence " I like to drink orange juice" can be analysed as follows: using unigrams (N=1), it is represented as I, like, drink, orange, juice. Each word appears individually. In the case of bigram (N=2), it is represented as Ilike, likedrink, drinkorange, orangejuice, where consecutive words are combined. The N-gram model used N values of 1,2, and 3, comprising almost the entire wording of the UR. In this way, the URL wording that sometimes only has the correct meaning, combined with other words such as "ice-cream", can be solved.

3.1.3. Naïve Bayes Classifier

This paper aims to improve the existing method of URL classification to identify trends within them. The classification algorithm used is the Naïve Bayes classifier. Following feature extraction that emphasises the wording, the extracted terms will be classified into appropriate categories. The historical websites are classified into 15 distinct fields: Adults, Art, Business, Computer, Game, Health, Home, Kid, News, Recreation, Reference, Science, Shopping, Society and Sports. Each URL will be grouped into the corresponding field category. In the Naïve Bayes classifier, each attribute is treated as independent of the given classes, making it particularly effective for text classification of URLs with

unconventional wording. Once the URLs are classified into their own category, the clustering process will aggregate all URLs that share the same classification.

3.2. Big data analysis approach

The results of the classification process are normalised before being input to the big data analysis approach. A graphical user interface (GUI), as shown in Figure 3, has been developed using Python, featuring four buttons: ConvertGUI, ReadGUI, CleaningGUI, and FilteringGUI, to facilitate the normalisation process. The ConvertGUI button first converts the datasets file running through the Multilevel N-gram model with Naïve Bayes classifier into a CSV or Excel format file. Then, the CSV version of the dataset can be accessed by pressing the ReadGUI button. Finally, the dataset undergoes a data cleaning process initiated by the CleaningGUI button. This process mainly identifies the errors that may arise during the classification or conversion processes.

The FilteringGUI button is linked to ReadGUI. When the FilteringGUI button is activated, datasets from the database, whether from the cloud or a local databank, are converted into a format that Python can easily access. At this stage, the datasets are categorised according to the specified group and saved in a designated location on the computer. This process helps to organise the datasets effectively, setting the stage for comprehensive analysis. The filtered data will then be imported into Jupyter Notebook tools for in-depth big data analysis. The actual valuable information and hidden data will be explored. The information will be visually represented, including trends in users' browsing habits. Additionally, uncovering hidden data, such as the trends in website visits, may point towards a potential career path. For example, a significant number of computer-related websites browsed, such as coding tutorials and information on the latest graphics cards, etc., indicates a strong interest in the field, suggesting that the user might consider it as a viable career option.

Figure 3. Running of GUI

In addition to using big data analysis tools, numerous graph-based libraries are available in Python for visualising the generated data. A more effective presentation of data can be achieved through the integration of the Jupyter Notebook and Matplotlib library. This combination allows for the valuable insights derived from the data to be displayed graphically using the graph library. The resulting visualisations showcase key information related to the imported dataset, including mean values, as well as maximum and minimum figures. These visual representations reflect the outcomes of the big data analysis conducted on the URL history datasets.

4. Result and Discussion

The study was conducted in two stages: a comparative validation experiment using the DMOZ dataset to identify the effectiveness of the proposed methods, and a validation of the model by deploying it in an applied analysis of a real-world browsing history dataset. Table 2 shows the performance of the proposed keyword-based multilevel n-gram model with the Naïve Bayes classifier. The proposed model demonstrated consistently high performance, with average precision, recall, and F1-scores of 90%, 88%, and 88%, respectively. The categories, such as Computers, Health, Home, Recreation, and Sports, scored above 95%, reflecting stable classification in well-structured domains. However, the categories of Adults and Kids performed poorly in terms of recall and F1-score, likely due to irregular URL patterns and high content variability. Similarly, Business and Games showed lower precision despite high recall, indicating a tendency toward false positives.

This performance disparity across categories can be further explained by the lexical characteristics of URLs and dataset distribution. Categories such as “Shopping” and “Computer” typically contain structured and descriptive keywords (e.g., product names, brands, or technical terms), which are well captured by N-gram features, leading to high classification performance. In contrast, categories such as “Adults” often include ambiguous, abbreviated, or obfuscated URL patterns with limited meaningful tokens, reducing the effectiveness of N-gram feature extraction.

Table 2. The performance of the proposed model

Category	Precision	Recall	F1-score
Adults	99%	14%	24%
Arts	99%	93%	77%
Business	72%	98%	83%
Computer	96%	98%	97%
Game	71%	98%	82%
Health	99%	98%	98%
Home	99%	97%	98%
Kids	92%	50%	65%
News	99%	87%	93%
Recreation	96%	99%	97%
Reference	91%	97%	94%
Science	95%	97%	96%
Shopping	98%	97%	99%
Society	87%	99%	93%
Sport	98%	97%	98%
Average	90%	88%	88%

The confusion matrix in Figure 4 shows strong diagonal dominance, with most categories exceeding 0.90 accuracy. This shows the model’s reliability for well-structured domains such as Computers, Health, Home, and Sports. Misclassifications mostly happen in the Adults and Kids categories, where samples are frequently assigned to News, Games, or Recreation, reflecting irregular URL patterns and overlapping vocabulary. The Business and Games classes also suffer from higher false positives due to shared terms with broader categories.

Table 3 compares the Fl-score between the proposed method and prior research. Most of the categories in the proposed method exceeded 80% except for Adults, Arts, and Kids. Only 24% of the proposed method in the Adults category was far behind the SVM+all-gram approach; however, the ME+n-gram method does not run in the adults’ category. SVM classifier served effectively in processing long sentences, which implied that the Adults website URLs needed a way to include the whole

combination of wording. The lack of keywords and difficulty identifying the wording for this kind of website causes the proposed method to score low values for certain categories. From the average F1-score, the proposed method achieved a 4.5% to 8.7% improvement in accuracy compared to prior methods. Thus, this paper's proposed method has successfully enhanced the classification of website URLs.

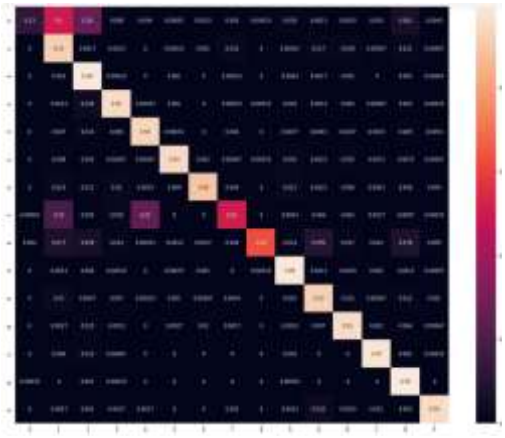


Figure 4. Confusion matrix of the proposed method

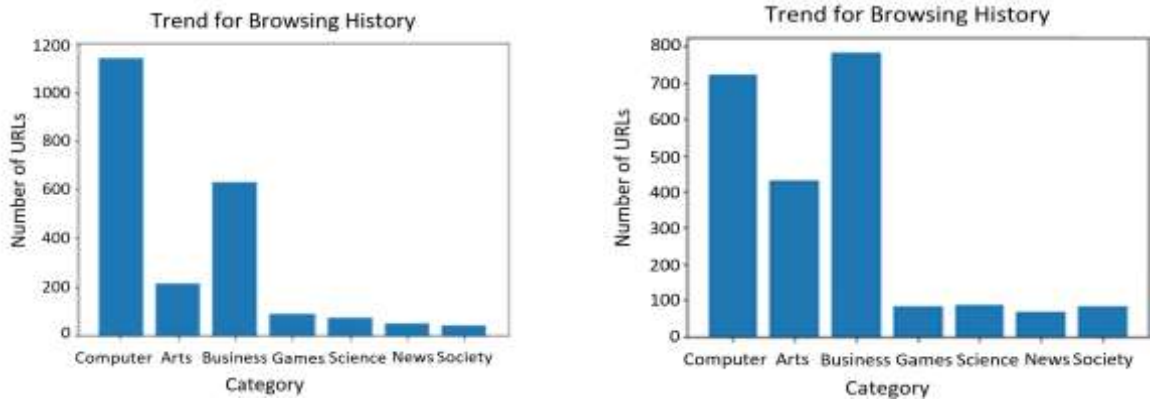
Table 3. Comparison of F1-score results with previous research

Category	SVM+all gram (%) [17]	ME+N-gram (%) [15]	Multilevel N-gram+NB Proposed (%)
Adults	87.60	null	24%
Arts	81.90	76.09	77%
Business	82.90	78.76	83%
Computer	82.50	77.82	97%
Game	86.70	83.09	82%
Health	82.40	89.98	98%
Home	81.00	78.97	98%
Kids	80.00	null	65%
News	80.00	76.30	93%
Recreation	79.70	75.08	97%
Reference	84.40	81.51	94%
Science	80.10	77.10	96%
Shopping	83.10	79.59	99%
Society	80.20	75.86	93%
Sport	84.00	82.39	98%
Average	82.44	78.23	88%

The trend of the browsing history for a personal website using big data analysis is determined here. The datasets of around 2000 obtained from Kaggle are employed. Based on Figure 5, big data analysis results of historical website datasets, most URL websites browsed are related to computers and business. The computer category website URLs rank first with 1144, continuing with 633 for the business category and 214 for arts. The remaining four categories have a gap difference of less than 20, where Games (92), Science (74), News (54), and Society (42). This showed that the remaining four categories are not this browsing history dataset's main interest or focus.

Compared with Figures 5 (a) and (b), the results of big data analysis with and without the proposed method have a similar trend. Computer, Arts, and business categories have far more in quantity than the remaining four categories, which are 723, 431, and 783, respectively. The Business category has 40% more URLs than the Computer category URLs, compared to those without and with the proposed big data analysis method. Besides, there is a significant drop in the Computer category URLs, around 40% compared to the proposed method. Meanwhile, the value for the remaining four categories also changes to 83, 85, 67, and 81. The gaming category URLs have decreased in big data analysis without the proposed method. Without the proposed method for performing big data analysis on historical website datasets, it will be hard to find the hidden information in the personal history website datasets.

The distribution of clicks in Figure 6 shows a highly skewed browsing pattern, where Business and Computer categories alone contribute nearly 85% of total interactions. This dominance suggests a strong preference for professional and technology-related content, reflecting the user's primary focus areas. In contrast, categories such as News, Society, and Science register minimal activity, pointing to limited engagement with general information or academic sources. The imbalance highlights how web usage is shaped by selective interests, but it also exposes blind spots in information consumption that could affect knowledge diversity and situational awareness.



(a) (b)

Figure 5. Big data analysis results of historical website datasets (a) with and (b) without the proposed method

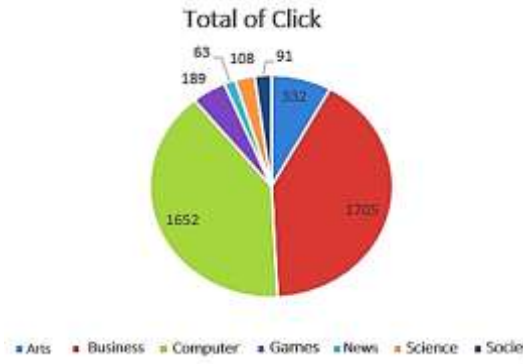


Figure 6. Total number of clicks in each category of historical website datasets

5. Conclusion

In conclusion, the multilevel N-gram model with the Naïve Bayes classifier successfully enhanced the big data analysis approach in classifying and clustering the URL website link. From the result, the proposed method had achieved an 88% F1 score compared to previous studies. It had improved around 5% compared to one of the similar studies. Based on Table 1, the categories affecting the F1 score were adults and kids. If both categories were, the F1-score would have boosted the accuracy to around 94%, which was why the multilevel N-gram model with the Naïve Bayes classifier method was considered a significant improvement in terms of performance. The trend of the browsing history datasets was explored and presented here. It showed that the individual browsing history datasets were highly interested in computer-related things. With the aid of this big data analysis method, much of the information hidden within can be seen more clearly and retrieved. A recommendation for future work will be adding another content-based filtering throughout the classification method.

Acknowledgements

The author would like to thank the Faculty of Engineering, Universiti Malaysia Sarawak (UNIMAS), for providing opportunities and facilities to support this project.

Conflict of Interest

We declare no conflict regarding the publication of the study.

References

[1] Ida Afriliana and Nurohim. (2021). Classification of Teachers and Lecturers Engagement on Webinar during the Pandemic using the Utilization of Big Data, *International Journal of Science, Technology & Management*, vol. 2, no. 3, pp. 673–684. doi: 10.46729/ijstm.v2i3.224.

[2] D. Ayata, Y. Yaslan, and M. E. Kamasak. (2020). Emotion Recognition from Multimodal Physiological Signals for Emotion Aware Healthcare Systems. *J Med Biol Eng*, vol. 40, no. 2, pp. 149–157. doi: 10.1007/s40846-019-00505-7.

[3] S. Ren. (2022). Optimization of Enterprise Financial Management and Decision-Making Systems Based on Big Data. *Journal of Mathematics*, vol. 2022, no. 1. doi: 10.1155/2022/1708506.

[4] H. Scells, Jimmy, and G. Zuccon. (2021, Jul). Big Brother: A Drop-In Website Interaction Logging Service. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: ACM, pp. 2590–2594. doi: 10.1145/3404835.3462781.

[5] Md. S. Rahman and H. Reza. (2022). A Systematic Review Towards Big Data Analytics in Social Media, *Big Data Mining and Analytics*, vol. 5, no. 3, pp. 228–244, doi: 10.26599/BDMA.2022.9020009.

[6] B. Chandramouli, J. Goldstein, and S. Duan. (2012, Apr). Temporal Analytics on Big Data for Web Advertising. In *2012 IEEE 28th International Conference on Data Engineering*, IEEE, pp. 90–101. doi: 10.1109/ICDE.2012.55.

[7] K. Maladkar. (2019, Dec). Content-Based Hierarchical URL Classification with Convolutional Neural Networks. In *2019 International Conference on Information Technology (ICIT)*, IEEE, pp. 263–266. doi: 10.1109/ICIT48102.2019.00053.

[8] H. Gomez, I. Markov, J. Baptista, G. Sidorov, and D. Pinto. (2017). Discriminating between Similar Languages Using a Combination of Typed and Untyped Character N-grams and Words. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 137–145. doi: 10.18653/v1/W17-1217.

[9] U. Mahor and A. Kumar. (2023). Authorship Attribution using Tf-Idf weight with Machine Learning Approaches. doi: 10.21203/rs.3.rs-2707585/v1.

[10] F. Ullah, X. Cheng, L. Mostarda, and S. Jabbar. (2023). Android-IoT Malware Classification and Detection Approach Using Deep URL Features Analysis. *Journal of Database Management*, vol. 34, no. 2, pp. 1–26. doi: 10.4018/JDM.318414.

[11] T. A. Abdallah and B. de La Iglesia. (2015). URL-Based Web Page Classification: With n-Gram Language Models. pp. 19–33. doi: 10.1007/978-3-319-25840-9_2.

[12] H. F. Mustika, A. F. Syafiandini, L. P. Manik, and Y. Rianto, "Evaluating Naïve Bayes Automated Classification for GBAORD," *Computer Engineering and Applications Journal*, vol. 9, no. 1, pp. 29–37, Feb. 2020, doi: 10.18495/comengapp.v9i1.320.

[13] D. Fahrudy and S. 'Uyun. (2022). Classification of Student Graduation using Navie Bayes by Comparing between Random Oversampling and Feature Selections of Information Gain and Forward Selection. *JOIV : International Journal on Informatics Visualization*, vol. 6, no. 4, p. 798. doi: 10.30630/joiv.6.4.982.

[14] Z. Xue, J. Wei, and W. Guo. (2020). A Real-Time Naïve Bayes Classifier Accelerator on FPGA. *IEEE Access*, vol. 8, pp. 40755–40766. doi: 10.1109/ACCESS.2020.2976879.

[15] R. Rajalakshmi and C. Aravindan. (2013, Dec). Web page classification using n-gram based URL features. In *2013 Fifth International Conference on Advanced Computing (ICoAC)*, IEEE, pp. 15–21. doi: 10.1109/ICoAC.2013.6921920.

[16] D. Shen, Z. Chen, Q. Yang, H.J. Zeng, B. Zhang, Y. Lu, and W. Y.Wa. (2004, Jul). Web-page classification through summarization. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, New York, NY, USA: ACM, pp. 242–249. doi: 10.1145/1008992.1009035.

[17] E. Baykan, M. Henzinger, L. Marian, and I. Weber. (2009, Apr). Purely URL-based topic classification. In *Proceedings of the 18th international conference on World wide web*, New York, NY, USA: ACM, pp. 1109–1110. doi: 10.1145/1526709.1526880.