Indonesian Journal of Computer Science
ISSN 2549-7286 (*online*)
Jln. Khatib Sulaiman Dalam No. 1, Padang, Indonesia
Website: ijcs.stmikindonesia.ac.id | E-mail: ijcs@stmikindonesia.ac.id

Next Word Prediction for Book Title Search Using Bi-LSTM Algorithm

Alwizain Almas Trigreisian[1], Nisa Hanum Harani[1], Roni Andarsyah[1]
alwizainalmastrigreisian@gmail.com, nisa@ulbi.ac.id, roniandarsyah@ulbi.ac.id
Universitas Logistik dan Bisnis Internasional

Article Information

Submitted : 12 Jun 2023
Reviewed: 16 Jun 2023
Accepted : 27 Jun 2023

Keywords

Next Word Prediction, Deep Learning, Bi-LSTM, Django, Python

Abstract

Finding a suitable book title is still quite difficult at the moment. We often guess what book title we want, but in reality the book title is often not available. This research aims to overcome these problems by producing an accurate and efficient prediction model in predicting the next words in book title search using a deep learning algorithm, namely Bidirectional Long Short Term Memory (Bi-LSTM). The research stages consist of data collection, data preprocessing, data modeling, evaluation, and implementation. This research uses a dataset of Indonesian book titles obtained from the bukukita.com online bookstore website with 5618 data. The results show that the resulting deep learning model can predict the next words in the book title search with an accuracy of 81.82%. The model is implemented in the form of a web application using the Django framework, Python language, and MySQL database.

A. Introduction

The Book title search is one of the activities often carried out by readers and academics in finding information in a book. By searching for book titles, a person can find the right book for their needs and goals. However, searching for book titles is not always easy. Book seekers are more likely to guess what book title they want, but the desired book title is often not available. In addition, difficulties are also experienced when determining what keywords are appropriate in searching for the desired book title. Then the search results are also sometimes irrelevant to the keywords inputted.

To overcome these problems, a next word prediction system for book title search was developed that can help users find the desired book title more effectively and efficiently. Next word prediction is a process of guessing the next word that appears in a sentence or text. Prediction utilizes data from the past to estimate what happens in the future pragmatically and systematically in the hope of providing great objectivity [1]. The next word is predicted based on the previous word that has been inputted. Thus, the developed system will be able to provide recommendations or suggestions for suitable words to complement keywords or text in searching for book titles.

The developed system utilizes the Deep Learning method with the Bidirectional Long Short Term Memory (Bi-LSTM) algorithm. Deep Learning is a branch of machine learning based on artificial neural networks and has multiple processing layers to learn features in data [2][3]. Deep learning is capable of learning complex tasks with high accuracy results such as natural language processing, speech recognition, and image recognition [4].

Bidirectional Long Short Term Memory (Bi-LSTM) is one type of algorithm in Recurrent Neural Network (RNN). Bi-LSTM can learn the context of text data in two directions because the algorithm is composed of two LSTMs that can run in opposite directions (forward and backward) in parallel. Then the results of the two LSTMs are put together as a final output that can produce predictions of the next word more accurately according to the context in the text data [5][6][7].

The system implementation was built in the form of a web-based application using the Python programming language. The programming language was chosen because it is ideal for development in artificial intelligence, machine learning, and deep learning [8]. Python has advantages in ease of development for software, hardware, and web applications because it has good code readability [9]. Then related to the backend in making a book title search web using the Django framework and MySQL database.

Django is a web framework that uses the Python development language. Django has a fast, efficient, and practical development framework by offering security, scalability, and versatility that can help in creating web applications easily [10]. In addition, it can also help in representing an Object Relational Mapper (ORM) which when changes occur in the database does not need to adjust the query again [11]. MySQL is a database server program or database management system with SQL (Structured Query Language) commands that can send and receive data quickly, multithread, and multi-user [12][13].

In previous research, the prediction of the next word has been done with several RNN methods, such as research conducted by Afika Rianti et al. [14] who

used the LSTM algorithm in predicting the next word with Indonesian-language data about tourist destinations in Indonesia. The model created has an accuracy of 75% with 200 epochs. Further research was conducted by Radhika Sharma et al. [15] with the Bi-LSTM model for predicting the next word using the Hindi language obtained an accuracy of 79.54%. Research with English data has also been conducted by S. Rajakumar et al. [16] and S. Ramya et al. [17] using the Bi-LSTM algorithm as the next word prediction model by obtaining an accuracy of 81.07% and 72% respectively.

Research conducted by Karma Wangchuk et al. [18] used the Dzongkha syllable as its input data and the Bi-LSTM algorithm obtained an accuracy of 73.89%. Research using other syllables was also conducted by K. Chakradhar et al. [19] using Amharic syllable data and the Bi-LSTM model can get an accuracy of 76.1%. Other research with the Bi-LSTM algorithm for predicting the next word was conducted by Milind Soam et al. [20] with an accuracy of 66.1%.

By looking at the results of previous research, this research will develop a next word prediction system using the Bidirectional Long Short Term Memory (Bi-LSTM) algorithm. The method has been widely used to predict the next word, but in this study it is proposed to use Indonesian data in the form of book titles to obtain the best accuracy. So that the model and system can be applied to facilitate book seekers in searching for appropriate book titles.

B. Research Method

The research method used in this research consists of several stages, starting from data collection, data preprocessing, data modeling, evaluation, and implementation. The stages of the research method can be seen in Figure 1.

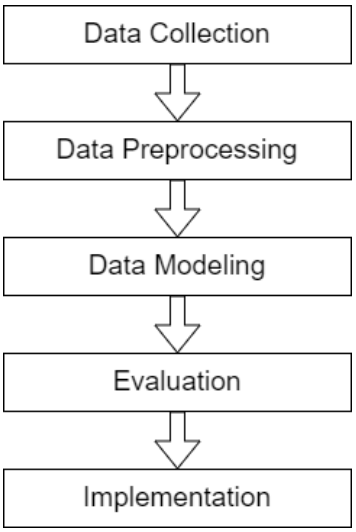

Figure 1. Stages of the Research Method [21]

The explanation of the stages of the research method used is as follows:

A. Data Collection

Data collection is done by retrieving book title data on the bukukita.com online bookstore website using scraping techniques. Scraping is done using the Go web

scraping framework. The data obtained in this study amounted to 5618 data and was stored in a dataset file in CSV format. Then the results of scraping data have also been stored in the MySQL database on PHPMyAdmin to be integrated with the application.

B. Data Preprocessing

The data preprocessing stage is done by cleaning and preparing the data before it is used in the modeling stage. The process carried out at this stage consists of case folding, removing punctuation, tokenizing, sequencing, and padding.

C. Data Modeling

Data modeling is done by implementing the Bidirectional Long Short Term Memory (Bi-LSTM) algorithm in Deep Learning. Bi-LSTM consists of two LSTM networks that have a function to process data sequences from the forward and backward directions [22]. LSTM itself will learn the data that must be removed and stored in each neuron [23]. The algorithm is a development of RNN by overcoming problems about vanishing gradients [24]. Based on the two LSTM networks owned by Bi-LSTM will produce one output as a result of combining the two LSTM networks owned and will obtain past and future information simultaneously [25]. The architecture of the Bi-LSTM algorithm can be seen in Figure 2.

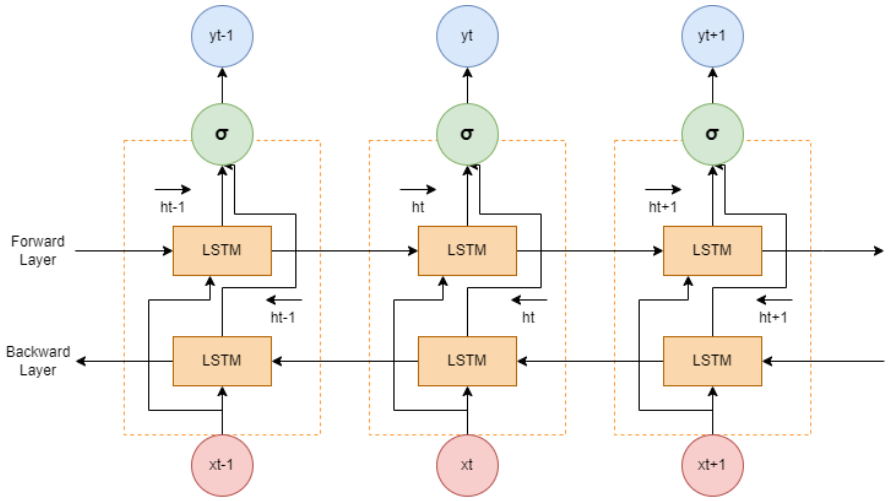

Figure 2. Bidirectional LSTM Architecture [26]

The calculation formula of the Bi-LSTM architecture is presented in Formula 1.

$$y_t = W_{\overrightarrow{hy}} \overrightarrow{ht} + W_{\overleftarrow{hy}} \overleftarrow{ht} \qquad (1)$$

Based on Formula 1, the notation y_t can be interpreted as the output gate value of Bi-LSTM. The notation $W_{\overrightarrow{hy}}$ is defined as the weight value of the output gate LSTM forward layer and the notation $W_{\overleftarrow{hy}}$ is defined as the weight value of the output gate LSTM backward layer. Then for the notation $\overrightarrow{ht}$ is interpreted as the output value of the LSTM forward layer and the notation $\overleftarrow{ht}$ is interpreted as the output value of the LSTM backward layer [27].

D. Evaluation

Evaluation is used to measure how well the model has been made. This evaluation is done to determine the performance of the model in predicting the next word. The measurement method used in this evaluation uses the accuracy matrix in the confusion matrix method. The Confusion matrix is a measurement of machine learning model performance by comparing actual values and predicted values [28]. The Confusion matrix can be seen in Figure 3.

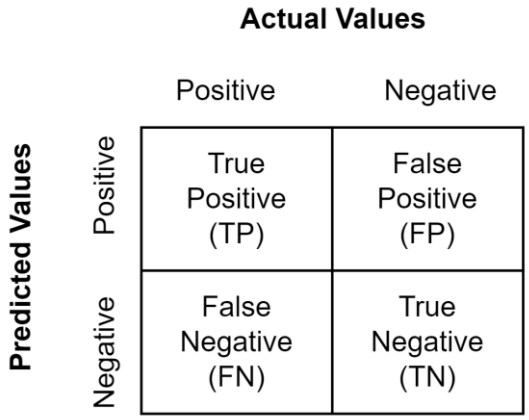

Figure 3. Confusion Matrix [29]

The accuracy formula of the confusion matrix can be seen in Formula 2.

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} x100\% \quad (2)$$

The Accuracy formula can also be defined as in Formula 3.

$$Accuracy = \frac{Total\ correct\ predictions}{Total\ predictions} x100\% \quad (3)$$

E. Implementation

Implementation is the final step in the development of this research. Implementation is intended so that the model that has been made can be implemented on the system and can be used by users to facilitate the search for book titles. The implementation is in the form of a book title search website built using the Python language and the Django framework. Then implement the MySQL database as a book data storage.

C. Result and Discussion

The data in the research conducted was taken from the bukukita.com online bookstore website using the scraping technique. The data has never been used for testing in previous studies. The data taken is only in the form of book titles available on the website which are stored in CSV format files and also stored in the MySQL database. Based on the scraping results, 10438 data were obtained and then data cleaning was carried out on the CSV dataset in the form of data duplication and only

selecting Indonesian language book title data. After the process, 5618 data on Indonesian book titles were obtained. Then the data that has been obtained is processed by converting it into data frame form to facilitate the data preprocessing stage.

Data that has been stored in the form of a data frame, then data preprocessing is carried out so that the data can be processed by the model algorithm. The data preprocessing process starts with data cleaning in the form of converting data into lowercase letters and removing all punctuation marks and symbols. The process is carried out so that the model to be created obtains good performance results. Then the cleaned data is tokenized to separate each word (token) in the sentence in the data. Words or tokens that have been separated from the sentence are sequenced to be given an n-gram sequence according to the words that appear together in the sentence. N-gram will calculate the frequency of occurrence of the sequence of words that appear in the dataset, then will estimate the probability [30]. Furthermore, words that have been given a sequence are padded so that each sentence or sequence has the same data length. The data resulting from the padding process is numerical so that computation can be done at the data modeling stage and will be used as data features.

The data modeling stage is carried out using the Bidirectional Long Short Term Memory (Bi-LSTM) algorithm. Data modeling with the Bi-LSTM algorithm has conducted several tests to find the best model parameters as in Table 1.

Table 1. Model Parameter Testing

No	Layer	Epoch	Loss	Accuracy
1	Bi-LSTM : 512	51	63.92%	80.29%
2	Bi-LSTM : 256	48	69.72%	79.80%
3	Bi-LSTM : 256	75	60.57%	80.68%
4	Bi-LSTM : 256	200	50.37%	81.38%
5	Bi-LSTM : 512	200	45.47%	81.72%
6	Bi-LSTM : 1000	200	45.36%	81.82%

Based on the test results in Table 1, the best model parameters are obtained by using a Bidirectional layer with 1000 inputs and an Epoch of 200. So that the layer structure in the model consists of an embedding layer which is useful for embedding word vectors with input 16 as the maximum length of the word sequence used as input and vector 10 as the dimension of the vector space in representing words. Then the Bidirectional layer with an input layer of 2000 is the result of the two-way readability of the algorithm based on 1000 inputs. Furthermore, the Dense layer with input in the form of a total of 6887 words that use softmax activation. The layer structure in the model that has been created can be seen in Figure 4.

```
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_1 (Embedding)     (None, 16, 10)            68870

 bidirectional_1 (Bidirectio  (None, 2000)             8088000
 nal)

 dense_1 (Dense)             (None, 6887)              13780887

=================================================================
Total params: 21,937,757
Trainable params: 21,937,757
Non-trainable params: 0
_________________________________________________________________
```

Figure 4. Model Layer Structure

Training on the model that has been made applies loss in the form of categorical crossentropy with adam optimizer and accuracy metrics. The results of the model training obtained a loss value of 0.4536 or 45.36% and an accuracy value of 0.8182 or 81.82%. The graph of the accuracy results of the model training process can be seen in Figure 5.

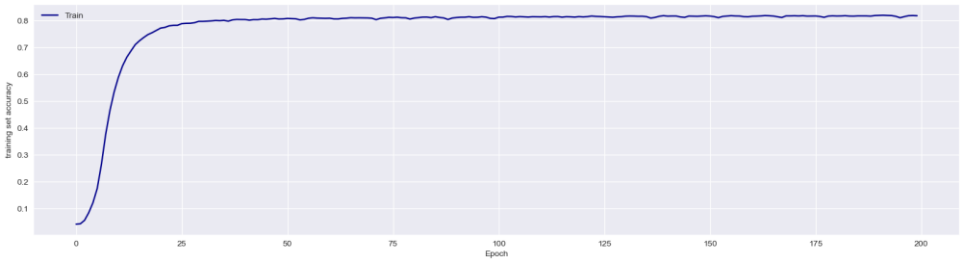

Figure 5. Training Accuracy Result Chart

Then the graph of the loss results from the model training process can be seen in Figure 6.

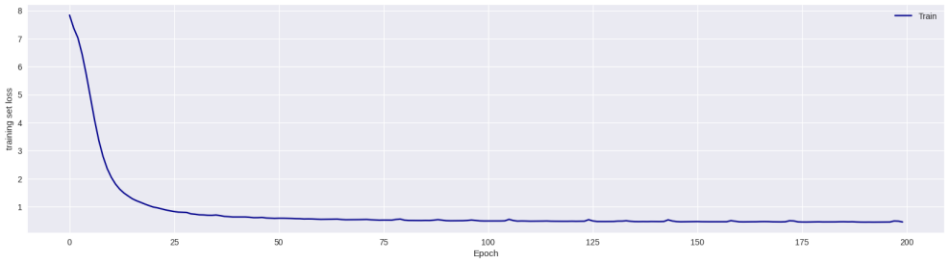

Figure 6. Training Loss Result Chart

After knowing the results of the training evaluation, testing is then carried out by predicting the inputted words. Testing is done by inputting a word and predicting the next few words. The test results can be seen in Figure 7.

```
Enter your line: nasihat
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 21ms/step
1/1 [==============================] - 0s 19ms/step
1/1 [==============================] - 0s 17ms/step
nasihat langit penentram jiwa 4
```

Figure 7. Testing Results

The model that has been trained is then implemented in the application. This implementation is intended so that the next word prediction system for book title search can be enjoyed or used by users in finding the appropriate book title easily. The implementation is in the form of a book title search website that has been integrated with the book collection database obtained from the bukukita.com website. The book title search website was built using the Django framework, Python language, and MySQL database. Then on the website page use HTML and CSS. Figure 8, is the appearance of the book title search website that has been created.

Figure 8. Book Title Search Web View

The database that has been created is used to store book data and display book search results from the next word prediction model in web applications. The database consists of one table with the structure as in Table 2.

Table 2. Book Table Structure

#	Name	Data Type	Data Length
1	book_id (pk)	Integer	11
2	book_cover	Varchar	500
3	book_title	Varchar	255
4	stock	Varchar	20

D. Conclusion

Based on the research that has been done, the next word prediction system for book title search has been successfully built using the Bidirectional Long Short Term Memory (Bi-LSTM) algorithm and obtained good accuracy results. This is proven by

the results of the model evaluation which obtained an accuracy matrix of 81.82%. The model has been able to predict the next word for a book title search correctly. In addition, the model that has been made has better results than previous studies using Indonesian-language data. However, the model that has been made still has shortcomings in the form of some predictions that have not been appropriate due to the lack of diverse data obtained. Thus, further research is highly expected for better development.

E. References

[1] I. Admirani, "Penerapan Metode Fuzzy Time Series Untuk Prediksi Laba Pada Perusahaan," *Jurnal JUPITER*, vol. 10, no. 1, pp. 19–31, 2018.

[2] Sarker and I. H, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Comput Sci*, vol. 2, no. 6, p. 420, Nov. 2021, doi: 10.1007/s42979-021-00815-1.

[3] L. Alzubaidi *et al.*, "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, vol. 8, no. 1, p. 53, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

[4] C. Janiesch, P. Zschech, and K. Heinrich, "Machine learning and deep learning," *Electronic Markets*, vol. 31, no. 3, pp. 685–695, 2021, doi: 10.1007/s12525-021-00475-2/Published.

[5] R. L. Abduljabbar, H. Dia, and P.-W. Tsai, "Development and evaluation of bidirectional LSTM freeway traffic forecasting models using simulation data," *Sci Rep*, vol. 11, no. 1, p. 23899, Dec. 2021, doi: 10.1038/s41598-021-03282-z.

[6] Yugesh Verma, "Complete Guide To Bidirectional LSTM (With Python Codes)," *Analytics India Magazine*, Jul. 17, 2021. https://analyticsindiamag.com/complete-guide-to-bidirectional-lstm-with-python-codes/ (accessed Jun. 04, 2023).

[7] Q. Cheng, Y. Chen, Y. Xiao, H. Yin, and W. Liu, "A dual-stage attention-based Bi-LSTM network for multivariate time series prediction," *J Supercomput*, vol. 78, no. 14, pp. 16214–16235, 2022, doi: 10.1007/s11227-022-04506-3.

[8] N. Thaker and A. Shukla, "Python as Multi Paradigm Programming Language," *Int J Comput Appl*, vol. 177, no. 31, pp. 38–42, Jan. 2020, doi: 10.5120/ijca2020919775.

[9] T. M. Kadarina and M. H. Ibnu Fajar, "Pengenalan Bahasa Pemrograman Python menggunakan Aplikasi Games untuk Siswa/i di Wilayah Kembangan Utara," *Jurnal Abdi Masyarakat (JAM)*, vol. 5, no. 1, p. 11, Jul. 2019, doi: 10.22441/jam.2019.v5.i1.003.

[10] J. Kavander, "Developing Kanban board backend by using Django web framework," Laurea University of Applied Sciences, 2022.

[11] K. Duisebekova, in Physics, A. Professor, R. Khabirov, master student, and A. Zholzhan, "Django as Secure Web-framework in Practice," *The Bulletin of KazATC Вестник КазАТК*, vol. 1, no. 116, pp. 275–281, 2021, doi: 10.52167/1609-1817-2020-116-1-275-281.

[12] M. Ahmadar, P. Perwito, and C. Taufik, "Perancangan Sistem Informasi Penjualan Berbasis Web pada Rahayu Photo Copy dengan Database MySQL,"

Dharmakarya: Jurnal Aplikasi Ipteks untuk Masyarakat, vol. 10, no. 4, pp. 284–289, Dec. 2021, doi: 10.24198/dharmakarya.v10i4.35873.

[13] H. Dhika, N. Isnain, and M. Tofan, "Manajemen Villa Menggunakan Java Netbeans dan Mysql," *Jurnal IKRA-ITH Informatika*, vol. 3, no. 2, 2019.

[14] A. Rianti, S. Widodo, A. D. Ayuningtyas, and F. B. Hermawan, "Next Word Prediction Using LSTM," *Journal of Information Technology and Its Utilization*, vol. 5, no. 1, 2022.

[15] R. Sharma, N. Gael, N. Aggarwal, and C. Prakash, "Next Word Prediction in Hindi Using Deep Learning Techniques," 2019.

[16] S. Rajakumar, V. Rameshbabu, D. Usha, N. K, R. Shree, and S. Priya, "EXO Next Word Prediction Using Machine Learning," *J Surv Fish Sci*, vol. 10, no. 3S, pp. 4112–4118, 2023.

[17] Ms. S. Ramya and Dr. C. S. K. Selvi, "Recurrent Neural Network based Models for Word Prediction," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 4, pp. 7433–7437, Nov. 2019, doi: 10.35940/ijrte.D5313.118419.

[18] K. Wangchuk, T. Wangchuk, and T. Namgyel, "Dzongkha Next Words Prediction Using Bidirectional LSTM," *Bhutan Journal of Research and Development*, no. 2, Feb. 2023, doi: 10.17102/bjrd.rub.se2.038.

[19] K. Chakradhar, K. S. Kiran, K. Shanmukh, K. Sharath Kumar, K. Dinesh Sagar, and] Gmr, "Next Word Prediction Using Deep Learning," 2022. [Online]. Available: www.ijrpr.com

[20] M. Soam and S. Thakur, "Next Word Prediction Using Deep Learning: A Comparative Study," in *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, Jan. 2022, pp. 653–658. doi: 10.1109/Confluence52989.2022.9734151.

[21] N. Afrianto, D. H. Fudholi, and S. Rani, "Prediksi Harga Saham Menggunakan BiLSTM dengan Faktor Sentimen Publik," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, pp. 41–46, Feb. 2022, doi: 10.29207/resti.v6i1.3676.

[22] D. I. Puteri, "Implementasi Long Short Term Memory (LSTM) dan Bidirectional Long Short Term Memory (BiLSTM) Dalam Prediksi Harga Saham Syariah," *Euler : Jurnal Ilmiah Matematika, Sains dan Teknologi*, vol. 11, no. 1, pp. 35–43, May 2023, doi: 10.34312/euler.v11i1.19791.

[23] A. A. Ningrum *et al.*, "Algoritma Deep Learning-LSTM untuk Memprediksi Umur Transformator," vol. 8, no. 3, pp. 539–548, 2021, doi: 10.25126/jtiik.202184587.

[24] I. Nyoman, K. Wardana, N. Jawas, I. Komang, and A. A. Aryanto, "Prediksi Penggunaan Energi Listrik pada Rumah Hunian Menggunakan Long Short-Term Memory," 2020. [Online]. Available: http://journal.undiknas.ac.id/index.php/tiers

[25] H. Nurrohmah, "Klasifikasi Berita Hoax Berbahasa Indonesia Menggunakan Bidirectional Long Short Term Memory (Bi-LSTM)," Jakarta, Aug. 2022.

[26] D. Naik and C. D. Jaidhar, "A novel Multi-Layer Attention Framework for visual description prediction using bidirectional LSTM," *J Big Data*, vol. 9, no. 1, p. 104, Nov. 2022, doi: 10.1186/s40537-022-00664-6.

[27] M. G. Rizky, "Analisis Perbandingan Metode LSTM dan BiLSTM untuk Klasifikasi Sinyal Jantung Phonocardiogram," Surabaya, Aug. 2021.
[28] M. S. Anggreany, "Confusion Matrix," *School of Computer Science BINUS University* , Nov. 01, 2020. https://socs.binus.ac.id/2020/11/01/confusion-matrix/ (accessed Jun. 16, 2023).
[29] R. M. Sagala, "Prediksi Kelulusan Mahasiswa Menggunakan Data Mining Algoritma K-Means," *TeIKa*, vol. 11, no. 2, pp. 131–142, Oct. 2021, doi: 10.36342/teika.v11i2.2610.
[30] P. Niharika and S. J. J. Thangaraj, "Long Short Term Memory model-based automatic next word generation for text-based applications In contrast to the N-gram model," *J Surv Fish Sci*, vol. 10, no. 1S, 2023.