Baghdad Science Journal

Volume 22 | Issue 7 Article 29

7-25-2025

Performance of CNN and LSTM Model on COVID-19 News Headline Data Classification

Dian Kurniasari
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia, dian.kurniasari@fmipa.unila.ac.id

Luthfia Nur Azizah
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia, luthfiaanurazizah@gmail.com

Purnomo Husnul Khotimah
National Research and Innovation Agency of the Republic of Indonesia, Bandung, Indonesia, purn005@brin.go.id

Warsono
Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia, warsono.1963@fmipa.unila.ac.id

How to Cite this Article

Kurniasari, Dian; Azizah, Luthfia Nur; Khotimah, Purnomo Husnul; and Warsono (2025) "Performance of CNN and LSTM Model on COVID-19 News Headline Data Classification," *Baghdad Science Journal*: Vol. 22: Iss. 7, Article 29.
DOI: https://doi.org/10.21123/2411-7986.5009

BAGHDAD SCIENCE JOURNAL 2025;22(7):2458–2468

RESEARCH ARTICLE

Performance of CNN and LSTM Model on COVID-19 News Headline Data Classification

Dian Kurniasari [1,*], **Luthfia Nur Azizah** [1], **Purnomo Husnul Khotimah** [2], **Warsono** [1]

[1] Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Lampung, Lampung, Indonesia
[2] National Research and Innovation Agency of the Republic of Indonesia, Bandung, Indonesia

ABSTRACT

During the COVID-19 pandemic, mass media, especially online news portals, have been essential in disseminating health information and governmental policies, serving as the primary reference for the general public. Unfortunately, not all news articles are relevant to COVID-19 case monitoring. Some sources provide information that is less useful for tracking the pandemic's progression. Thus, it is crucial to develop a methodology that allows news articles to effectively aid stakeholders in monitoring COVID-19 developments. This study proposes using Deep Learning (DL) models to classify news headlines for this purpose. The aim is to identify suitable and reliable DL models for classifying Indonesian-language news headlines related to COVID-19 by comparing two popular models: Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) under various data imbalance scenarios. To improve model performance and reduce overfitting during training, hyperparameter tuning is applied to parameters such as epochs, batch size, dropouts, and LSTM units. Furthermore, the model uses the Count-Vectorizer approach for word embedding with the Bag of Words (BoW) technique to effectively understand the text's vocabulary. The results indicate that the CNN model outperforms the LSTM model in terms of precision, efficiency, and reliability, especially in scenarios with imbalanced data. The CNN model proves superior across all levels of data balance when evaluating its capacity to classify imbalanced data.

Keywords: COVID-19, CNN, Imbalanced data, LSTM, News headline data classification

Introduction

COVID-19, caused by the highly transmissible SARS-CoV-2 virus, swiftly spread worldwide after its initial detection in Wuhan, Hubei Province, China, in late December 2019. Given its severity, the World Health Organization (WHO) declared it a global pandemic on March 11, 2020.[1] SARS-CoV-2 has continued to evolve, resulting in various strains with differing levels of infectivity and mortality. This situation has raised concerns among government bodies, health institutions, and the general public. Consequently, these stakeholders must understand the current state of COVID-19 to effectively disseminate necessary information and develop strategies to combat this global health crisis.

Amid the pandemic, online media platforms, particularly news portals, play a crucial role in distributing health information and governmental policies, serving as the primary source of information for the general public.[2,3] That is due to their ability to deliver news quickly, conveniently, and with frequent updates on local events. Additionally, online news portals offer substantial potential as a reliable and informal data source for monitoring the real-time dynamics of COVID-19 status.[4] Unfortunately, not all broadly categorized news reports using terms like "coronavirus", "COVID-19", and "pandemic" cover

Received 15 April 2024; revised 11 August 2024; accepted 13 August 2024.
Available online 25 July 2025

* Corresponding author.
E-mail addresses: dian.kurniasari@fmipa.unila.ac.id (D. Kurniasari), luthfiaanurazizah@gmail.com (L. N. Azizah), purn005@brin.go.id (P. H. Khotimah), warsono.1963@fmipa.unila.ac.id (Warsono).

https://doi.org/10.21123/2411-7986.5009

relevant topics comprehensively. Indeed, some sources provide COVID-19 information that lacks relevance in monitoring the changing state of the pandemic. Therefore, a methodology is needed to enable news articles to serve as a reliable point of reference for stakeholders in monitoring the evolving status of COVID-19.

A commonly used method involves categorizing news texts. However, comprehensively analysing the entirety of a news article to determine its main subject can be challenging. An alternative approach is to focus on the headline. News headlines are succinct and offer a rapid means to ascertain the primary subject matter. Khotimah et al.[5] presented evidence that dengue fever outbreaks can be detected by analyzing the titles of Indonesian-language news articles from online news portals.

Text classification proves highly effective in categorizing text based on its content.[6,7] However, this process often faces challenges due to imbalanced class distributions, where one class significantly outweighs another. This imbalanced dataset condition complicates information extraction and introduces bias into classification results.[8] The severity of this issue hinges on the imbalance ratio of data used in training, exacerbating bias against the majority class as the level of imbalance intensifies.[9]

The problem of data imbalance has been a prominent focus in classical Machine Learning (ML) research for more than a decade.[10] Handling imbalanced data poses challenges in traditional data science and ML, especially with large datasets. Deep Learning (DL) models have effectively addressed this issue.[11] These models allow computational systems to learn data representations through multiple processing layers, capturing varying levels of abstraction. Moreover, the non-linear nature of DL models and their ability to seamlessly incorporate word embeddings often result in higher classification accuracy than traditional linear classifiers. Consequently, DL models have demonstrated exceptional performance across various text classification tasks, such as sentiment analysis, topic modelling, and question answering.[12]

Various DL model architectures, such as Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN), are categorized based on their neural network structure. Researchers widely acknowledge RNN models as prominent architectures for solving text classification challenges.[13] Despite the vulnerability of basic RNNs to vanishing gradient issues, several improved versions have been developed to mitigate these drawbacks. One notable example is the Long Short-Term Memory (LSTM) model. The utilization of CNNs for text categorization was pioneered by Collobert et al.,[14] Kalchbrenner et al.,[15] and Kim.[16] This approach has gained considerable traction in subsequent studies, leading to a significant proliferation of CNN-based networks in recent literature.[17–19]

Several research studies have shown that LSTM and CNN models achieve higher accuracy by incorporating word embedding into their layers, particularly when handling complex and uneven textual data. For instance, Sharma et al.[20] conducted a study using the LSTM model to detect and categorize fake news. Their approach involved using the GloVe word embedding model to create an embedding matrix for the dataset, which was then fed into the LSTM layer. The proposed model achieved an accuracy rate of 84.1% in learning outcomes.

In their research, Khotimah et al.[5] employed several DL models, including the Multi-Layer Perceptron (MLP), LSTM, and CNN, to detect dengue occurrences using data from internet news headlines. The dataset they utilized exhibited a class imbalance ratio of 1:2. Word embedding served as an input feature in these DL models. The CNN model demonstrated the highest accuracy among the investigated models, achieving 88.69%. Bhuiyan et al.[21] introduced the LSTM model for categorizing Bengali news headlines. Their dataset comprised 4500 news titles categorized into four distinct types, collected from multiple newspapers using web scraping techniques. The text underwent multiple pre-processing stages to optimize the training process, followed by the application of word embedding. Their classification of news titles achieved a high degree of accuracy.

Dogru et al.[22] classified the TTC-3600 dataset containing Turkish news texts and the BBC News datasets containing English news texts. They utilized the Doc2Vec word representation method to propose multiple distinct models, including CNN-based models, Gauss Naïve Bayes, Random Forest, Naïve Bayes, and Support Vector Machine. The CNN-based model achieved an impressive accuracy of 94.17% on the Turkish dataset and 96.41% on the UK dataset. Khuntia et al.[23] employed various word embedding techniques and DL models to enhance the accuracy of their model in categorizing a news headline dataset acquired from Kaggle. They utilized Word2Vec and GloVe for word embedding and employed LSTM and Bidirectional LSTM (Bi-LSTM) for DL modelling. The results indicated that the GloVe with LSTM model was most effective for classifying the news headline dataset.

Previous research has not extensively explored the statistical analysis of the effectiveness of DL methodologies and appropriate architectures in addressing

class imbalances. Researchers have identified a gap in comprehensive research on DL due to imbalanced data.[24–26] Therefore, this study seeks to determine the most effective DL model for accurately classifying imbalanced textual data. The research introduces two advanced DL architectures, specifically CNN and LSTM, for classifying Indonesian-language online news headlines related to the COVID-19 pandemic. Additionally, word embedding techniques are employed to enhance the model's efficiency. The performance of the classification models was evaluated and compared using key metrics such as accuracy, precision, recall, and f1-score. This evaluation encompassed various class imbalance ratios: 37%, 30%, 20%, 10%, and 1%.

Although the primary focus of this study is on analyzing COVID-19 data, the optimized models derived from this research apply to similar tasks involving the classification of online news related to different diseases or pandemics. This capability can assist stakeholders in formulating targeted strategies to address specific health crises effectively.

Materials and methods

Methodology

This research aims to determine the most effective DL model for classifying Indonesian-language news headlines about COVID-19, considering different class imbalance levels. Python software will evaluate each classifier's performance, comprehensively comparing accuracy, precision, recall, and f1-score metrics. The research design, illustrating the various phases of this research, is presented in Fig. 1 and detailed further in subsequent sections.

Input data

The first phase of this study involves importing data from Excel files into Python using the Pandas module. The research data is stored on Google Drive and accessed through Python scripts. This dataset is the property of the Information Retrieval Research Group at the BRIN Informatics Research Center and has been appropriately categorized. The dataset

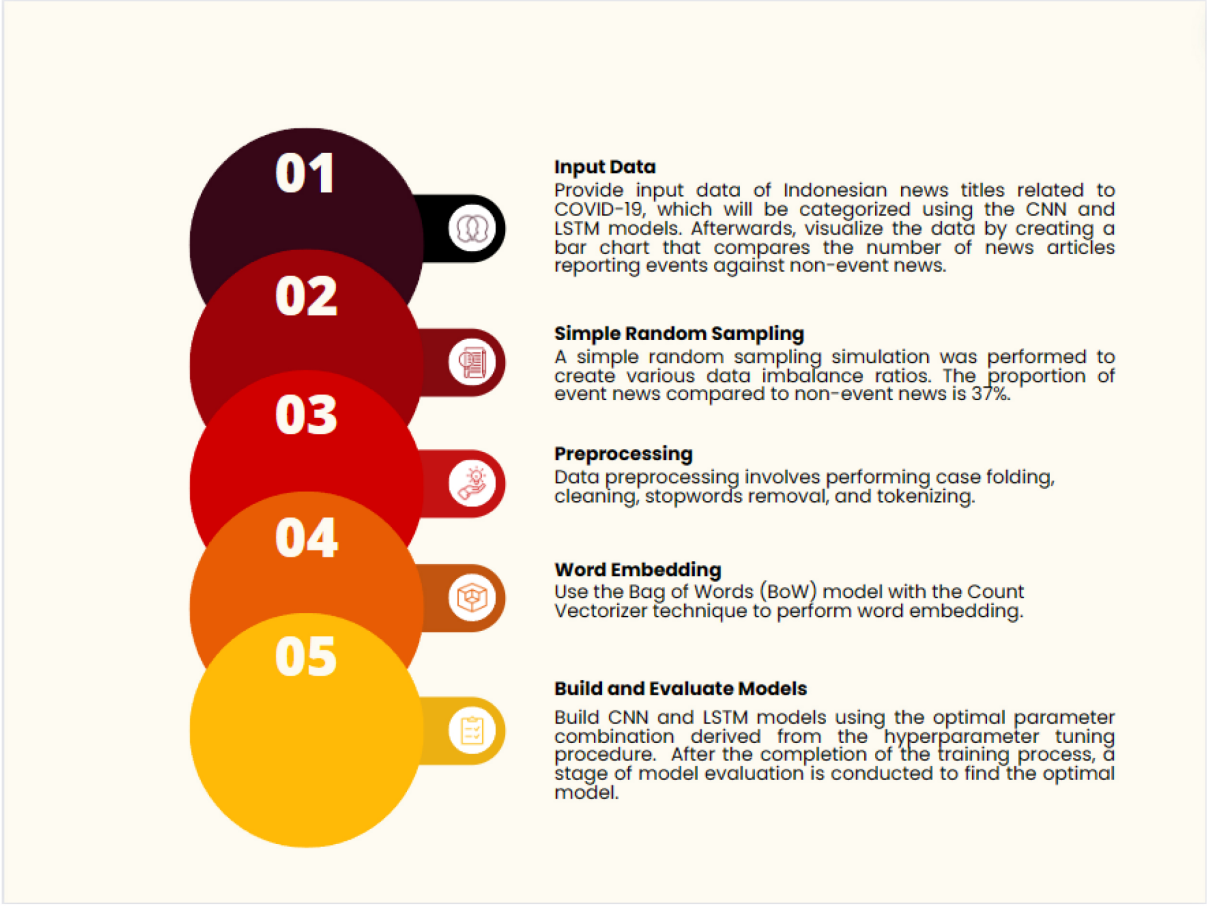

Fig. 1. Research design.

Table 1. COVID-19 news heafigdline data sample.

Id	News Portal	Publish time	News headline	News category
778	detik	2020-01-24T11:11:00.000Z	*Heboh Virus Corona yang Mirip Film 'Contagion'* Eng. title: The excitement of the coronavirus is similar to the film 'Contagion'	0
2283	antara	2020-01-24T11:14:00.000Z	*Seorang pasien "suspect" virus corona di Jakarta diisolasi* Eng. title: A "suspected" corona virus patient in Jakarta is isolated	1
4449	merdeka	2020-01-24T12:00:00.000Z	*Garuda Indonesia Siap Siaga Cegah Virus Corona Masuk Indonesia* Eng.title: Garuda Indonesia is Ready to Prevent the Corona Virus from Entering Indonesia	0
6272	tempo	2020-01-24T14:03:00.000Z	*Virus Corona Misterius: 20 Juta Warga Wuhan dan Hubei Diisolasi* Eng. title: Mysterious Corona Virus: 20 Million Wuhan and Hubei Residents Isolated	1
3614	detik	2020-01-25T05:05:00.000Z	*Sedih, Perayaan Imlek di China Dibatalkan akibat Wabah Corona* Eng. title: Sad, Chinese New Year celebrations in China have been cancelled due to the Corona outbreak	0
...	...	...	...	...
7571	kompas	2020-01-25T05:05:00.000Z	*Virus Corona Menyebar Lewat Droplet, Kenapa Kita Perlu Cuci Tangan?* Eng.title: Corona Virus Spreads Through Droplets, Why Do We Need to Wash Our Hands?	0
8112	tempo	2020-01-25T07:01:00.000Z	*Bank Cina Beri Pinhouran Rp 4 triliun untuk Atasi Virus Corona* Eng.title: Chinese Bank Gives Pinhouran IDR 4 trillion to Overcome Corona Virus	0
5550	antara	2020-01-26T04:53:00.000Z	*Data baru virus corona: 56 orang meninggal, 2.000 tertular* Eng.title: New coronavirus data: 56 people died, 2,000 infected	1

comprises Indonesian-language news headlines concerning COVID-19, collected between January 2020 and May 2020, totalling 16,863 entries. News directly related to COVID-19 events is labelled "1," while news containing COVID-19 information but not directly related to events is labelled "0." Table 1 presents the data used in this research.

The ratio of event-related news to non-event newsstands is approximately 1:3, with 4,547 event news and 12,289 instances of non-event news. That indicates a noticeable imbalance in the research data. Moreover, this study will selectively consider factors deemed to be significant. Therefore, the data selection process focused on two variables—News Title and News Category—that most closely align with the research objectives.

Simple random sampling

The next phase involves applying random sampling to achieve a balanced data distribution. Simple random sampling represents the most fundamental method of selecting a sample, where each member has an equal chance of being chosen. Therefore, the quality of the sample remains unaffected by bias. [27]

Table 2. Data imbalance level composition.

	Number of News		
Sample size	*Event*	*Non-event*	Total data
37%	4.547	12.289	16.836 data
30%	3.686	12.289	15.975 data
20%	2.458	12.289	14.747 data
10%	1.228	12.289	13.517 data
1%	123	12.295	12.412 data

The data collected consists of news headlines categorized by events. Five sampling scenarios were utilized to maintain consistent data volumes across non-event-related categories, varying in proportions: 30%, 20%, 10%, and 1%. The specific configuration of these sampling scenarios is outlined in Table 2.

Pre-processing

Kerner et al.[28] discovered that pre-processing techniques on specific datasets can significantly enhance data quality, especially in text categorization. Their research utilized several methods for preparing text data, which include:

a. Lowercasing: Converting uppercase letters to lowercase.
b. Cleaning: Removing irrelevant elements from the data, such as misspellings, non-essential characters, punctuation marks, and numerical values.
c. Stopword Removal: Eliminating stopwords—words considered non-informative, such as pronouns, prepositions, time adverbs, and conjunctions.
d. Tokenization: Breaking down sentences from each headline into individual units (words or phrases) called tokens.

Word embedding

Word embedding is a method used to convert words into low-dimensional vectors.[29] The Bag of Words (BoW) model represents words using the count-vectorizer algorithm. This algorithm transforms the BoW into vector form. Initially, the document's sentences are extracted, creating a unique vocabulary. The frequency of each word in each document is then calculated. As described by Khomsah and Aribowo,[30] each document is represented by a vector of the same size as the vocabulary. The components of this vector correspond to the word counts within that document.

Convolutional neural network (CNN)

A CNN is a DL model that integrates the convolution layer as a foundational component within its neural network architecture. The convolution layer operates as a sparse matrix with reduced dimensions compared to the input it processes. This characteristic enhances computational efficiency, distinguishing the CNN model's capability to extract essential information from textual data. Moreover, the convolution layer minimizes computational overhead. The CNN model typically comprises an input layer, a pooling layer, a fully connected layer, an output layer, and the convolution layer.[31,32]

Long-short term memory (LSTM)

LSTM was first introduced by Hochreiter and Schmidhuber in 1997 to tackle the challenge of capturing long-term dependencies in RNN. LSTM has garnered significant attention as a neural network architecture for text classification.[31] Fundamentally, the LSTM architecture resembles RNN, but it distinguishes itself through the design of its recurrent module, which includes specialized gates. These gates—namely, the forget gate, input gate, and output gate—are strategically incorporated to manage the flow of information more effectively, thereby mitigating issues related to long-term dependencies.[33]

Table 3. Model evaluation metrics.

Evaluation metrics	Formulation
Accuracy	$\frac{TP+TN}{TP+FP+TN+FN}$
Precision	$\frac{TP}{TP+FP}$
Recall	$\frac{TP}{TP+FN}$
F1-Score	$\frac{2\times recall\times precision}{(recall+precision)}$

Model evaluation

The model evaluation aims to quantitatively assess the classification model's performance using evaluation metrics such as accuracy, precision, recall, and f1-score. Accuracy is commonly used as a standard measure to gauge the efficacy of classification models. However, the f1-score is a more robust evaluation metric for assessing classification performance, especially in models addressing imbalanced datasets.[5] The mathematical formulations of these evaluation metrics are presented in Table 3.

Results and discussion

This section presents a comprehensive overview of the results obtained during each phase of the classification process using two separate DL models: the CNN and the LSTM model.

Simple random sampling

Before pre-processing, the data undergoes initial random sampling using five distinct scenarios, as detailed in the methodology section. As the sample size decreases, the imbalance in data increases. The data selected through a simple random sampling technique will then be used to compare the performance of CNN and LSTM models in classification. Table 4 presents the word counts obtained from this primary random sampling method with a predefined sample size.

Pre-processing

Data pre-processing is conducted to prepare the data for DL algorithms. This pre-processing includes four stages: converting to lowercase, cleaning,

Table 4. Word count of simple random sampling results.

Sample size	Number of words
37%	11,169 word
30%	11,019 word
20%	10,840 word
10%	10,597 word
1%	10,320 word

removing stopwords, and tokenization. An illustration of pre-processed data is presented in Table 5.

After completing all stages of data pre-processing, the next step involves padding to standardize the length of each news headline for input into the embedding layer. However, accurately determining the maximum headline length is crucial to avoid introducing unnecessary noise or losing information due to inaccuracies in this measurement.

The padding procedure sets a maximum length of 20 words. Hence, any headline shorter than 20 words will be padded with zeros to ensure uniform length across all inputs to the embedding layer.

Word embedding

The headline text data has been converted into vector representations referred to as feature vectors. A BoW model using the Count Vectorizer technique was employed for word embedding. The BoW model constructs an index of unique terms from the entire text (news headlines) arranged alphabetically, thus capturing the vocabulary used in the content. Each document's feature vector is generated by counting the occurrences of each word within that specific document. The resulting numerical features are represented as vectors, as detailed in Table 6.

Classification process with CNN and LSTM models

Selecting the appropriate model is crucial for developing an effective classification system for imbalanced textual data. Therefore, a comparison was conducted between the CNN and LSTM models to identify this study's most suitable approach for addressing data imbalances.

In this research, the CNN architecture includes two one-dimensional convolutional layers to enhance model performance, as illustrated in Fig. 2. A one-dimensional max pooling layer with a pool size of 3 is employed to select the highest values from feature maps within each pooling window. These pooling layers help reduce feature map size and alleviate overfitting issues commonly encountered in CNNs. Additionally, batch normalization and dropout techniques are applied to standardize output from preceding layers and prevent overfitting. Furthermore, the fully connected layer incorporates a Rectified Linear Unit (ReLU) activation function with 32 nodes. The final layer, crucial for determining classification outcomes, utilizes a Sigmoid activation function suitable for binary classification tasks.

Fig. 3 presents the LSTM architecture, which contrasts with CNN's architecture. This research utilized a neural network configuration featuring two LSTM layers and two dense layers with activation functions. The objective was to mitigate gradient vanishing issues and improve the model's effectiveness. Each LSTM layer has 128 LSTM units, and the output gate employs a Sigmoid activation function. The use of Sigmoid has shown better results compared to other activation functions. The final layer consists of a single node activated by a sigmoid function, which outputs values of either 0 or 1, ideal for binary classification tasks.

Before starting the training process, adjustments will be applied to the CNN and LSTM model architectures concerning hyperparameters. The objective is to identify the optimal parameter settings that can enhance model performance and mitigate overfitting and underfitting. While hyperparameter tuning

Table 5. Text data pre-processing results.

Pre-processing stages	Results
Initial data	Data baru virus corona: 56 orang meninggal, 2.000 tertular
Lowercasing	data baru virus corona: 56 orang meninggal, 2.000 tertular
Cleaning	data baru virus corona orang meninggal tertular
Stopword removal	data virus corona orang meninggal tertular
Tokenization	69, 3, 1, 7, 13, 129

Table 6. Word embedding results sample with BoW model.

	Index tip					
News headline	aa	aaji	abad	abai	abaikan	...
aa gym ajak masyarakat muliakan jenazah pasien covid	1	0	0	0	0	...
aaji asuransi mengcover pasien virus corona ditetapkan pandemik	0	1	0	0	0	...
abaikan gejala strok pandemi covid	0	0	0	0	1	...

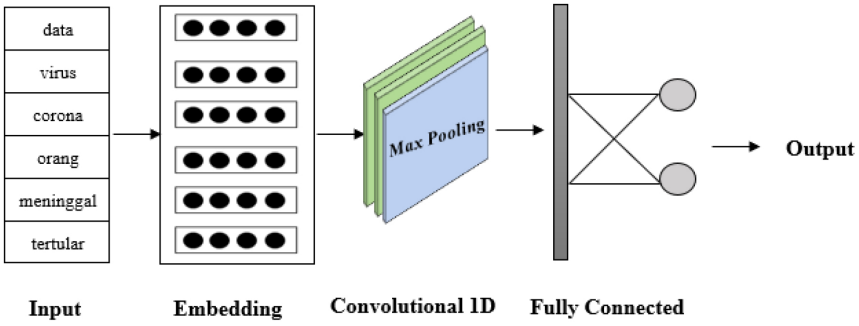

Fig. 2. CNN architecture.

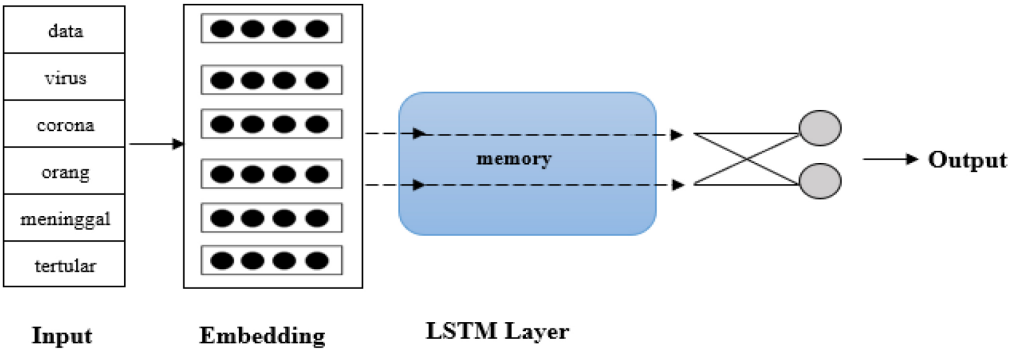

Fig. 3. LSTM architecture.

Table 7. Parameters for hyperparameter tuning.

Parameter	Parameter values
Epochs	50, 100, 200
Batch size	32, 64, 128
Dropout	0.2, 0.3, 0.4
LSTM unit	32, 64, 128

through trial and error is straightforward, it is inefficient due to the increasing complexity and time required to find the best parameter combination. Therefore, this study utilized Grid Search CV for hyperparameter optimization, systematically evaluating all possible parameter combinations and employing k-fold cross-validation to ensure accurate estimation. Detailed parameters involved in the hyperparameter tuning process are outlined in Table 7.

The results from the hyperparameter optimization process using Grid Search CV for CNN and LSTM models are detailed in Tables 8 and 9, respectively.

After determining the optimal parameter estimates, the next step involves training the models using CNN and LSTM architectures. However, it is crucial to validate the data through k-fold cross-validation during training to minimize errors or inaccuracies arising from hyperparameter tuning. K-fold cross-validation enhances the training process by utilizing all available training and test data. Increasing the value of k enhances the accuracy of the validation model. In this study, a value of k equal to 10 was employed.

The model evaluation included accuracy, precision, recall, and f1-score, calculated as the average classification report values across folds 1 to 10. The

Table 8. CNN model best parameters.

Data 37%	Data 30%	Data 20%	Data 10%	Data 1%	
K-fold	10 fold	10 fold	10 fold	10 fold	10 fold
Epochs	200	50	200	50	100
Batch size	32	128	64	64	32
Dropout	0.4	0.2	0.3	0.4	0.3
Layer	**Conv.** : 2 layer (64 and 64) **Hidden**: 2 layers (32 and 32)	**Conv.** : 2 layer (64 and 32) **Hidden**: 2 layers (32 and 32)	**Conv.** : 2 layer (64 and 64) **Hidden**: 2 layers (64 and 32)	**Conv.** : 2 layer (128 and 128) **Hidden**: 2 layers (64 and 32)	**Conv.** : 2 layer (128 and 128) **Hidden**: 2 layers (32 and 32)
Time	9 hour	7.5 hour	7.5 hour	9.5 hour	9.5 hour

Table 9. LSTM model best parameters.

Parameter	Data 37%	Data 30%	Data 20%	Data 10%	Data 1%
K-fold	10 fold	10 fold	10 fold	10 fold	10 fold
Epochs	100	200	200	200	100
Batch size	64	32	32	32	32
Dropout	0.4	0.4	0.2	0.4	0.2
LSTM unit	128 dan 128	128 dan 64	64 dan 32	128 dan 128	32 dan 32
Layer	**Hidden**: 2 layers (64 and 64)	**Hidden**: 2 layers (64 and 64	**Hidden**: 2 layers (32 and 32)	**Hidden**: 2 layers (128 and 64)	**Hidden**: 2 layers (32 and 32)
Time	48 hour	92 hour	40 hour	24 hour	4.5 hour

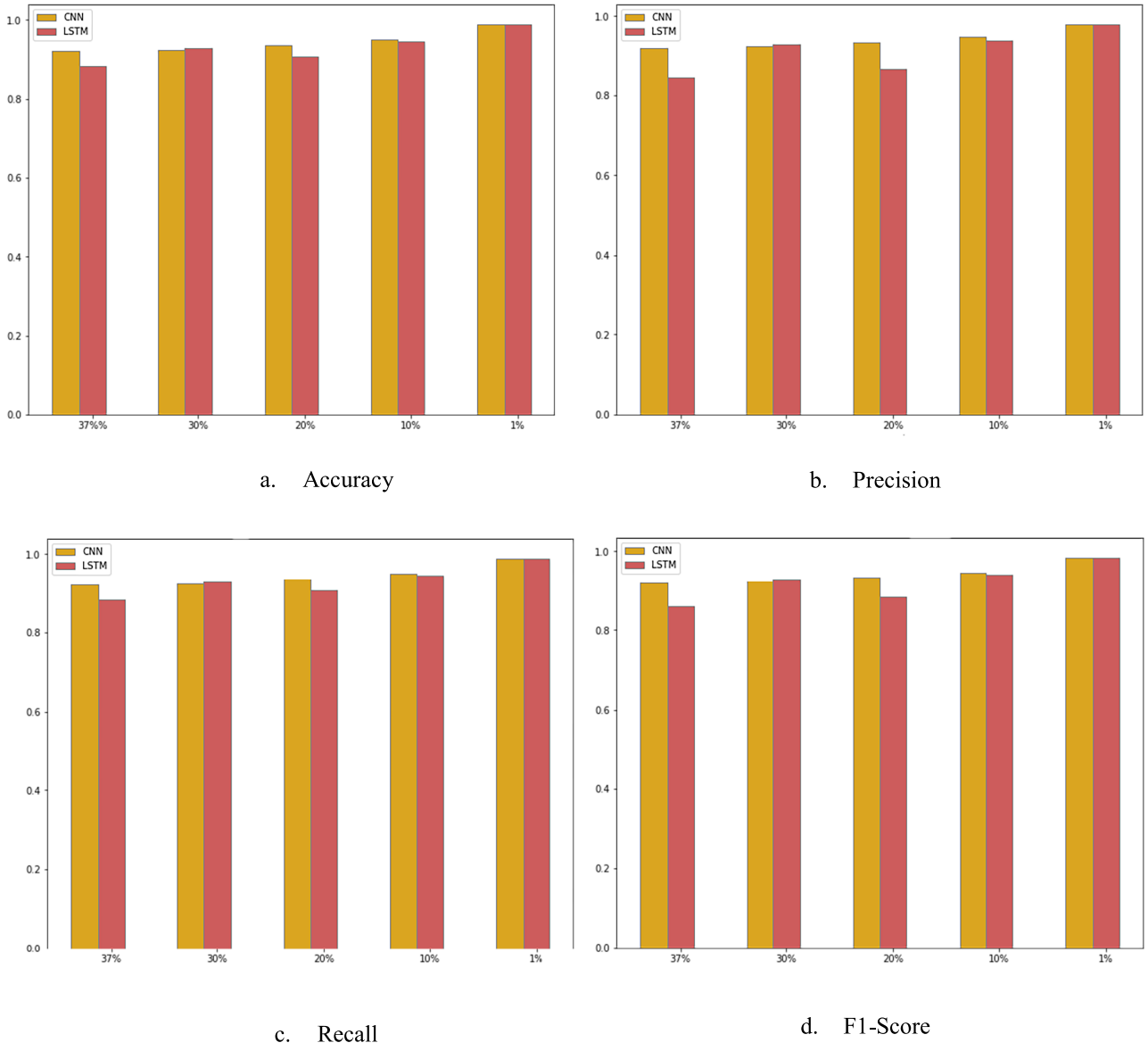

Fig. 4. Comparison of evaluation metrics on CNN and LSTM models.

evaluation and comparative results of the models are illustrated in Fig. 4.

Fig. 4 presents the evaluation metrics across all five data scenarios. The x-axis denotes the degree of data imbalance, while the y-axis represents the evaluation metrics—accuracy, precision, recall, or f1-score. As depicted in Fig. 4, it is apparent that both proposed models perform competitively. This

competitiveness arises from the distinct strengths and weaknesses of CNN and LSTM. Specifically, CNN models leverage their convolutional layers to extract significant information from news text, including phrases and relationships between adjacent words in sentences.

Furthermore, it reduces computational load due to its smaller dimensions than processed data. Overfitting issues commonly observed in CNN models were effectively mitigated by implementing techniques such as dropouts, early stopping, and fine-tuning hyperparameters. The LSTM model offers several advantages, including retaining relevant information over longer sequences while filtering out irrelevant details and recognizing connections between contexts in textual materials. However, the LSTM model is less efficient due to its significantly longer processing time than the CNN model. Across all evaluation criteria at data imbalance levels of 37%, 30%, and 20%, the CNN model consistently outperformed the LSTM model. At a data imbalance level of 1%, however, the LSTM model demonstrated superior performance over the CNN model. Therefore, it can be concluded that the CNN model exhibits higher accuracy, efficiency, and robustness when classifying textual data, especially in scenarios involving imbalanced data.

Conclusion

Based on the findings and discussion presented earlier, it can be concluded that the CNN model is the appropriate deep-learning model for classifying Indonesian COVID-19 news headlines. Through the training and evaluation conducted, the CNN model has shown its superiority over the LSTM model in handling imbalanced data across all levels of data distribution. However, the CNN model often overlooks the contextual interdependencies within the text. Conversely, the LSTM model excels in capturing relationships between different contexts. Therefore, the author suggests integrating both CNN and LSTM models to enhance future research efforts and achieve optimal performance. This approach harnesses the strengths of both models simultaneously.

Acknowledgement

This study is supported by Riset dan Inovasi untuk Indonesia Maju (RIIM) Batch 4 of National Research and Innovation Agency (BRIN) and Indonesia Endowment Fund for Education Agency (LPDP); Contract Number: B-3836/II.7.5/FR.06.00/11/2023.

Authors' declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for republication, which is attached to the manuscript.
- Authors sign on ethical consideration's approval.
- No animal studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Lampung, Lampung, Indonesia.

Authors' contribution statement

All authors collaborated on this study, contributing to various aspects such as ideation, methodology, model development, validation, formal analysis, resource acquisition, data curation, and drafting of the original manuscript.

References

1. Sharma A, Ahmad FI, Lal SK. Covid-19: A review on the novel coronavirus disease evolution, transmission, detection, control and prevention. Viruses. 2021 Jan;13(2):1–25. https://doi.org/10.3390/v13020202.
2. Mach KJ, Salas RR, Pentz B, Taylor J, Costa CA, Cruz SG, *et al.* News media coverage of COVID-19 public health and policy information. Humanit Soc Sci Commun. 2021 Sept;8(1):1–11. https://doi.org/10.1057/s41599-021-00900-z.
3. Wahyudi MDR, Fatwanto A, Kiftiyani U, Galih Wonoseto M. Topic modeling of online media news titles during COVID-19 emergency response in indonesia using the latent dirichlet allocation (LDA) algorithm. Telematika. 2021 August;14(2):101–111. https://doi.org/10.35671/telematika.v14i2.1225.
4. Nugraheni E, Khotimah PH, Arisal A, Rozie AF, Riswantini D, Purwarianti A. Classifying aggravation status of COVID-19 event from short-text using CNN. In Proceeding 2020 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications. IEEE. 2020 Dec;240–5. https://doi.org/10.1109/ICRAMET51080.2020.9298674.
5. Khotimah PH, Fachrur Rozie A, Nugraheni E, Arisal A, Suwarningsih W, Purwarianti A. Deep learning for dengue fever event detection using online news. In Proceeding 2020 International Conference on Radar, Antenna, Microwave, Electronics and Telecommunications. IEEE. 2020 Dec;261–6. https://doi.org/10.1109/ICRAMET51080.2020.9298630.
6. Alqahtani A, Ullah KH, Alsubai S, Sha M, Almadhor A, Iqbal T, *et al.* An efficient approach for textual data classification using deep learning. Front Comput Neurosci. 2022 Sept;16:1–9. https://doi.org/10.3389/fncom.2022.992296.
7. Zhu H, Lei L. The research trends of text classification studies (2000–2020): A Bibliometric Analysis. SAGE Open. 2022 April;12(2):1–16. https://doi.org/10.1177/21582440221089963.

8. Kulkarni A, Chong D, Batarseh FA. Data democracy: At the nexus of artificial intelligence, software development, and knowledge engineering. Academic Press; 2020. Chap 5, Foundations of data imbalance and solutions for a data democracy;83–106. https://doi.org/10.1016/B978-0-12-818366-3.00005-8.
9. Liu H, Zhou M, Liu Q. An embedded feature selection method for imbalanced data classification. IEEE/CAA J Autom Sin. 2019 May;6(3):703–15. https://doi.org/10.1109/JAS.2019.1911447.
10. Johnson JM, Khoshgoftaar TM. Survey on deep learning with class imbalance. J Big Data. 2019 March;6(1):1–54. https://doi.org/10.1186/s40537-019-0192-5.
11. Bathla G, Aggarwal H, Rani R. Deep learning for big data analytics. (Eds). Advances in Computing and Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. 2020 Jan;391–399. https://doi.org/10.1007/978-981-15-0222-4_36.
12. Akpatsa SK, Li X, Lei H. A survey and future perspectives of hybrid deep learning models for text classification. Artificial intelligence and security. In ICAIS 2021. Lecture Notes in Computer Science. Springer, Cham.x 2021 July;358–69. https://doi.org/10.1007/978-3-030-78609-0_31.
13. Lan Y, Hao Y, Xia K, Qian B, Li C. Stacked residual recurrent neural networks with cross-layer attention for text classification. IEEE Access. 2020 April;8:70401–70410. https://doi.org/10.1109/ACCESS.2020.2987101.
14. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, Kuksa P. Natural language processing (almost) from scratch. J Mach Learn Res. 2011 Nov;12:2493–537. https://dl.acm.org/doi/10.5555/1953048.2078186.
15. Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics. 2014 June;655–65. https://doi.org/10.3115/v1/p14-1062.
16. Kim Y. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics. 2014 Oct;1746–51. https://doi.org/10.3115/v1/d14-1181.
17. Zheng J, Zheng L. A hybrid bidirectional recurrent convolutional neural network attention-based model for text classification. IEEE Access. 2019 August;7:106673–85. https://doi.org/10.1109/ACCESS.2019.2932619.
18. Ce P, Tie B. An analysis method for interpretability of CNN text classification model. Futur Internet. 2020 Dec;12(12):1–14. https://doi.org/10.3390/fi12120228.
19. Zhao W, Zhu L, Wang M, Zhang X, Zhang J. WTL-CNN: A news text classification method of convolutional neural network based on weighted word embedding. Connect Sci. 2022 Aug;34(1):2291–312. https://doi.org/10.1080/09540091.2022.2117274.
20. Sharma R, Agarwal V, Sharma S, Arya MS. An LSTM-based fake news detection system using word embeddings-based feature extraction. Lect Notes Netw Syst. Springer, Singapore. 2021 Dec;247–55. https://doi.org/10.1007/978-981-15-8354-4_26.
21. Bhuiyan MR, Keya M, Masum AKM, Hossain SA, Abujar S. An approach for Bengali news headline classification using LSTM. Adv Intell Syst. Comput Springer, Singapore. 2021 June;299–308. https://doi.org/10.1007/978-981-15-9927-9_30.
22. Dogru HB, Tilki S, Jamil A, Ali Hameed A. Deep learning-based classification of news texts using Doc2Vec model. In 1st International Conference on Artificial Intelligence and Data Analytics. IEEE. 2021 May;91–6. https://doi.org/10.1109/CAIDA51941.2021.9425290.
23. Khuntia M, Gupta D. Indian news headlines classification using word embedding techniques and LSTM model. Procedia Comput Sci. 2023;218:899–907. https://doi.org/10.1016/j.procs.2023.01.070.
24. Basha SJ, Madala SR, Vivek K, Kumar ES, Ammannamma T. A review on imbalanced data classification techniques. In International Conference on Advanced Computing Technologies and Applications. 2022 Apr;1–6. https://doi.org/10.1109/ICACTA54488.2022.9753392.
25. Fu S, Su D, Li S, Sun S, Tian Y. Linear-exponential loss incorporated deep learning for imbalanced classification. ISA Trans. 2023 Sept;140:279–92. https://doi.org/10.1016/j.isatra.2023.06.016.
26. Chen W, Yang K, Yu Z, Shi Y, Chen CLP. A survey on imbalanced learning: latest research, applications and future directions. Artif Intell Rev. 2024 May 1;57(137):1–51. https://doi.org/10.1007/s10462-024-10759-6.
27. Bhardwaj P. Types of sampling in research. J Pract Cardiovasc Sci. 2019 Jan;5(3):157–163. https://doi.org/10.4103/jpcs.jpcs_62_19.
28. HaCohen-Kerner Y, Miller D, Yigal Y. The influence of pre-processing on text classification using a bag-of-words representation. PLoS One. 2020 May;15(5):1–22. https://doi.org/10.1371/journal.pone.0232525.
29. Nurdin A, Anggo Seno Aji B, Bustamin A, Abidin Z. Comparative analysis of the performance of word embedding models Word2Vec, Glove, and FastText in text classification (Perbandingan kinerja word embedding Word2Vec, Glove, dan FastText pada klasifikasi teks). J Tekno Kompak. 2020 Aug;14(2):74–9. https://doi.org/10.33365/jtk.v14i2.732.
30. Khomsah S, Sasmito Aribowo A. Text-preprocessing model Youtube comments in Indonesian. J Resti. (Rekayasa Sist dan Teknol Informasi). 2020 Aug;4(4):648–54. https://doi.org/10.29207/resti.v4i4.2035.
31. Wang Q, Li W, Jin Z. Review of text classification in deep learning. OALib Journal. 2021 March;8(3):1–8. https://doi.org/10.4236/oalib.1107175.
32. Widhiyasana Y, Semiawan T, Mudzakir Noor MR. Application of convolutional long short-term memory for text classification of Indonesian news (Penerapan convolutional long short-term memory untuk klasifikasi teks berita Bahasa Indonesia). J Nas Tek Elektro dan Teknol Inf. 2021 Nov;10(4):354–61. https://doi.org/10.22146/jnteti.v10i4.2438.
33. Lindemann B, Müller T, Vietz H, Jazdi N, Weyrich M. A survey on long short-term memory networks for time series prediction. Procedia CIRP. 2021;99:650–5. https://doi.org/10.1016/j.procir.2021.03.088.

أداء نموذج CNN و LSTM في تصنيف بيانات عناوين أخبار كوفيد-19

ديان كورنياساري[1]، لطفية نور عزيزة[1]، بورنومو حسنول خوتيمة[2]، وارسونو[1]

[1]قسم الرياضيات، كلية الرياضيات والعلوم الطبيعية، جامعة لامبونج، لامبونج، إندونيسيا.
[2]الوكالة الوطنية للبحث والابتكار في جمهورية إندونيسيا، باندونج، إندونيسيا.

الخلاصة

خلال جائحة كوفيد-19، كانت وسائل الإعلام، وخاصة بوابات الأخبار عبر الإنترنت، ضرورية في نشر المعلومات الصحية والسياسات الحكومية، حيث كانت بمثابة المرجع الأساسي لعامة الناس. لسوء الحظ، ليست كل المقالات الإخبارية ذات صلة بمراقبة حالات كوفيد-19. تقدم بعض المصادر معلومات أقل فائدة لتتبع تقدم الوباء. وبالتالي، من الأهمية بمكان تطوير منهجية تسمح للمقالات الإخبارية بمساعدة أصحاب المصلحة بشكل فعال في مراقبة تطورات كوفيد-19. تقترح هذه الدراسة استخدام نماذج التعلم العميق (DLلتصنيف عناوين الأخبار لهذا الغرض. والهدف هو تحديد نماذج التعلم العميق المناسبة والموثوقة لتصنيف عناوين الأخبار باللغة الإندونيسية المتعلقة بكوفيد-19 من خلال مقارنة نموذجين شائعين: الشبكة العصبية التلافيفية (CNNوالذاكرة طويلة المدى القصيرة (LSTMفي ظل سيناريوهات مختلفة لعدم توازن البيانات. لتحسين أداء النموذج وتقليل الإفراط في التجهيز أثناء التدريب، يتم تطبيق ضبط المعلمات الفائقة على معلمات مثل العصور وحجم الدفعة والانقطاعات ووحدات LSTM. علاوة على ذلك، يستخدم النموذج نهج Count-Vectorizer لتضمين الكلمات باستخدام تقنية Bag of Words (BoW) لفهم مفردات النص بشكل فعال. تشير النتائج إلى أن نموذج CNN يتفوق على نموذج LSTM من حيث الدقة والكفاءة والموثوقية، خاصة في السيناريوهات ذات البيانات غير المتوازنة. يثبت نموذج CNN تفوقه عبر جميع مستويات توازن البيانات عند تقييم قدرته على تصنيف البيانات غير المتوازنة.

الكلمات المفتاحية: COVID-19، CNN، البيانات غير المتوازنة، LSTM، تصنيف بيانات عناوين الأخبار.