COMPILER
Vol. 11, No. 2, November 2022, pp. 51~58
ISSN: 2252-2403 (Print), ISSN: 2549-2403 (Online)
Accredited Fourth Grade by Kemenristekdikti, Decree No:200/M/KPT/2020
DOI: 10.28989/compiler.v11i2.1327

Analysis of Deep Learning Approach Based on Convolution Neural Network (CNN) for Classification of Web Page Title and Description Text

Aris Wahyu Murdiyanto[1,*], Muhammad Habibi[2]
[1]Department of Information System, Universitas Jenderal Achmad Yani Yogyakarta, Indonesia
[2]Department of Informatics, Universitas Jenderal Achmad Yani Yogyakarta, Indonesia

Article Info

Article history:

Received August 20, 2022
Accepted November 25, 2022
Published December 31, 2022

Keywords:

Deep learning
Convolution Neural Networks
Web page classification
Text classification
DMOZ dataset

ABSTRACT

The volume of digital documents available online is growing exponentially due to the increasing use of the internet. Categorization of information obtained online is needed to make it easier for recipients of information to determine and filter which information is needed. Classification of web pages can be based on titles and descriptions, which are text data that can be done by utilizing deep learning technology for text classification. This study aimed to conduct data training and analysis experiments to determine the accuracy of the proposed deep learning architecture in classifying web page titles and descriptions. In this research, we proposed a Convolution Neural Network (CNN) architecture that generates few parameters. The training and evaluation set was conducted on the web page dataset provided by DMOZ. As a result, the proposed CNN architecture with the number of N (Dropout + 1D Convolution + ReLU activation) equal to 1 achieves the best validation accuracy. It achieves 79.51% with only generates 825,061 parameters. The proposed CNN architecture achieved outperformed performance on the accuracy of the five other technologies in the state-of-the-art.

Corresponding Author:

Aris Wahyu Murdiyanhto,
Department of Information System,
Universitas Jenderal Achmad Yani Yogyakarta,
Jl. Siliwangi, Ringroad Barat, Banyuraden, Gamping, Sleman, Daerah Istimewa Yogyakarta, Indonesia.
Email: *ariswahyumurdiyanto@gmail.com

1. INTRODUCTION

In today's increasingly digital era, the internet has become an inseparable part of people's lives. People can easily obtain useful information through the internet in a relatively short time [1]. It is also reinforced by the community's demands to get the latest information as a basis for decision-making. News web pages and blogs are one of the most common sources of information used to gather information about current events. However, collecting and categorizing information from various web pages seems to be a challenging task. It is further compounded by the volume of online digital documents growing massively due to increased internet use [2].

The emergence of technology such as web scraping allows the acquisition and collection of information from various web pages to be made automatically in a short time [3]–[7]. In addition to performing it, categorizing information obtained is also needed to facilitate the recipients of information in determining and filtering which information is required. Information obtained from web pages or blogs, of course, comes in several categories where users are only interested in specific topics, for example, the categories of business, entertainment, sports, or politics. Web page classification (WPC) technology is automatically a solution to answer these problems. Web page classification assigns web pages to one or more predefined category labels [8].

Several previous researchers have carried out the application of web page classification techniques with various approaches. Le et al., in 2015 [9], proposed a new simplified Simplified Swarm Optimization

(SSO) for classifying web pages by studying the best weights for each feature in the training data set and adopting the best weights. The Taguchi method was applied to determine parameter settings. Based on the experimental results, SSO performs better than the other three approaches: Genetic Algorithm (GA), Bayesian classifier, and K-nearest neighbor (KNN) classifiers. Purnama in 2020 [10] applied and compared two methods, namely the Multilayer Perceptron and Naive Bayes methods, to classify websites automatically. His research shows that the Naive Bayes method has an accuracy rate of 89%, which is better than the Multilayer Perceptron method, which produces an accuracy of 80%.

Matošević et al. [11] utilized machine learning technology to classify web pages based on the knowledge of experts. It classified web pages into three predetermined classes. It is based on the level of content adjustment from Search Engine Optimization (SEO) recommendations. A classifier was constructed and trained to classify an unknown sample (web page) into one of three predefined classes and to identify the important factors influencing the level of page customization. The domain expert manually labels the data in the training set. According to the experimental results, machine learning technology can predict the degree to which a web page conforms to SEO recommendations. The resulting classifier accuracy ranges from 54.59% to 69.67%. Buber and Diri [12] applied deep learning technology based on recurrent neural networks (RNN) to classify web pages. In their research, classification is based on meta tag information available on web pages, such as title, description, and keywords. Based on the results of the study obtained an accuracy rate of 85%.

Apandi et al., in 2021 [13], also applied deep learning technology based on Convolutional Neural Network (CNN) to classify web pages and identify whether they related to Game or Online Video Streaming based on the text content of a web page. It used the dataset from the Center of Information Technology and Communication (PTMK), which is part of Universiti Malaysia Pahang (UMP). The dataset comprises 640 web pages in the Game category and 407 in the Online Video Streaming category. A manually designed CNN architecture was proposed in the study, which resulted in an accuracy of 82.22% for detecting previously classified web pages.

Classification of web pages can be based on titles and descriptions, which can be obtained automatically using web scraping technology. Classifying titles and descriptions, which are text data, can be conducted using deep learning technology for text classification [14]–[21]. This technology uses a dataset of titles and descriptions from various web pages to train classifiers. The dataset has been labeled according to its categories. After the training is complete, the trained classifier can be used to classify the title and text of a web page. However, deep learning technology based on CNN, which has many layers tends to generate large parameters. It makes the architecture not efficient.

Based on these problems, we proposed a new CNN architecture to classify a web page's title and description data, which generates a few parameters. This study aims to conduct data training and analysis experiments to determine the efficient deep learning architecture for classifying web page titles and descriptions. In this research, deep learning architecture will be designed based on CNN. In this study, we use the web page dataset provided by DMOZ as training and validation. The novelty and contribution of this study can summarize as follow:

1) In terms of approaches, we proposed a new CNN architecture for web page classification and conducted experiments and analysis of the proposed architecture. The proposed CNN architecture generates a few parameters that make the architecture more efficient.
2) In terms of implementing datasets, experiments and analysis in this study were carried out on URL Classification Dataset provided by DMOZ.

2. RESEARCH METHOD

The analysis of web page classification using deep learning in this research consists of four components. They are data pre-processing, word embedding, feature extraction, and classification. In general, the proposed method can be seen in Figure 1.

2.1. Dataset

In this study, we used web page dataset provided by DMOZ. The dataset selection was determined through several procedures as follows:

1) The dataset must have many instances. The web page dataset provided by DMOZ has 1,195,851 so that can be used as data training and validation.
2) The dataset must have some categories as classes. The web page dataset provided by DMOZ consist of 13 categories: Business, Society, Arts, Science, Computers, Sport, Recreation, Shopping, Health, Reference, Games, Home, and News.

2.2. Data Pre-Processing

In the early stages, data preprocessing was carried out to process the database that will be used for the feature extraction and classification process. At this stage, the web page title and description text was combined. Further, the tokenizing process was carried out, which served to cut or separate each sentence into several tokens/parts [22] so that the input sentence becomes a collection of words in a list.

2.3. Word Embedding Method

Word embedding is a method employed to map words from vocabulary to vectors of real numbers [23]. It represents words that encode semantic, statistic, or context information [24]. This method takes the corpus of text as input through the pre-processing stage and then produces a vector representation of each word in the word corpus as output.

2.4. Feature Extraction and Classification

Convolutional Neural Network (CNN) is one of the deep learning techniques that can be applied to extract features and classify text documents [25]–[27]. The proposed architecture for web page classification, which can be seen in Figure 1, implements a series of 1D convolution layers. The function of this layer is to extract sentence features and then predict the sentence categories at the end of the network. 1D convolution is used to extract spectral features. The 1D kernel is exploited to capture intrinsic semantic content along the 1D spectral dimension effectively. In this operation, a one-dimension input data is convoluted with a one-dimension filter (one-dimension kernel length is the size of the receptive field) and produces a one-dimension output. After that, the activation function forms the output data (feature vector).

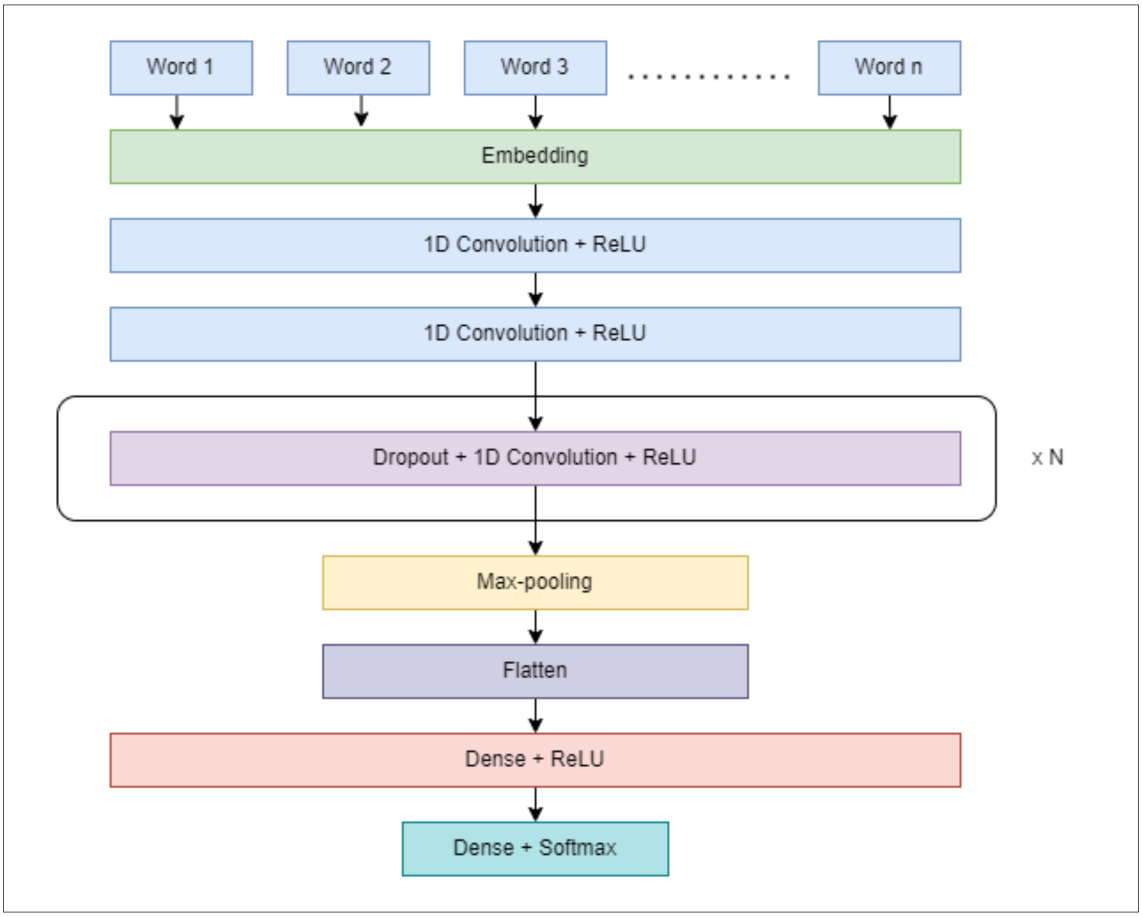

Figure 2. The proposed CNN architecture for web page classification

In the proposed architecture, the input is a vector from the results of the previous word embedding process, and then it will be entered for the embedding stage to reduce the vocab size. Then, several layers of convolution with the number of kernel 64, kernel size 5, and padding same are applied. In this study, the number of convolution layers will be analyzed to determine their effect on the accuracy of validation results. After each convolution layer, an activation function is used to decide which values to pass to the next layer.

In this study, the Rectified Linear Unit (ReLU) activation function was applied, passing positive values and make the layer decides the output. The ReLU activation function is described in (1).

$$f(x) = \begin{cases} x, & if\ x \geq 0 \\ 0, & if\ x < 0 \end{cases}. \tag{1}$$

x indicates the input function. In order to prevent overfitting [28], the dropout operation by 25% is also used before the third convolution layer and so on. After the last convolution layers, the pooling layer with size two is applied to down sampling each feature map generated by the convolution layer, and the maximum value is taken for each channel. In this study, we use the max-pooling operation, which is described in (2).

$$p(x) = \max(0, x). \tag{2}$$

x indicates the input function. The result of the pooling process from each channel is then set to a flatten operation which reduces the dimensions to a one-dimensional vector. The last step is the classification demanded to predict the category of the web page title and description text at the network's end. This step consists of two fully connected or dense layers followed by activation functions. The first dense layer contains 64 units, followed by ReLU activation. The second dense layer has 13 units representing the category class of the news category. At last, the softmax activation function is applied to transform the input value into possibilities representing the category class of the news category. The softmax function is described in (3).

$$s(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}. \tag{3}$$

x_i indicates the input function, n indicates the number of the class and e indicates the exponential operation. The proposed CNN architecture in this study is trained on NVIDIA GTX 1650 4GB as an accelerator. It is trained with a total of 20 epochs using Adam as the optimizer with a batch size of 256 to accelerate computing on high-performance GPU parallelism. It is implemented on the Tensorflow 2.0 and Keras 2.3.1 frameworks.

2.5. Evaluation

In this work, the performance of the CNN architecture was measured based on the validation accuracy. The validation accuracy was calculated by using an equation described in (4).

$$Validation\ Accuracy = \frac{\sum correct\ prediction}{\sum all\ of\ the\ data}. \tag{4}$$

Precision, recall, and F1-score was also used to measure the detail performance of the proposed CNN architecture based on the distribution of the dataset. The precision was calculated by using an equation described in (5), the recall is calculated by using an equation described in (6), and the F1-score is calculated by using an equation described in (7). TP indicates true positive, FP indicates false positive, and FN indicates false negative.

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN}. \tag{6}$$

$$F1 - Score = \frac{2 \times precision \times recall}{recision + recall}. \tag{7}$$

3. RESULTS AND ANALYSIS

This section describes the dataset analysis and the proposed CNN architecture performance analysis result. For the first time, we analyzed the dataset to obtain the data distribution. After that, we analyzed the performance of the CNN architecture based on the validation accuracy, precision, recall, and f1-score.

3.1. Dataset Analysis

In this work, the web page dataset provided by DMOZ was used as a training and validation set. It contains several attributes such as category, title, and description. It consists of more than one million data divided into 13 categories: Business, Society, Arts, Science, Computers, Sport, Recreation, Shopping, Health, Reference, Games, Home, and News. The example of the web page data of the DMOZ dataset is shown in Table 1.

Table 1. The Example of the Web Page Data of the DMOZ Dataset

Web Page Categories	Titles	Descriptions
Arts	The Art Institute of Chicago	Displays a small list of Rousseau's artworks.
Business	CPA Marketing Tips	Easy to use marketing system created especially for smaller practices.
Computers	RJ String Matching	New exact single string-matching algorithms
Games	Chess strategy	Winning chess tactics, puzzles, and instruction.
Health	Addiction Recovery	Information about alcohol and drug dependency, by David Hazen.
Home	Roommate Nation	Find a roommate to share housing with anywhere in the U.S.A.
News	OneWorld News	Daily news about human rights and environmental issues.
Recreation	Country Craft	Narrowboat holidays in Wales on the Monmouthshire and Brecon Canal.
Reference	These United States	Offers state information profiles for each of the 50 United States.
Science	Science Emu Farming India	Provides information about emu bird breeding and emu farming.
Shopping	Pink and Blue Bebe	Offering comforters, shoes, duvet covers, sleeping sacks, sheet covers and silkies.
Society	Society Diary of an Addict	One person's attempt to break free from his television addiction.
Sports	Irish Adventure Racing	The home site of the Irish Adventure Racing Squad.

In this experiment, the DMOZ dataset was divided into 80% and 20% as training and validation sets, respectively. The distribution of the web page dataset is shown in Figure 2. The detail of the web page dataset is shown in Table 2. The dataset consists of web page data with a significant imbalance between each category. It will affect the accuracy of the learning process, which will be discussed in the performance analysis section below.

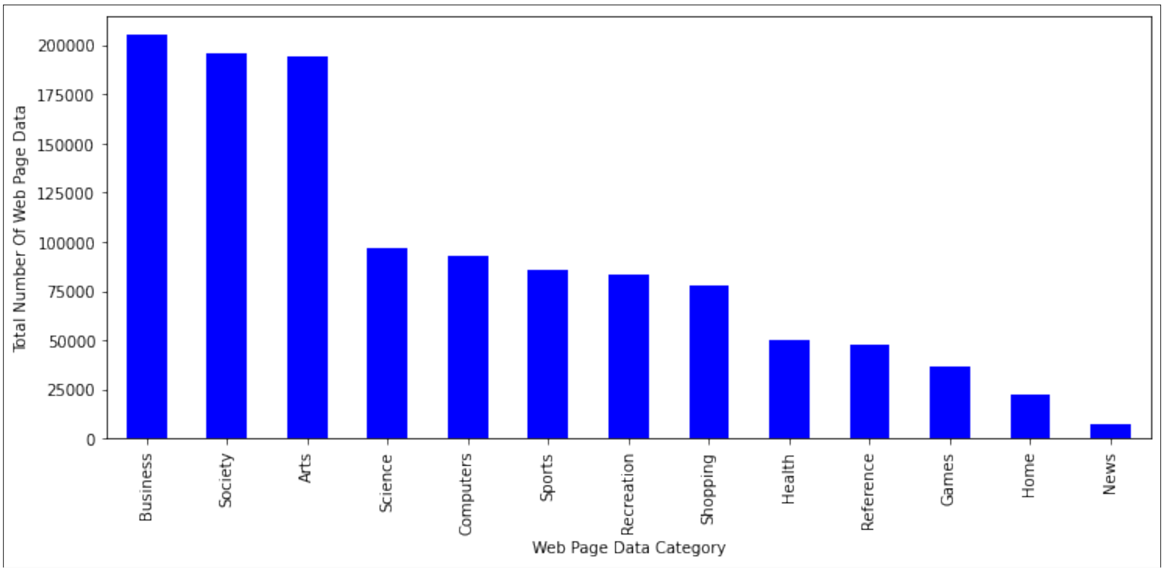

Figure 2. Distribution of the Web Page Title and Description Data of DMOZ Dataset

Table 2. Number of Web Page Title and Description Data from DMOZ Dataset based on Category

Web Page Categories	The Number of Data
Arts	193,914
Business	204,910
Computers	93,204
Games	36,454
Health	50,388
Home	22,241
News	7,488
Recreation	83,618
Reference	47,428
Science	96,766
Shopping	78,184
Society	195,880
Sports	85,376

3.1. CNN Performance Analysis

This section will discuss the performance of the proposed CNN architecture which is based on validation accuracy. As can be seen, the proposed CNN architecture shown in Figure 1 consists of several 1D convolution layers. This work investigated the effect of CNN architectural depth, which refers to the number of convolution layers used. This study analyzed five settings for the number of convolution layers, which can be seen in Table 3. It shows that the proposed CNN architecture with the number of N (Dropout + 1D Convolution + ReLU activation) equal to 1 achieves the best validation accuracy. It achieved 79.51% with only generates 825,061 parameters. It is above the performance of the proposed CNN architecture with the number of N equal to 2 with 845,605 parameters and N equal to 0 with 804,517 parameters, which differed by 0.11% and 0.17%, respectively. It proved that in this case, the deeper the CNN architecture shown from the more layers of convolution does not guarantee a better result the validation accuracy.

Table 3. Evaluation Result of Validation Accuracy on DMOZ dataset

No.	Number of N	Total Number of Convolution Layers	Number of Parameters	Validation Accuracy (%)
1	0	2	804,517	79.34
2	1	3	825,061	79.51
3	2	4	845,605	79.40
4	3	5	866,149	79.27
5	4	6	886,693	79.08

In order to measure the detail performance of the proposed CNN architecture based on the distribution of the dataset each web page category, we calculated the precision, recall, and f1-score. The analysis result of the precision, recall, and f1-score can be seen in Table 4. It shows that the some web page categories with the largest number of data tends to rank at the top. They are Business, Society, and Art with the F1-score 0.82, 0.80, and 0.83, respectively. It also shows that News category, the web page category with the smallest largest number of data, is ranked at the lowest with the F1-score 0.63. It indicates that the more data used will increase the knowledge learned by the architecture and vice versa, the less data used will reduce the knowledge learned by the architecture, which will affect the classification performance results.

Table 4. Evaluation Result of Precision, Recall, and F1-Score on DMOZ dataset

Web Page Category	Precision	Recall	F1-Score
Arts	0.85	0.81	0.83
Business	0.80	0.84	0.82
Computers	0.79	0.76	0.77
Games	0.78	0.77	0.78
Health	0.79	0.77	0.78
Home	0.69	0.71	0.70
News	0.61	0.65	0.63
Recreation	0.76	0.72	0.74
Reference	0.64	0.71	0.68
Science	0.65	0.77	0.71
Shopping	0.74	0.71	0.72
Society	0.82	0.77	0.80
Sports	0.86	0.91	0.89

Compared with studies that performed web page classification in previous works, the proposed CNN architecture achieved outperformed performance on the accuracy of the five other technologies in the state-of-the-art. Table 5 shows the accuracy results of the comparison with previous studies.

Table 5. Comparison of Previous Web Page Classification Studies on DMOZ dataset

No.	References	Accuracy (%)
1	k-nearest neighbors (KNN) [18]	57
2	Support Vector Machine (SVM) [18]	60
3	Artificial Neural Network (ANN) [18]	61
4	BERT Base + CNN [18]	65
5	BERT Base + DRIMN [18]	66
6	Proposed CNN	79

4. CONCLUSION

This study proposes a deep learning architecture to classify web page titles and descriptions. In this research, we proposed a Convolution Neural Network (CNN) architecture that generates few parameters. We also analyze the effect of the number of convolution layers on the proposed CNN architecture performance in classifying web page titles and descriptions. The training and evaluation set is conducted on the web page dataset provided by DMOZ. As a result, the proposed CNN architecture with the number of N (Dropout + 1D Convolution + ReLU activation) equal to 1 achieves the best validation accuracy. It achieves 79.51% with only generates 825,061 parameters. This study also shows that that in this case, the deeper the CNN architecture shown from the more layers of convolution does not guarantee a better result the validation accuracy. The proposed CNN architecture achieved outperformed performance on the accuracy of the five other technologies in the state-of-the-art. In the future work, the novel CNN architecture will be designed to increase the classifier accuracy.

ACKNOWLEDGEMENTS

The author would like to thank the Directorate of Research and Community Service Directorate General of Research and Development Strengthening (DRPM) of the Ministry of Research and Higher Education (Kemristekdikti) of the Republic of Indonesia for the support provided to the author in the form of research funding assistance in 2022 of Penelitian Dosen Pemula (PDP) scheme.

REFERENCES

[1] A. Priyanto and M. R. Ma'arif, "Implementasi Web Scrapping dan Text Mining untuk Akuisisi dan Kategorisasi Informasi dari Internet (Studi Kasus: Tutorial Hidroponik)," *Indonesian Journal of Information Systems*, vol. 1, no. 1, pp. 25–33, Aug. 2018, doi: 10.24002/ijis.v1i1.1664.

[2] J. Kristiyono and A. Nurrosyidah, "ANALISIS PERILAKU PENCARIAN INFORMASI DI INTERNET MELALUI FITUR VISUAL SEARCH," *Scriptura*, vol. 11, no. 2, pp. 96–104, Dec. 2021, doi: 10.9744/SCRIPTURA.11.2.96-104.

[3] M. I. Akrianto, A. D. Hartanto, and A. Priadana, "The Best Parameters to Select Instagram Account for Endorsement using Web Scraping," in *2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Nov. 2019, pp. 40–45. doi: 10.1109/ICITISEE48480.2019.9004038.

[4] A. Priadana and A. W. Murdiyanto, "Instagram Hashtag Trend Monitoring Using Web Scraping," *Journal Pekommas*, vol. 5, no. 1, p. 23, Apr. 2020, doi: 10.30818/jpkm.2020.2050103.

[5] A. W. Murdiyanto and A. Priadana, "Analysis of web scraping techniques to get keywords suggestion and allintitle automatically from Google Search Engines," *Compiler*, vol. 10, no. 2, pp. 71–78, Nov. 2021, doi: 10.28989/COMPILER.V10I2.1064.

[6] A. Himawan, A. Priadana, and A. Murdiyanto, "Implementation of Web Scraping to Build a Web-Based Instagram Account Data Downloader Application," *IJID (International Journal on Informatics for Development)*, vol. 9, no. 2, pp. 59–65, Dec. 2020, doi: 10.14421/IJID.2020.09201.

[7] A. I. Abdullah, A. Priadana, M. Muhajir, and S. N. Nur, "Data Mining for Determining The Best Cluster Of Student Instagram Account As New Student Admission Influencer," *Telematika : Jurnal Informatika dan Teknologi Informasi*, vol. 18, no. 2, pp. 255–266, Oct. 2021, doi: 10.31315/TELEMATIKA.V18I2.5067.

[8] L. Deri, M. Martinelli, D. Sartiano, and L. Sideri, "Large scale web-content classification," in *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, 2015.

[9] J. H. Lee, W. C. Yeh, and M. C. Chuang, "Web page classification based on a simplified swarm optimization," *Appl Math Comput*, vol. 270, pp. 13–24, Nov. 2015, doi: 10.1016/J.AMC.2015.07.120.

[10] I. N. Purnama, "Perbandingan Klasifikasi Website Secara Otomatis Menggunakan Metode Multilayer Perceptron dan Naive Bayes," *Jurnal Sistem Komputer dan Informatika (JSON)*, vol. 2, no. 2, pp. 155–161, Jan. 2021, doi: 10.30865/JSON.V2I2.2703.

[11] G. Matošević, J. Dobša, and D. Mladenić, "Using Machine Learning for Web Page Classification in Search Engine Optimization," *Future Internet 2021, Vol. 13, Page 9*, vol. 13, no. 1, p. 9, Jan. 2021, doi: 10.3390/FI13010009.

[12] E. Buber and B. Diri, "Web Page Classification Using RNN," *Procedia Comput Sci*, vol. 154, pp. 62–72, Jan. 2019, doi: 10.1016/J.PROCS.2019.06.011.

[13] S. H. Apandi, J. Sallim, R. Mohamed, and A. Madbouly, "Web Page Classification Using Convolutional Neural Network (CNN) towards Eliminating Internet Addiction," *Proceedings - 2021 International Conference on Software Engineering and Computer Systems and 4th International Conference on Computational Science and Information Management, ICSECS-ICOCSIM 2021*, pp. 149–154, Aug. 2021, doi: 10.1109/ICSECS52883.2021.00034.

[14] J. Gong *et al.*, "Hierarchical Graph Transformer-Based Deep Learning Model for Large-Scale Multi-Label Text Classification," *IEEE Access*, vol. 8, pp. 30885–30896, 2020, doi: 10.1109/ACCESS.2020.2972751.

[15] J. Wang, Y. Li, J. Shan, J. Bao, C. Zong, and L. Zhao, “Large-Scale Text Classification Using Scope-Based Convolutional Neural Network: A Deep Learning Approach,” *IEEE Access*, vol. 7, pp. 171548–171558, 2019, doi: 10.1109/ACCESS.2019.2955924.

[16] J. Cai, J. Li, W. Li, and J. Wang, “Deeplearning Model Used in Text Classification,” *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2018*, pp. 123–126, Jan. 2019, doi: 10.1109/ICCWAMTIP.2018.8632592.

[17] R. Wang, Z. Li, J. Cao, T. Chen, and L. Wang, “Convolutional Recurrent Neural Networks for Text Classification,” *Proceedings of the International Joint Conference on Neural Networks*, vol. 2019-July, Jul. 2019, doi: 10.1109/IJCNN.2019.8852406.

[18] A. Gupta and R. Bhatia, “Ensemble approach for web page classification,” *Multimedia Tools and Applications 2021 80:16*, vol. 80, no. 16, pp. 25219–25240, Apr. 2021, doi: 10.1007/S11042-021-10891-3.

[19] M. Hashemi, “Web page classification: a survey of perspectives, gaps, and future directions,” *Multimedia Tools and Applications 2020 79:17*, vol. 79, no. 17, pp. 11921–11945, Jan. 2020, doi: 10.1007/S11042-019-08373-8.

[20] S. Moriya and C. Shibata, “Transfer Learning Method for Very Deep CNN for Text Classification and Methods for its Evaluation,” *Proceedings - International Computer Software and Applications Conference*, vol. 2, pp. 153–158, Jun. 2018, doi: 10.1109/COMPSAC.2018.10220.

[21] C. Li, G. Zhan, and Z. Li, “News Text Classification Based on Improved Bi-LSTM-CNN,” *Proceedings - 9th International Conference on Information Technology in Medicine and Education, ITME 2018*, pp. 890–893, Dec. 2018, doi: 10.1109/ITME.2018.00199.

[22] A. Priadana and A. A. Rizal, “Sentiment Analysis on Government Performance in Tourism During The COVID-19 Pandemic Period With Lexicon Based,” *CAUCHY: Jurnal Matematika Murni dan Aplikasi*, vol. 7, no. 1, pp. 28–39, Nov. 2021, doi: 10.18860/CA.V7I1.12488.

[23] S. Selva Birunda and R. Kanniga Devi, “A review on word embedding techniques for text classification,” *Lecture Notes on Data Engineering and Communications Technologies*, vol. 59, pp. 267–281, 2021, doi: 10.1007/978-981-15-9651-3_23/COVER.

[24] M. Habib, M. Faris, A. Alomari, and H. Faris, “Altibbivec: A word embedding model for medical and health applications in the arabic language,” *IEEE Access*, vol. 9, pp. 133875–133888, 2021, doi: 10.1109/ACCESS.2021.3115617.

[25] P. Song, C. Geng, and Z. Li, “Research on Text Classification Based on Convolutional Neural Network,” *Proceedings - 2nd International Conference on Computer Network, Electronic and Automation, ICCNEA 2019*, pp. 229–232, Sep. 2019, doi: 10.1109/ICCNEA.2019.00052.

[26] K. H. Chan, S. K. Im, and W. Ke, “Variable-Depth Convolutional Neural Network for Text Classification,” *Communications in Computer and Information Science*, vol. 1333, pp. 685–692, 2020, doi: 10.1007/978-3-030-63823-8_78/COVER.

[27] S. Yang and Y. Tang, “Text Classification Based on Convolutional Neural Network and Attention Model,” *2020 3rd International Conference on Artificial Intelligence and Big Data, ICAIBD 2020*, pp. 67–73, May 2020, doi: 10.1109/ICAIBD49809.2020.9137447.

[28] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.