IJASST
International Journal of
Applied Sciences and Smart Technologies
p-ISSN 2655-8564, e-ISSN 2685-943, pp. 153-168, 2026

Volume 08
Issue 01

# Tourism News Classification Using Convolution Long Short-Term Memory (C-LSTM)

Yoga Dwitya Pramudita[1], Husni [1], Mohammad Syarief [1], Eka Mala Sari Rochman [1], Arif Muntasa [1], Zahra Arwananing Tyas [2], Ika Oktavia Suzanti [1]

[1] *Informatics Engineering Department, Faculty of Engineering, University of Trunodjoyo Madura, Bangkalan 69162, Indonesia*
[2] *Information Technology Program, Faculty of Business Management and Information Technology (FBIT), Universiti Muhammadiyah Malaysia, Perlis, 02100, Malaysia*
[*]*Corresponding Author: yoga@trunojoyo.ac.id*

(Received 23-10-2026; Revised 21-01-2026; Accepted 02-03-2026)

**Abstract**
Along with the rapid development of information technology, news about Indonesian tourism destinations can now be accessed widely through various platforms such as social media and online news, making news easily accessible. With the increase in tourism news, manual news classification is less effective in dividing data into various subcategories, such as natural tourism, artificial tourism, cultural tourism, and non-tourism. An algorithm is needed to address this problem, one of which uses an algorithm from deep learning. This study developed a tourism news classification model using Convolutional Long Short-Term Memory (C-LSTM) and Word2Vec Representation with Continuous Bag of Words (CBOW) architecture to obtain better accuracy and computational efficiency, and is used to produce better word vectorization, so that semantic relationships between words can be captured. This study used a news dataset of 5261 and news with an 80:20 ratio for training and testing. With this approach, the highest accuracy value of 94% was obtained with a time of 1140 seconds.

**Keywords**: News classification, tourism, C-LSTM, CBOW, computational efficiency.

## 1 Introduction

Classification is the process of categorizing an object into a certain category based on the similarity of the characteristics of the data it has [1]. For example, in the context of tourism in East Java, classification is used to group tourist destinations into three categories, namely, natural tourism, artificial tourism, and cultural tourism. This

classification concept is also applied in various fields, including news classification, which aims to group news according to certain characteristics or categories [2]. Considering the increasing number of tourism news, manual news classification is not effective in grouping news into good tourism subcategories, which can affect the reader's experience [3]. Classification of news texts can be done through two approaches, namely binary classification (two categories) and multiclass categories (more than two categories)[4].

News is a crucial source of information for tourists planning their trips, providing guidance on destinations, accommodations, and activities [5]. Tourism also plays a crucial role in increasing regional income. Emphasis is needed on developing the tourism sector through improving facilities, human resources, and preserving culture and traditions to create a comfortable holiday environment[6]. With the rapid advancement of information technology, news about tourism destinations is now widely accessible through various platforms such as social media and online news [7].

This study uses Word2Vec with the Continuous Bag of Words (CBOW) method, to generate word vector representations based on the semantic context around the word [8], allowing the model to capture deeper meaning and context from news texts. Although CBOW is effective in improving classification accuracy, its weakness is its limited ability to capture complex sentence contexts or variations in word meanings [9]. In addition, the combination of CBOW and C-LSTM allows for more efficient handling of tourism news, because CBOW helps improve accuracy without increasing the computational burden. Research [10] previously showed that CBOW can improve model performance in text classification by providing more accurate word representations, especially in documents with diverse content.

Several algorithms fall into the binary classification and multiclass classification categories with deep learning, including Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Convolutional LSTM (C-LSTM). This study develops a tourism news classification model using Convolutional Long Short-Term Memory (C-LSTM) [9] [11] [12] [13]. The combination of these two approaches is often used for text classification, allowing the model to utilize

both sequence and spatial information. Thus, these various algorithms can produce better performance in text processing [8] [14] [15].

Convolutional Neural Networks (CNNs), which are often used in image data processing, have also been shown to have significant capabilities in text processing, even surpassing the performance of other methods based on the results of research that has been done [12]. The main weakness of CNNs is their limitation in handling sequential data that has strong and long-time dependencies. Training CNN models often requires large computing power and significant time due to the complexity of their architecture and the large number of parameters. The process of recurrent word processing on long text data takes quite a long time because the model must analyze each word and its relationship to the context thoroughly. Meanwhile, Long Short-Term Memory (LSTM) is a modification of Recurrent Neural Networks (RNNs), which are specifically designed to overcome the obstacles encountered in RNNs, namely gradient vanishing and exploding. LSTMs have three important memory cell components, namely forget gates, input gates, and output gates that are able to handle long-term dependencies in sequential data [15].

C-LSTM method is a combination of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) and has been proven effective in classifying Indonesian-language news [16]. The research results showed that C- LSTM achieved an f1-score of 93.27%, outperforming the LSTM method by 2.4% and the CNN method by 3.42%. This study also observed that using a larger batch size can reduce the f1-score accuracy compared to a smaller batch size. Hyperparameters such as a batch size of 16 and a learning rate of 0.001 and the adam optimizer also played a significant role in achieving the optimal f1-score value.

The literature review section of this study examines research that applies CNN and LSTM models. A summary of the methods, feature extraction techniques, datasets used, and the results of these studies is presented in Table 1.

**Table 1.** Comparison of Previous Research

| *Study* | *Method* | *News Article Language* | *Features extraction* | *Dataset size* | *Category* | *Eval* |
|---|---|---|---|---|---|---|
| Li et al. , 2018 [8] | Bi-LSTM-CNN | Chinese | Word2vec - CBOW | 65000 | 10 | 96.45 |
| Jianping Li et al. 2019 [14] | Bi-LSTM + hierarchical attentions | Chinese | One hot encoding | 45000 | 14 | 98.65 |
| Minyong Shi et al. , 2019 [9] | C-LSTM + Word Embedding | Chinese | Word2vec - CBOW | 228000 | 18 | 86.53 |
| Widhiyasana Y et al. 2021 [16] | C-LSTM | Indonesian | - | 5000 | 4 | 93.27 |
| Amalia J et al. , 2022 [11] | Bi-LSTM + Word2Vec | English | Word2vec - CBOW | 9805 | 2 | 92.8 |
| Alroobaea, R. 2023 [12] | CNN+LSTM | Arabic | Bag of word | 5000 | 6 | 93.15 |
| Sultan D., 2023 [15] | CNN+LSTM | English | Word2vec | 22,324 | 6 | 97.52 |
| Sirait et al., 2024[13] | Bi-LSTM + GloVe | English | Glove | 4000 | 2 | 95 |

Referring to the previous explanation, this study focuses on developing an effective tourism news classification model using the Convolutional Long Short-Term Memory (C-LSTM) algorithm and Word2Vec vector representation. This approach aims to classify tourism news to obtain accuracy, precision, recall, accuracy, and F1-score of C-LSTM in handling news text classification.

# 2 Material and Methods

### Text Classification

Text classification is a type of machine learning problem that aims to classify data into more than two classes or categories. In the context of the research entitled "Tourism News Classification Using the Convolutional Long Short-Term Memory (C-LSTM) Method with Word2Vec Representation," text classification refers to the process of grouping tourism news into several different categories based on the content of the news

text. For example, tourism news can be categorized into categories such as "nature tourism," "cultural tourism," "artificial tourism," and so on.

**Word Embedding**

Word Embedding is a method in Natural Language Processing (NLP) that converts words into high-dimensional numeric vector representations, designed to capture the meaning and relationships between words. In this study, Word Embedding was created using the Word2Vec method. This method is used to identify hidden relationships between words. Word2Vec consists of two architectural models, namely Continuous Bag-of-Words (CBOW) and Skip-Gram. The CBOW model predicts target words based on the context of surrounding words, while the Skip-Gram model predicts potential words that can be context for the target word. At this stage, text or strings are converted into numbers or vectors for processing by the deep learning architecture. Input must have a consistent shape and size in the form of numbers or vectors. Because sequential data often has varying lengths, the padding process is applied using pad_sequences from keras. Padding is a special form of masking that can be added at the beginning or end of a sequence. In addition, the maximum number of words per sentence can be specified, and the input sentences must be tokenized using the tokenizing tool from keras [17].

**C-LSTM**

CNN and LSTM are two approaches frequently used to address text classification problems. However, in text processing, CNN often ignores the contextual relationships between words in a text document. Due to this limitation, many researchers prefer to use LSTM, which is superior in handling contextual relationships in data. Nevertheless, CNN has advantages over LSTM in the process of feature extraction or retrieving features from news. Therefore, a combination of CNN and LSTM is often used to exploit the advantages of each method [16].

By combining CNN and LSTM, news is fed into a deep learning model and both are obtained simultaneously. The CNN and LSTM processes are combined to produce classification predictions. This combined approach demonstrates higher accuracy than when the two methods are run separately.

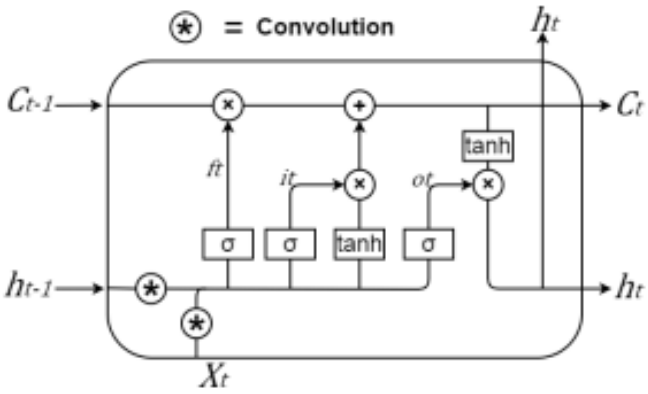


**Figure 1.** C-LSTM diagram

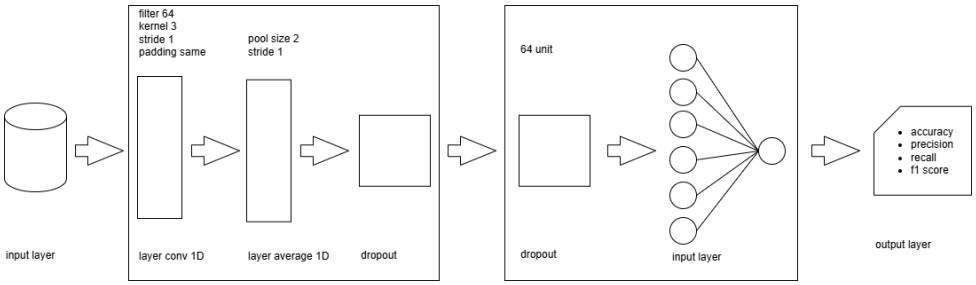


**Figure 2.** C-LSTM Model with Word2Vec Representation [11]

The LSTM model can be combined with a convolution process. This model is called C-LSTM. C-LSTM is almost the same as a regular LSTM, the difference being that the matrix multiplication in LSTM is replaced with a convolution operation on each gate. The C-LSTM diagram is as shown in Fig. 1. Based on Fig. 2. in the C-LSTM model with a modified Word2Vec representation from previous research [16]. The C-LSTM model uses convolution and pooling layers to extract features from the input data. Next, the results from the CNN will be fed into the LSTM. LSTM is used to process the input data from the results of the CNN and continued using fully connected + ReLu. Then the results of the LSTM will be used to produce accuracy, precision, recall, and f1-score values.

**System Architecture**

This chapter explains the stages of the research, from data collection to system implementation with Input Process Output (IPO), as shown in Fig. 3.

Fig. 3 shows the system architecture (IPO) where there are research steps starting from data collection from the detik.com site. After the data is collected, the data will be labeled to be grouped into several categories. After the data is labeled, the Pre-processing stage continues. In the Pre-processing stage there are 4 stages, namely Cleaning, Case Folding, tokenizing, and Stopword Removal. Data that has passed the Pre-processing process will continue to the Feature Extraction stage, namely Word2Vec with the CBOW architecture where the results of the Pre-processing in the form of words will be processed in vector form. Next, the data that has been converted into vector form will be divided into training data and testing data. The data will be divided into 80% as training data and 20% as testing data. At this stage, the data that has been separated into training data and test data will be classified into the C-LSTM model. The C-LSTM model is trained with the training data, after which the test data is fed into the trained C-LSTM model. The trained C-LSTM model will then be evaluated using test data. After the evaluation process is complete, the model will produce accuracy, precision, recall, and f1-score.

**Dataset**

The first stage is data collection as input for the research. The dataset used is the result of crawling using tourism keywords include tags information from the detik.com news portal, with a total of 5261 news documents. This news text is then saved in a comma-separated value (CSV) file. Saving to a CSV file makes it easier to read news documents in the text pre-processing stage. At the data stage, the news text labels are divided into four categories based on article tags, namely "Natural Tourism", "Artificial Tourism", "Cultural Tourism", "Non-Tourism". Can be seen in Table 2. Data was taken from October 16, 2023 to November 20, 2023. The data can be accessed at the link datasets.

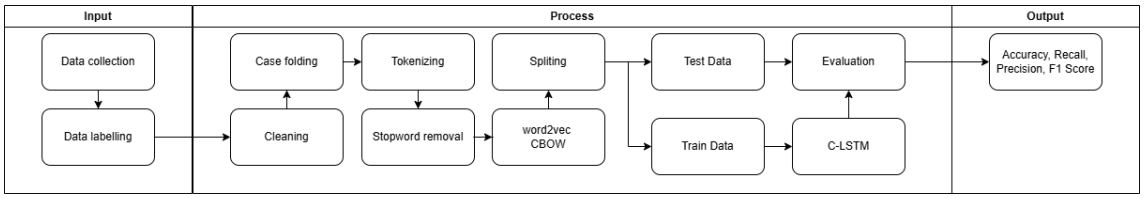


**Figure 3.** Input Process Output (IPO)

**Table 2.** Number of data for each category

| *No* | *News Category* | *Total* |
|---|---|---|
| 1. | Natural tourism | 807 |
| 2. | Artificial Tourism | 1441 |
| 3. | Cultural Tourism | 1168 |
| 4. | Not Tourism | 1845 |
| Total | | 5261 |

Table 2 shows the number of each news category that has been obtained with a total of 5,261 tourism news data, there are four news categories used in this study, namely natural tourism, artificial tourism, cultural tourism and non-tourism. News categories are determined based on the tag information on the article provided by the author. The next process involves experts (from the Bangkalan Tourism Office) in the news category labeling process.

Based on Fig. 2 in the C-LSTM model that has been modified from previous research [16]. In the C-LSTM model, the input comes from the embedding layer generated by Word2Vec (CBOW) to convert text into numeric vectors. This method converts text into high-dimensional numeric vectors that capture the semantic relationships and deeper meanings between words. By providing more accurate word representations, CBOW helps improve classification accuracy without increasing the computational burden, which is particularly useful for limited dataset with diverse content. Next, the CNN layer extracts spatial features through Conv1D (64 filters, kernel 3, ReLu, padding same) and MaxPooling (pool size 2) to reduce the feature dimension. The results are processed by LSTM (128 units) to understand the order and context of words in the text, then passed to the dense and dropout layers to prevent overfitting. The model finally produces tourism news classification output that is evaluated using accuracy, precision, recall, and F1-score.

**Test Scenario**

The test scenario was conducted to answer the research questions. To identify model performance issues, including computation, accuracy, precision, recall, and f1-score, a training process was conducted using a baseline model using C-LSTM. The CNN parameters used in this test can be seen in Table 3 [16].

The parameters for LSTM classification used in testing are taken from previous research where the batch sizes used are 32, 64, and 128. Where a smaller batch size will produce more batches, when the amount of data processed remains the same. Similarly to the batch size, the learning rate used is 0.001 and 0.0001. The epochs used are 150 and 200 [16]. Using the accuracy, precision, recall, and F1-score matrices on the test data for each model. As seen in Table 4.

**Table 3**. Parameters on CNN [16]

| Parameter | Test |
|---|---|
| *Convolutional layer* | *Filter =64*<br>*Kernel = 3*<br>*Padding = same*<br>*ReLu* |
| *Pooling* | *Max pooling, size pooling =2, stride =1* |

**Table 4.** Parameters in LSTM [16]

| Testing | *epoch* | *Batch size* | *Learning rate* |
|---|---|---|---|
| 1 | 150 | 32 | 0.001 |
| | | 64 | |
| | | 128 | 0.0001 |
| 2 | 200 | 32 | 0.001 |
| | | 64 | |
| | | 128 | 0.0001 |

# 3 Results and Discussions

### Classification of C-LSTM Methods

In the test scenario results, the best value was searched with epoch [150, 200], batch size [32, 64, 128], and learning rate [0.001, 0.0001]. In this training, the accuracy obtained after evaluation by finding the best value of the epoch, batch size, and learning rate that had been determined was around 94%, where the highest accuracy was obtained when using a combination of epoch value = [200], batch size = [64], and learning rate = [0.001] with an estimated time of 1140 seconds. The fastest combination of execution was epoch

value = [200], batch size = [128], and learning rate = [0.0001] as much as 640 seconds. The results of finding the best epoch value can be seen in Table 5.

The confusion matrix of the test results with C-LSTM on each parameter with the best accuracy can be seen in Fig. 5. By looking at this confusion matrix, it can be understood that the model predicts the correct and incorrect categories in each test. Fig. 5 illustrates the performance of the classification model using epoch 200 in four different categories. For category 1, the model successfully identified 204 data points correctly as category 1 (True Positives) and only misidentified 9 data points (False Negative) and 13 data points (False Positive). In addition, the model correctly identified 1213 data points as not being in category 1 (True Negatives). Similar performance was seen in category 2 with 380 True Positives, 15 False Positives, 1018 True Negatives, and 32 False Negatives. In category 3 with 304 True Positives, 35 False Positives, 1078 True Negatives, and 13 False Negatives. Finally, for category 4, the model showed 465 True Positives, 23 False Positives, 919 True Negatives, and 32 False Negatives.

**Table 5.** Testing Scenario

| ***Epoch*** | ***Batch size*** | ***Learning rate*** | ***Time (second)*** | ***accuracy*** |
|---|---|---|---|---|
| 150 | 32 | 0.001 | 1260 | 90% |
| 150 | 32 | 0.0001 | 1120 | 92% |
| 150 | 64 | 0.001 | 850 | 93% |
| 150 | 64 | 0.0001 | 916 | 90% |
| 150 | 128 | 0.001 | 916 | 88% |
| 150 | 128 | 0.0001 | 911 | 82% |
| 200 | 32 | 0.001 | 1260 | 90% |
| 200 | 32 | 0.0001 | 1020 | 92% |
| **200** | **64** | **0.001** | **1140** | **94%** |
| 200 | 64 | 0.0001 | 1020 | 90% |
| 200 | 128 | 0.001 | 720 | 90% |
| 200 | 128 | 0.0001 | 640 | 88% |

After obtaining the test results from all scenarios, a graph was then created for the overall results. The graph of the scenario results can be seen in Fig. 6. Based on Fig. 6, it shows a comparison graph of accuracy (%) from 12 different scenarios in one trial. This graph has a fluctuating trend with the highest accuracy in scenario 3 (93.26%) and scenario 9 (93.9%), and one significant low point in scenario 6 (81.35%). The initial accuracy in scenario 1 starts from 89.8%, then increases until it reaches its first peak in scenario 3. After that, there is a gradual decline until it reaches its lowest point in scenario 6. After this decline, the accuracy increases again and reaches its second peak in scenario 9, before fluctuating again until scenario 12. Changes in each parameter or method used in the scenario have a significant impact on the accuracy of the model.

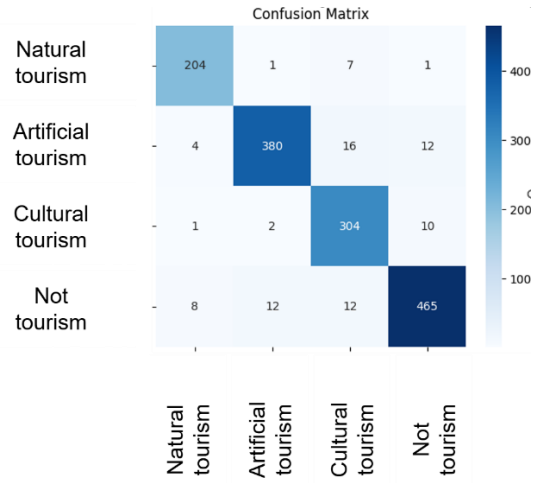


**Figure 5.** Confusion matrix epoch 200

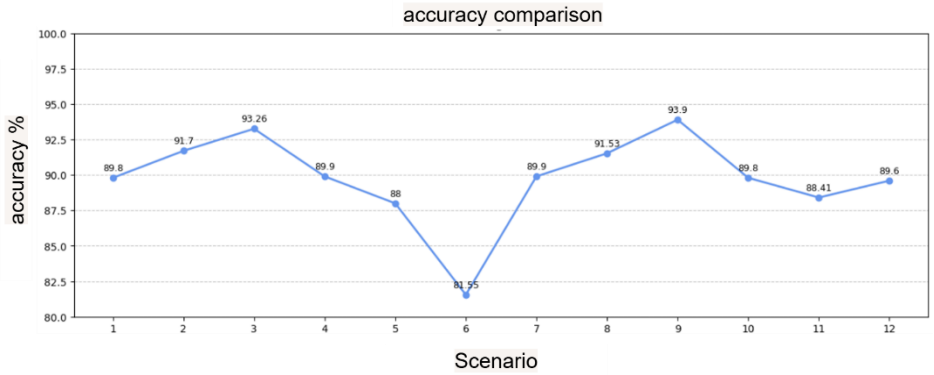


**Figure 6.** Model Analysis

Scenarios 3 and 9 are the best configurations in this trial test, while scenario 6 shows the worst performance which may require further evaluation of the factors causing the decrease in accuracy.

# 4 Conclusions

From the research that has been carried out, the results of parameter tuning, the best combination for the model is epoch 200, batch size 64, and learning rate 0.001, which produces 94% accuracy with an execution time of 1140 seconds. Although epoch 200, batch size 128 and learning rate 0.0001 provide faster execution time, the accuracy is the smallest at 82% of epoch 150, batch size 128, and learning rate 0.0001. The scenario result graph also supports that this best combination provides optimal performance. Therefore, to achieve the best results, it is recommended to use epoch 200, batch size 64, and learning rate 0.001, because it provides the best balance between accuracy and execution time.

First, our dataset is limited to a single news article source, which may not represent other regions or languages. In future work, it would be beneficial to collect and analyze datasets from a wider variety of sources to improve the generalization of our method.The resulting model still has prediction errors. These limitations are caused by several factors, such as mixed news content, less specific words, and language ambiguity.

# Acknowledgements

Researcher say thank you to the Research and Community Service Institute to the Community (LPPM) University of Trunojoyo Madura, for the opportunity to do research.

# References

[1] Haidar Ali Laudza, Julius Tegar Aji Putra, M. Daffa' Atstsaqif, Irvino Kent Setiawan, and Muhamad Sahal Annabil, "Sentiment Analysis and Topic Modelling of Tourist Reviews on Heritage Destinations of Lawang Sewu," *International Journal of Advances in Scientific Research and Engineering (IJASRE), ISSN:2454-8006, DOI: 10.31695/IJASRE*, vol. 11, no. 9, pp. 38–48, Sep. 2025, doi: 10.31695/IJASRE.2025.9.3.

[2] P. Zhang, J. Wang, and R. Li, "Tourism-type ontology framework for tourism-type classification, naming, and knowledge organization," *Heliyon*, vol. 9, no. 4, p. e15192, Apr. 2023, doi: 10.1016/J.HELIYON.2023.E15192.

[3] S. Karmila, V. Intan Ardianti Prodi Telematika Energi, and V. Intan Ardianti, "Metode Latent Dirichlet Allocation Untuk Menentukan Topik Teks Suatu Berita," *Jurnal Informatika dan Komputasi: Media Bahasan, Analisa dan Aplikasi*, vol. 16, no. 01, pp. 36–44, Nov. 2022, doi: 10.56956/JIKI.V16I01.100.

[4] K. Bhowmick and V. Sarvaiya, "A Comparative Study of The Different Classification Algorithms on Football Analytics," *Int. J. Adv. Res. (Indore).*, vol. 9, no. 08, pp. 392–407, Aug. 2021, doi: 10.21474/IJAR01/13280.

[5] Y. Agrawal *et al.*, "Presbyvestibulopathy: Diagnostic criteria Consensus document of the classification committee of the Bárány Society," *J. Vestib. Res.*, vol. 29, no. 4, pp. 161–170, 2019, doi: 10.3233/VES-190672.

[6] Y. Herawati, M. Sa'diyah, A. A. Nugroho, and D. Altin, "The Impact of Tourism Sector Development on the Socio- Economic Conditions of the Community in Supporting Quality Tourism in South Bangka Regency," *International Journal of Research and Innovation in Social Science*, vol. 9, no. 28, pp. 131–136, Dec. 2025, doi: 10.47772/IJRISS.2025.92800014.

[7] S. A. Karunia, "Online News Classification Using Naïve Bayes Classifier With Mutual Information for Feature Selection," 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:86589834

[8] 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.

[9] M. Shi, K. Wang, and C. Li, "A C-LSTM with word embedding model for news text classification," *Proceedings - 18th IEEE/ACIS International Conference on Computer and*

*Information Science, ICIS 2019*, pp. 253–257, Jun. 2019, doi: 10.1109/ICIS46139.2019.8940289.

[10] J. Zhou, Z. Ye, S. Zhang, Z. Geng, N. Han, and T. Yang, "Investigating response behavior through TF-IDF and Word2vec text analysis: A case study of PISA 2012 problem-solving process data," *Heliyon*, vol. 10, no. 16, Aug. 2024, doi: 10.1016/j.heliyon.2024.e35945.

[11] J. Amalia, J. Pakpahan, M. Pakpahan, Y. Panjaitan, F. Informatika dan Teknik Elektro, and I. Teknologi Del, "Model Klasifikasi Berita Palsu Menggunakan Bidirectional LSTM dan Word2vec sebagai Vektorisasi," *JATISI*, vol. 9, no. 4, pp. 3319–3331, Dec. 2022, doi: 10.35957/JATISI.V9I4.1332.

[12] R. Alroobaea, "An Empirical Deep Learning Approach for Arabic News Classification," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, pp. 1048–1055, Jun. 2023, doi: 10.14569/IJACSA.2023.01406112.

[13] E. M. Sirait, R. Silalahi, A. A. Tambunan, and J. Amalia, "News Classification Using Bidirectional Long Short Term Memory and GloVe," *Sinkron*, vol. 9, no. 1, pp. 112–124, Jan. 2024, doi: 10.33395/SINKRON.V9I1.13005.

[14] J. Li, Y. Xu, and H. Shi, "Bidirectional LSTM with Hierarchical Attention for Text Classification," *Proceedings of 2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference, IAEAC 2019*, pp. 456–459, Dec. 2019, doi: 10.1109/IAEAC47372.2019.8997969.

[15] D. Sultan, M. Mendes, A. Kassenkhan, and O. Akylbekov, "Hybrid CNN-LSTM Network for Cyberbullying Detection on Social Networks using Textual Contents," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 9, pp. 748–756, 2023, doi: 10.14569/IJACSA.2023.0140978.

[16] Y. Widhiyasana, T. Semiawan, I. Gibran, A. Mudzakir, and M. R. Noor, "Penerapan Convolutional Long Short-Term Memory untuk Klasifikasi Teks Berita Bahasa Indonesia," *Jurnal Nasional Teknik Elektro dan Teknologi Informasi*, vol. 10, no. 4, pp. 354–361, Nov. 2021, doi: 10.22146/JNTETI.V10I4.2438.

[17] M. R. F. Kamarula and N. Rochmawati, “Perbandingan CNN dan Bi-LSTM pada Analisis Sentimen dan Emosi Masyarakat Indonesia Di Media Sosial Twitter Selama Pandemik Covid-19 yang Menggunakan Metode Word2vec,” *Jurnal Informatika dan Komputasi Sains (JINACS)*, vol. 4, no. 2, pp. 219–228, 2022, doi: 10.26740/jinacs.v4n02.p219-228.

This page intentionally left blank