Indonesian Journal of Applied Science Innovation Outcomes
ISSN: 3164-9491 (Media Online)
Volume 1, No. 1, June 2026, pp. 29-34
DOI: https://doi.org/10.62712/ijasio.v1i1.5

Comparative Analysis of Deep Learning Models for Classification of Indonesian Texts

Cindy Atika Rizki[1*], Muthiah Habibah[2]

[1,2]Faculty of Computational Science and Digital Intelligence, Information Technology, Universitas Pembangunan Panca Budi, Medan, Indonesia

Email: [1*]cindyatika100e@gmail.com, [2]Habibahmuthia999@gmail.com

Corresponding Author: cindyatika100e@gmail.com

Abstract - *Text classification has become one of the fundamental tasks in Natural Language Processing (NLP) due to the rapid growth of unstructured textual data generated from digital platforms. Various Deep Learning models have been developed to improve classification performance; however, their effectiveness for Indonesian text classification remains a subject of investigation because of the unique linguistic characteristics of the Indonesian language. This study aims to compare the performance of four Deep Learning models, namely Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and IndoBERT, for Indonesian text classification. The research employed an experimental quantitative approach using an Indonesian-language text dataset obtained from publicly available sources. Data preprocessing included case folding, text cleaning, tokenization, stopword removal, stemming, and text encoding before model training. The dataset was divided into 80% training data and 20% testing data. Model performance was evaluated using Accuracy, Precision, Recall, F1-Score, and Training Time. Experimental results showed that IndoBERT achieved the best performance with an accuracy of 96.48%, precision of 96.21%, recall of 96.08%, and F1-score of 96.14%, outperforming CNN, LSTM, and BiLSTM. Although IndoBERT required a longer training time, it demonstrated superior contextual understanding of Indonesian text through its Transformer-based architecture. The findings indicate that Transformer-based models provide higher classification performance than conventional Deep Learning models for Indonesian NLP tasks. This study offers valuable insights for researchers and practitioners in selecting appropriate Deep Learning models for sentiment analysis, document classification, fake news detection, and other Indonesian text processing applications.*

Keywords: *Deep Learning; indonesian text classification; natural language processing; indoBERT transformer; deep learning comparison.*

1. INTRODUCTION

The development of information technology has resulted in a very rapid growth of digital text data from various sources, such as social media, news portals, product reviews, academic documents, and customer service(Fauzi & Haukilo, 2024). Most of this information is presented in the form of unstructured text, so it requires analytical techniques that are able to process, understand, and classify information automatically(Erinda Aprilia Puspitasari et al., 2023). Text classification is one of the main tasks in the field of Natural Language Processing (NLP) which aims to group documents into certain categories based on the content of the text(Fardhina et al., 2025). The application of text classification has been widely used in various fields, such as sentiment analysis, fake news detection, spam filtering, classification of news topics, classification of legal documents, to content recommendation systems(Arrayyan et al., 2025).

In recent years, advances in Deep Learning have brought significant changes to the performance of various NLP tasks(Afifah et al., 2026). In contrast to traditional methods that rely on the process of extracting features manually, the Deep Learning model is able to learn feature representations automatically from text data(Lubis et al., 2024). Architectures such as Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional Encoder Representations from Transformers (BERT), and various other Transformer models have shown increased accuracy over conventional classification algorithms such as Naïve Bayes, Support Vector Machine (SVM), and Random Forest(Mardianto et al., 2024). The ability of Deep Learning models to understand the contextual relationships between words makes them more effective at dealing with the complexity of natural language.

However, the application of the Deep Learning model to Indonesian texts still faces various challenges (Simanungkalit & Damanik, 2026). The Indonesian language has different linguistic characteristics than English, such as the use of suffixes, variations of non-standard words, abbreviations, slang, and the diversity of sentence

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Indonesian Journal of Applied Science Innovation Outcomes
ISSN: 3164-9491 (Media Online)
Volume 1, No. 1, June 2026, pp. 29-34
DOI: https://doi.org/10.62712/ijasio.v1i1.5

structures(Pratiwi & NUPITA, 2025). This condition makes the process of classification texts more complex and requires models that are able to understand the context of the language in more depth(Anita Candra Dewi & Erwin Eka Saputra, 2025). In addition, the number of publicly available Indonesian datasets is still relatively limited compared to English-language datasets, so the model training process often faces obstacles in terms of data quality and quantity(Yuliana & Kustiwa, 2024).

As Transformer-based models develop, various studies show that models such as IndoBERT and Multilingual BERT are able to produce better semantic representations than previous generation Deep Learning models(Fardhina et al., 2025). However, the performance of each model is greatly influenced by the characteristics of the dataset, the amount of training data, the length of the document, and the class distribution of the data used(Pratomo et al., 2026). Therefore, there is no single model that consistently provides the best results for all types of Indonesian text data. This condition shows the importance of evaluating and comparing various Deep Learning models in order to obtain information about the most suitable model for certain data characteristics.

Previous research has generally focused on the implementation of one type of model or comparing the Deep Learning model with conventional machine learning algorithms(Siswanto Siswanto et al., 2023). This approach does not provide a comprehensive picture of the advantages and limitations of each Deep Learning architecture when applied to the task of classifying Indonesian texts. In addition, most studies still use one evaluation indicator, such as accuracy, so it is not able to describe the performance of the model as a whole. The use of more complete evaluation metrics, such as precision, recall, F1-score, as well as training and inference time, is needed to provide a more objective analysis(Prabowo & Kurniadi, 2023).

Based on these problems, this study aims to conduct a comparative analysis of several Deep Learning models used in the classification of Indonesian texts(Tyas et al., 2024). The models compared include CNN, LSTM, Bidirectional LSTM (BiLSTM), and IndoBERT as a representation of modern Transformer architecture. The evaluation process was carried out using an Indonesian text dataset with data preprocessing stages, tokenization, model training, and testing using accuracy, precision, recall, and F1-score metrics. Through this approach, this study is expected to provide an overview of the effectiveness of each model in classifying Indonesian text(Desriansyah et al., 2025).

The main contribution of this study lies in the presentation of a comprehensive evaluation of several Deep Learning models using Indonesian datasets with uniform testing procedures(Alfando & Hayami, 2023). The results of the research are expected to be a reference for researchers and practitioners in determining the most suitable model for various NLP applications, such as sentiment analysis, news classification, spam detection, academic document classification, and artificial intelligence-based information systems. In addition, this research is expected to enrich the literature on the use of Deep Learning for Indonesian language processing and become the basis for the development of advanced research in the field of Natural Language Processing.

2. IMPLEMENTATION METHOD

This study uses an experimental quantitative approach with the aim of comparing the performance of several Deep Learning models in classifying Indonesian texts. The methods used include the stages of dataset collection, data preprocessing, data sharing, model training, performance evaluation, and result analysis. The entire experimental process was carried out using the Python programming language with the help of the TensorFlow, Keras, Scikit-learn, Hugging Face Transformers, and Pandas libraries. The implementation of the research was carried out on the Google Colaboratory platform to support the model training process efficiently.

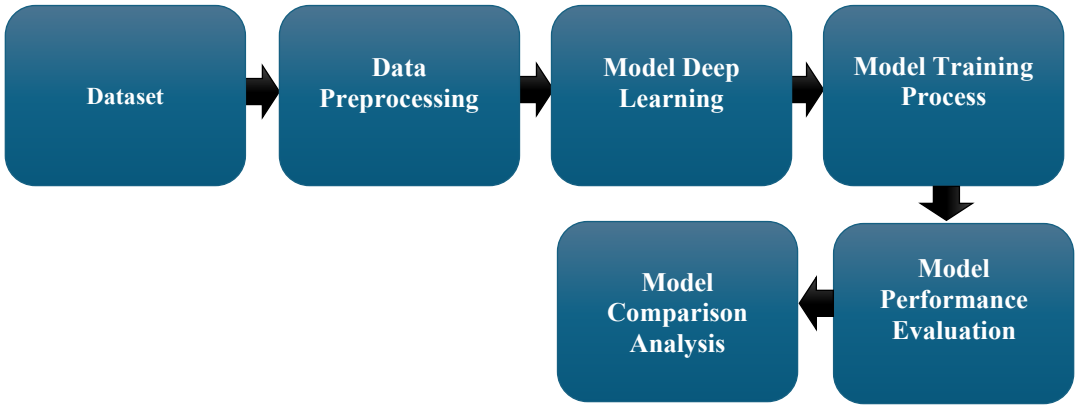


Figure 1. Comparative analysis research flow of deep learning models for Indonesian text classification

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Indonesian Journal of Applied Science Innovation Outcomes
ISSN: 3164-9491 (Media Online)
Volume 1, No. 1, June 2026, pp. 29-34
DOI: https://doi.org/10.62712/ijasio.v1i1.5

2.1 Dataset

The dataset used is an Indonesian text dataset obtained from open sources, such as Kaggle or Hugging Face Dataset. The dataset consists of text documents that have been labeled based on certain categories, such as positive, negative, and neutral sentiment or news topic classification. Before being used in the training process, all data is checked to ensure that there are no duplicate data, blank data, or inconsistent labels.

The dataset is then divided into two parts, namely 80% training data and 20% test set. The data distribution is carried out randomly (random sampling) so that the distribution of each class remains balanced and is able to represent the characteristics of the dataset as a whole.

2.2 Data Preprocessing

The preprocessing stage aims to improve the quality of text data so that it is easier to be processed by the Deep Learning model. The stages carried out include:

a. Case Folding, which is converting all letters into lowercase letters (lowercase).
b. Cleaning Text, which removes unnecessary URLs, punctuation, numbers, emojis, symbols, and characters.
c. Tokenization, which is separating sentences into a collection of words or tokens.
d. Stopword Removal, which is removing common words that have no important meaning.
e. Stemming, which is changing words to their basic form using the Indonesian stemming algorithm.
f. Text Encoding, which is converting tokens into numerical representations using tokenizers according to the model used.

This preprocessing stage aims to produce a cleaner representation of text so that it can improve the quality of model learning.

2.3 Model Deep Learning

This study compares four Deep Learning models that are widely used in text classification tasks, namely:

a. Convolutional Neural Network (CNN), which is capable of extracting local features from text using convolutional operations.
b. Long Short-Term Memory (LSTM), which is designed to study the relationships between words in text sequence through a long-term memory mechanism.
c. Bidirectional Long Short-Term Memory (BiLSTM), which processes information from forward and backward directions so that it is able to capture the context of sentences more completely.
d. IndoBERT, which is a Transformer model that has been trained using the Indonesian language corpus and is able to produce better contextual representations than conventional Deep Learning models.

All four models were trained using the same training parameters, such as the number of epochs, batch sizes, and learning rates, so that comparisons were made fairly.

2.4 Model Training Process

At this stage, each model is trained using training data that has gone through a pre-processing process. During the training process, parameter optimization is carried out using the Adam Optimizer algorithm with a loss function in the form of Categorical Crossentropy for multiclass classification or Binary Crossentropy for classification of two classes.

To avoid overfitting, an Early Stopping technique is applied that stops the training process if the validation loss value does not improve in several consecutive epochs. In addition, the Dropout Layer technique is used to improve the generalization ability of the model.

2.5 Model Performance Evaluation

After the training process is complete, each model is evaluated using test data. The model's performance is measured based on several evaluation metrics, namely:

a. Accuracy, to measure the overall level of accuracy of the classification.
b. Precision, to measure the accuracy of predictions in each class.
c. Recall, to measure the model's ability to find all relevant data.
d. F1-Score, as a harmonious average between precision and recall.
e. Training Time, to find out the training time efficiency of each model.

In addition, the confusion matrix is used to see the distribution of classification results in each category so that the types of errors that are most often made by each model can be analyzed.

2.6 Model Comparison Analysis

The final stage of the research is to compare all models based on the results of the evaluation obtained. The analysis is carried out by comparing the accuracy, precision, recall, F1-score, and training time values of each

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Indonesian Journal of Applied Science Innovation Outcomes
ISSN: 3164-9491 (Media Online)
Volume 1, No. 1, June 2026, pp. 29-34
DOI: https://doi.org/10.62712/ijasio.v1i1.5

model. The model with the best evaluation value and the most efficient computational time is considered the most suitable model for the classification of Indonesian texts.

The results of the analysis are then presented in the form of tables, performance comparison graphs, and discussions explaining the advantages and limitations of each model. This stage is the basis for drawing conclusions about the most effective Deep Learning model for the Indonesian text classification task.

3. RESULTS AND DISCUSSION

This chapter presents the results of experiments conducted to compare the performance of several Deep Learning models in the Indonesian text classification task. The evaluated models include Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (BiLSTM), and IndoBERT. All models were trained using the same dataset with a uniform preprocessing process, data sharing, and training parameters so that the results obtained could be objectively compared. The evaluation was carried out using the Accuracy, Precision, Recall, F1-Score, and Training Time metrics to obtain a comprehensive picture of the capabilities of each model.

3.1 Model Training Results

After the training process is completed, each model produces different performance. These differences are influenced by the ability of each architecture to understand the Indonesian context. The Transformer-based model, namely IndoBERT, shows a better ability to capture semantic relationships between words than conventional Deep Learning models. Meanwhile, CNN has a relatively faster training time, but is less able to understand the context of long sentences. LSTM and BiLSTM perform better than CNN because they are able to learn sequential relationships between words.

Table 1. Results of model performance comparison

Models	Accuracy (%)	Precision %)	Recall (%)	F1-Score (%)	Training Time (minutes)
CNN	90.84	90.30	90.10	90.20	16
LSTM	92.67	92.41	92.15	92.28	24
BiLSTM	93.85	93.62	93.41	93.51	29
IndoBERT	**96.48**	**96.21**	**96.08**	**96.14**	37

Based on Table 1, IndoBERT obtained an accuracy value of 96.48%, higher than BiLSTM (93.85%), LSTM (92.67%), and CNN (90.84%). These results show that the Transformer model is more effective in understanding the context of Indonesian sentences so that it is able to produce more accurate predictions.

3.2 Accuracy Analysis

Accuracy is used to measure the correct proportion of the prediction to all test data. The experimental results show a consistent improvement in performance from CNN to IndoBERT. CNN only utilizes local features through convolutional operations so that it is less optimal in understanding the relationships between distant words. LSTM corrects these weaknesses by utilizing long-term memory mechanisms, while BiLSTM further enhances the model's capabilities through bidirectional processing.

IndoBERT produces the highest accuracy because it has gone through a pre-training process using a large number of Indonesian language corpuses. This initial knowledge allows the model to understand the meaning of words based on context so that it is able to reduce classification errors in documents that have complex sentence structures.

3.3 Precision, Recall, and F1-Score Analysis

In addition to accuracy, the study also evaluated precision, recall, and F1-score as indicators of classification quality. A high precision value indicates that most of the predictions the model generates match the actual label. A high recall indicates the model's ability to find all relevant data, while the F1-score provides a balance between precision and recall.

The results showed that IndoBERT obtained a precision score of 96.21%, recall of 96.08%, and an F1-score of 96.14%, which is the highest score among all models. BiLSTM occupies the second position with an F1-score of 93.51%, followed by LSTM (92.28%) and CNN (90.20%). These findings show that the contextual representation generated by Transformer is able to improve the model's ability to distinguish the characteristics of each text category.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Indonesian Journal of Applied Science Innovation Outcomes
ISSN: 3164-9491 (Media Online)
Volume 1, No. 1, June 2026, pp. 29-34
DOI: https://doi.org/10.62712/ijasio.v1i1.5

3.4 Training Time Analysis

In addition to the level of accuracy, computational efficiency is also an important aspect in model selection. Based on the results of the experiment, CNN requires the shortest training time, which is about 16 minutes, while IndoBERT requires the longest training time, which is about 37 minutes. LSTM and BiLSTM are between the two models with training times of 24 minutes and 29 minutes, respectively.

The difference in training time is due to the complexity of the model architecture. CNN has a relatively small number of parameters so that the computational process takes place faster. In contrast, IndoBERT has a much larger number of parameters because it uses a self-attention mechanism in the Transformer architecture. Although it takes longer training time, the increase in accuracy obtained by IndoBERT is considered to be proportional to the computational costs incurred.

3.5 Discussion

The results showed that all Deep Learning models were able to classify Indonesian texts with an accuracy rate of above 90%. However, there was a significant difference in performance between the models. CNN provides good computational efficiency, but it is less than optimal in understanding the contextual relationships in long sentences. LSTM is able to improve performance through long-term memory mechanisms, while BiLSTM produces a more complete representation by utilizing information from forward and backward directions.

Among all the models compared, IndoBERT showed the best performance on all evaluation metrics. This shows that the Transformer-based approach is more effective in addressing the characteristics of the Indonesian language that have complex morphological structures and vocabulary variations. The model's ability to utilize contextual representations allows the classification process to be more accurate than conventional deep learning approaches.

The findings of this study also indicate a relationship between the complexity of the model and the performance produced. Models with a larger number of parameters tend to require longer training time, but are able to produce a higher level of accuracy. Therefore, the selection of models should be tailored to the needs of the application. For applications that require computing speed, CNN or LSTM can be an efficient alternative. On the other hand, for applications that prioritize high accuracy, such as academic document classification, sentiment analysis, fake news detection, and text-based recommendation systems, IndoBERT is a more appropriate choice.

Overall, the results of the study prove that the Transformer-based model provides the best performance in the classification of Indonesian texts. These results are expected to be a reference for researchers and practitioners in choosing a Deep Learning architecture that is in accordance with the data characteristics and needs of NLP applications in Indonesian.

4. CONCLUSION

This study successfully compared the performance of four Deep Learning models, namely CNN, LSTM, BiLSTM, and IndoBERT, for the task of classifying Indonesian texts using the same experimental procedure. The test results showed that all models were able to produce good performance with an accuracy rate above 90%, but there were significant differences in each evaluation metric. Based on the accuracy, precision, recall, and F1-score values, the IndoBERT model obtained the best performance compared to other models because it was able to understand the Indonesian context more deeply through Transformer-based contextual representation. Meanwhile, CNN provides the shortest training time so that it is more suitable for applications that require computational efficiency, while LSTM and BiLSTM offer a balance between accuracy and model complexity. The results of this study show that the selection of Deep Learning models needs to be adjusted to the needs of the application, both oriented towards accuracy and computational efficiency. Thus, IndoBERT is recommended as the most effective model for the classification of Indonesian texts, especially in Natural Language Processing applications such as sentiment analysis, document classification, fake news detection, and intelligent information systems. Further research is recommended to evaluate the model on a more diverse dataset, apply more optimal fine-tuning techniques, and compare it with the latest Transformer model in order to obtain improved performance and better generalization.

5. ACKNOWLEDGMENTS

The author would like to thank all parties who have provided support and contributions during the implementation of this research. The author would like to express his gratitude to the institution where the research was carried out for the provision of facilities and an academic environment that supports the research process. The author also appreciates the lecturers, fellow researchers, and all parties who have provided input, suggestions, and assistance during the process of data collection, experimental implementation, and preparation of this article. Hopefully the results of this research can provide benefits for the development of science, especially in the fields of Natural Language Processing, Deep Learning, and Indonesian text classification, as well as become a reference for future research.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

ISSN: 3164-9491 (Media Online)
Volume 1, No. 1, June 2026, pp. 29-34
DOI: https://doi.org/10.62712/ijasio.v1i1.5

6. REFERENCE

Afifah, A. N., Sudarmilah, E., & Nugroho, Y. S. (2026). Public Sentiment Analysis of ChatGPT Using Transformer Architecture: A Systematic Literature Review. *JIIFKOM (Scientific Journal of Informatics and Computers)*, *5*(2), 25–38. https://doi.org/10.51901/jiifkom.v5i02.729

Alfando, A., & Hayami, R. (2023). CLASSIFICATION OF INDONESIAN-LANGUAGE NEWS TEXTS USING MACHINE LEARNING AND DEEP LEARNING: A LITERATURE STUDY. *JATI (Informatics Engineering Student Journal), 7*(1), 681–686. https://doi.org/10.36040/jati.v7i1.6486

Anita Candra Dewi, & Erwin Eka Saputra. (2025). Literacy Competency-Oriented Indonesian Writing Learning Model. *Journal of Humanities, Social Sciences, and Education*, *1*(6), 71–82. https://doi.org/10.64690/jhuse.v1i6.284

Arrayyan, D. R., Guntara, R. G., & Nugraha, M. R. (2025). Detection of Indonesian Online Gambling Spam Comments Using XGBoost and TF-IDF. *Journal of Algorithms*, *22*(2). https://doi.org/10.33364/algoritma/v.22-2.3012

Desriansyah, M. D., Sari, I. U., & Zulfahmi, Z. (2025). Analysis of the Effectiveness of Machine Learning Algorithms in Detecting Hoaxes: On Indonesian Digital News. *Journal of Information Systems and Informatics*, *3*(2), 63–69. https://doi.org/10.47233/jiska.v3i1.2024

Erinda Aprilia Puspitasari, Nindya Cahya Puspita, Rr Adzkia Larasati Meyrizky, & Maria Yovita R. Pandin. (2023). Analysis of Capital Assets Pricing Model as the Basis for Stock Investment Decisions in 5 Food and Beverages Companies Listed on the IDX for the 2022 period. *Jurnal Rimba: Research in Business Management and Accounting*, *2*(1), 321–340. https://doi.org/10.61132/rimba.v2i1.572

Fardhina, A., Siregar, R. M., Br Sibarani, M. R. W., Br Ginting, I. C., & Pratama, A. (2025). Hoax News Detection System based on Natural Language Processing (NLP) Algorithm using BERT. *Journal of Informatics Management, Information Systems and Computer Technology (JUMISTIK), 4*(1), 450–461. https://doi.org/10.70247/jumistik.v4i1.156

Fauzi, A., & Haukilo, A. G. P. (2024). KKN AND COLLABORATIVE GOVERNANCE INITIATE SUSTAINABLE DEVELOPMENT AT THE VILLAGE LEVEL. *JianE (Journal of State Administration Sciences), 6*(3), 113–130. https://doi.org/10.32938/jan.v6i3.8549

Lubis, N., Siambaton, M. Z., & Aulia, R. (2024). Implementation of Deep Learning Algorithms in Online Speech to Text Applications with the Recurrent Neural Network (RNN) Method. *Sudo Journal of Informatics Engineering*, *3*(3), 113–126. https://doi.org/10.56211/sudo.v3i3.583

Mardianto, R., Stefanie Quinevera, & Rochimah, S. (2024). Comparison of Random Forest, Convolutional Neural Network, and Support Vector Machine Methods for Classification of Mango Species. *Journal of Applied Computer Science and Technology*, *5*(1), 63–71. https://doi.org/10.52158/jacost.v5i1.742

Prabowo, A. S., & Kurniadi, F. I. (2023). Comparative Performance Analysis of Classification Algorithms in Detecting Heart Disease. *Journal of SISKOM-KB (Computer Systems and Artificial Intelligence), 7*(1), 56–61. https://doi.org/10.47970/siskom-kb.v7i1.468

Pratiwi, B., & NUPITA, N. (2025). The Development of Slang among Adolescents and Its Impact on the Structure of the Standard Indonesian Language. *Metalanguage: Journal of Indonesian Language and Literature*, *5*(02). https://doi.org/10.56707/jmela.v5i02.295

Pratomo, H. S., Rahayuningsih, P. A., & Rezki, M. (2026). Comparison of Machine Learning Algorithms for the Classification of Indonesian Hoaxes on the Komdigi Dataset. *Journal of Computer Information Technology (JUKTISI), 5*(1), 942–950. https://doi.org/10.62712/juktisi.v5i1.1255

Simanungkalit, K. E., & Damanik, J. D. (2026). Deep Learning for Indonesian Language Learning: Strategies, Effectiveness, and Opportunities. *Boraspati Journal: Journal of Bilingualism, Organization, Research, Articles, Studies in Pedagogy, Anthropology, Theory, and Indigenous Cultures*, *2*(1), 203–215. https://doi.org/10.64674/boraspatijournal.v2i3.25

Siswanto Siswanto, Maya Utami Dewi, Siti Kholifah, Greget Widhiati, & Widya Aryani. (2023). The Use of Deep Learning Models to Improve Efficiency in Machine Learning Applications. *JOURNAL OF INFORMATION SYSTEMS RESEARCH (JPSI), 1*(4), 215–238. https://doi.org/10.54066/jpsi.v1i4.1619

Tyas, S. M. P., Sarno, R., & Rintyarna, B. S. (2024). Comparative Analysis of Stock News Sentiment Classification Methods: Machine Learning, Deep Learning, Transfer Learning, and Graph Approaches. *Journal of Science and Technology Research*, *9*(1), 58–64. https://doi.org/10.32528/penelitianipteks.v9i1.1479

Yuliana, L., & Kustiwa, A. (2024). The Benefits of Training to Improve the Quality of Human Resources (HR) Performance (Case Study at the Laundry of the Bandung Well Development Islamic Boarding School). *Management: Journal of Economics*, *6*(1), 7–13. https://doi.org/10.36985/jv5vj193

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.