The IJICS (International Journal of Informatics and Computer Science)
Vol 10 No 2, July 2026, Page 161-166
ISSN 2548-8384 (online), ISSN 2548-8449 (print)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/ijics/index
DOI 10.30865/ijics.v10i2.9899

LSTM-Based Deep Learning Approach for Hoax Detection on Indonesian Social Media

Nuraisana*, R. Mahdalena Simanjorang, Thania Rizky Ristanti, Della Triyani
Informatics Engineering, STMIK Pelita Nusantara, Medan, Indonesia
Email: 1'*nuraisana94@gmail.com, 2lenasinaga@gmail.com, 3thaniarizkyr@gmail.com, 4dellatriyani1708@gmail.com
(*: nuraisana94@gmail.com)
Submitted: **25/06/2026**; Accepted: **18/07/2026**; Published: **29/07/2026**

Abstract− The widespread adoption of social media, particularly Facebook, has drastically accelerated information dissemination while simultaneously amplifying the spread of misinformation. Hoax content poses a serious threat to public trust and social cohesion. This study proposes a hoax classification system based on the Long Short-Term Memory (LSTM) architecture, selected for its inherent ability to model sequential dependencies in textual data. The experimental workflow covered text preprocessing (case folding, tokenization, stopword removal, and stemming), word vectorization using the Keras Tokenizer, and end-to-end LSTM model training on the "Indonesian Fact and Hoax Political News" dataset from Kaggle, totaling 4,502 samples with a balanced 80:20 train-test split. Experimental results demonstrate the proposed model achieves 92.44% accuracy, 92.63% precision, 92.22% recall, and a 92.42% F1-score on the test set. These findings confirm that LSTM effectively captures contextual and sequential linguistic patterns in Indonesian-language content, offering a viable and scalable solution for automated hoax detection on social media platforms.
Keywords: Deep Learning; Hoax Detection; Indonesian Text Classification; LSTM; Social Media

1. INTRODUCTION

The rapid evolution of digital communication infrastructure has fundamentally transformed how information is created, distributed, and consumed. Social media platforms, most notably Facebook, have emerged as dominant channels for real-time content sharing, enabling individuals to broadcast information to vast audiences without geographical constraints. While this democratization of information offers significant societal benefits, it has also created an environment where unverified content can spread unchecked [1].

A hoax, by definition, refers to deliberately fabricated information intended to mislead the public by presenting falsehoods as credible facts [2]. The proliferation of such content has triggered measurable societal harm, including the erosion of public trust, polarization of opinion, and disruption of informed decision-making. Studies have established that conventional machine learning methods, such as Support Vector Machines (SVM) combined with Term Frequency-Inverse Document Frequency (TF-IDF) feature extraction, can partially address this challenge by classifying deceptive news across specific thematic domains [3]. Nevertheless, these approaches are constrained by their reliance on manually engineered features and their inability to capture complex sequential relationships within text. The challenge is further compounded by low digital literacy levels in many communities, which makes individuals particularly vulnerable to deceptive content [4].

Social media ecosystems are architecturally grounded in Web 2.0 principles, which facilitate bidirectional interaction and user-generated content dissemination across interconnected digital communities [5]. This infrastructure inherently amplifies the virality of content, regardless of its veracity, making automated screening mechanisms indispensable.

In response, the research community has increasingly turned to deep learninga class of artificial intelligence techniques employing multi-layered neural networks capable of autonomously learning hierarchical feature representations from raw data [5][6][7]. Among the various deep learning architectures, Long Short-Term Memory (LSTM) networks have proven particularly effective for natural language processing (NLP) tasks. As a sophisticated variant of Recurrent Neural Networks (RNNs), LSTM addresses the vanishing gradient problem through its gated memory cell mechanism, which selectively retains and discards information over long input sequences [8][9]. This property renders LSTM especially adept at modeling semantic dependencies between words in a sentencea critical capability for understanding the context required to distinguish genuine news from fabricated narratives [6].

Prior work in this space has demonstrated promising results. A notable study employing a hybrid CNN-LSTM architecture reported classification accuracy reaching 99% by exploiting the complementary strengths of convolutional feature extraction and sequential context modeling [10]. However, the majority of existing research remains confined to model benchmarking under controlled conditions, with limited attention paid to practical deployment scenarios, particularly for Indonesian-language content circulating on platforms like Facebook.

This study addresses that gap by designing and implementing an LSTM-based classification pipeline specifically targeting Indonesian hoax content. Rather than focusing solely on performance benchmarking, the research emphasizes the construction of an end-to-end, deployable system capable of autonomously distinguishing credible information from misinformation. The ultimate objective is to contribute a practical tool that supports digital media literacy and assists in mitigating the societal impact of online hoaxes.

The IJICS (International Journal of Informatics and Computer Science)
Vol 10 No 2, July 2026, Page 161-166
ISSN 2548-8384 (online), ISSN 2548-8449 (print)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/ijics/index
DOI 10.30865/ijics.v10i2.9899

2. RESEARCH METHODOLOGY

2.1 Research Design

This research adopts a quantitative experimental design grounded in the text classification paradigm of machine learning. The methodology is structured to maximize reproducibility and scientific rigor. The central objective is to construct an automated hoax detection model by systematically combining NLP preprocessing techniques with a deep learning classifier, enabling the model to generalize effectively to unseen social media content.

2.2 Dataset and Data Collection

This research utilized the *Indonesian Fact and Hoax Political News* dataset, which is publicly available through the Kaggle data repositor [11]. The dataset contains manually annotated Indonesian-language news articles labeled as factual or hoax, making it suitable for supervised text classification using the Long Short-Term Memory (LSTM) algorithm. The factual articles were collected from trusted Indonesian online news portals, including *Kompas*, *Tempo*, and *CNN Indonesia*, whereas the hoax articles were sourced from *TurnBackHoax (MAFINDO)*, an Indonesian fact-checking organization.

The dataset comprised 4,502 news articles, including 2,251 factual articles and 2,251 hoax articles. Each record contains the news content and its corresponding class label. Prior to preprocessing, the dataset was divided into 80% training data (3,602 articles) and 20% testing data (900 articles) using stratified sampling to preserve the class distribution. The resulting datasets were subsequently processed through the preprocessing pipeline before being transformed into numerical representations for LSTM model training and evaluation.

2.3 Text Preprocessing Pipeline

Before feeding text data into the neural network, each document underwent a multi-stage preprocessing pipeline to reduce noise and normalize the input representation:

1. Case folding: All characters were converted to lowercase to eliminate case-based ambiguity.
2. Data cleaning: Non-linguistic elementsincluding numerical digits, punctuation marks, hyperlinks, and special symbolswere removed to retain only semantically meaningful tokens.
3. Tokenization: Each document was segmented into individual word tokens, forming the fundamental unit of text representation.
4. Stopword removal: High-frequency, semantically low-value words (e.g., conjunctions and prepositions) were filtered out to reduce dimensionality.
5. Stemming: Words were reduced to their morphological root forms using the Sastrawi stemmer, consolidating lexical variants and compressing the vocabulary space.

This pipeline collectively reduces vocabulary redundancy and computational overhead while sharpening the signal available to the downstream classification model [12].

3. RESULTS AND DISCUSSION

3.1 Dataset Description

The study utilized the "Indonesian Fact and Hoax Political News" dataset, publicly accessible via Kaggle. This corpus aggregates political news articles from reputable Indonesian outlets (Kompas, Tempo, CNN Indonesia) alongside debunked misinformation entries from TurnBackHoax. To counteract class imbalance and prevent biased model learning, random undersampling was performed to equalize both categories. The final balanced corpus comprises 4,502 recordsequally distributed as 2,251 hoax and 2,251 factual entries. Dataset partitioning followed an 80:20 ratio, allocating 3,602 samples for training and 900 samples for testing. The distribution is summarized in Table 1.

Table 1. Dataset Distribution

Training Data	Testing Data	Total Data
3,602	900	4,502

Hoax data examples:

a. The government will distribute IDR 10 million in cash to all citizens through the following link.
b. Drinking salt water can cure all types of cancer.
c. Vaccines cause humans to turn into robots.

Non-hoax data examples:

a. The government officially inaugurated the construction of a new toll road in North Sumatra.

The IJICS (International Journal of Informatics and Computer Science)
Vol 10 No 2, July 2026, Page 161-166
ISSN 2548-8384 (online), ISSN 2548-8449 (print)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/ijics/index
DOI 10.30865/ijics.v10i2.9899

b. The Meteorology, Climatology, and Geophysics Agency (BMKG) issued a weather forecast for the Medan area.
c. The university held a national seminar on Artificial Intelligence.

3.2 Preprocessing Output

The preprocessing pipeline was applied uniformly across all samples. To illustrate its effect, consider the input sentence: *"VAKSIN COVID-19 DAPAT MENGUBAH DNA MANUSIA"*. After sequential application of case folding, cleaning, tokenization, stopword removal, and stemming, the output token array is: ["vaksin", "covid", "ubah", "dna", "manusia"]. This transformation significantly reduces noise while preserving the core semantic content of the original text, demonstrating the effectiveness of the pipeline in preparing raw social media data for classification.

3.3 Word Vectorization

Preprocessed tokens were mapped to integer indices using the Keras Tokenizer, which constructs a vocabulary based on corpus-wide term frequencies and assigns a unique index to each word up to a defined vocabulary limit. To ensure uniform input dimensions required by the LSTM layer, all sequences were padded or truncated to a fixed maximum length. This standardization step is critical for enabling efficient batch processing during model training.

3.4 LSTM Model Architecture

The neural network architecture was designed to balance model capacity with generalization capability. Table 2 and Table 3 detail the layer configuration and training hyperparameters, respectively.

Table 2. LSTM Model Architecture

Layer / Component	Configuration
Embedding Layer	Vocabulary Size = 5,000; Dimension = 128
LSTM Layer	64 Units
Dropout	Rate = 0.2
Dense Output Layer	1 Neuron; Activation = Sigmoid

Table 3. Training Hyperparameters

Parameter	Value
Training Epochs	10
Batch Size	32
Optimizer	Adam
Loss Function	Binary Cross-Entropy

The embedding layer transforms integer-indexed tokens into dense, continuous vector representations, capturing semantic similarity between words. The LSTM layer then processes these embeddings sequentially, leveraging its gated memory mechanism to capture both short- and long-range syntactic and semantic dependencies. The dropout layer introduces regularization by randomly deactivating a fraction of neurons during training, reducing the risk of overfitting. The final sigmoid-activated dense layer maps the learned representation to a binary probability output.

3.5 LSTM Gate Computation

The LSTM unit governs information flow through three primary gates. At each time step t, given the previous hidden state h_{t-1} and the current input x_t, the gate computations proceed as follows:

1. Forget Gate: $f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$ determines what proportion of the previous cell state to discard.
2. Input Gate: $i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$ controls how much of the candidate information is incorporated.
3. Candidate Cell State: $\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$ computes potential new memory content.
4. Cell State Update: $C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t$ integrates retained and new information.
5. Output Gate: $o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$ governs how much of the cell state is exposed as the hidden state.
6. Hidden State Output: $h_t = o_t \times \tanh(C_t)$ the output passed to the next time step and to the classification layer.

The IJICS (International Journal of Informatics and Computer Science)
Vol 10 No 2, July 2026, Page 161-166
ISSN 2548-8384 (online), ISSN 2548-8449 (print)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/ijics/index
DOI 10.30865/ijics.v10i2.9899

Numerical Example:

Assume: $f_t = 0.80$, $i_t = 0.70$, $\tilde{C}_t = 0.60$, $C_{t-1} = 0.50$, $o_t = 0.75$

Cell state: $C_t = (0.80 \times 0.50) + (0.70 \times 0.60) = 0.40 + 0.42 = 0.82$

Hidden state: $h_t = 0.75 \times \tanh(0.82) = 0.75 \times 0.6761 \approx 0.507$

This resulting hidden state is propagated to the subsequent time step, with the final hidden state used by the dense output layer to generate the binary classification probability.

3.6 Training Performance

The model was trained across 10 epochs on the 3,602-sample training partition. Table 4 presents the progression of accuracy and loss metrics, illustrating the model's learning trajectory over successive epochs.

Table 4. Training Accuracy and Loss per Epoch

Epoch	Training Accuracy	Training Loss
1	0.74	0.58
2	0.79	0.49
3	0.83	0.42
4	0.86	0.37
5	0.88	0.32
6	0.90	0.28
7	0.91	0.25
8	0.92	0.22
9	0.93	0.19
10	0.94	0.17

A consistent upward trend in accuracy (0.74 → 0.94) accompanied by a steady reduction in loss (0.58 → 0.17) over the training period confirms that the model was learning meaningful representations from the data, with no evidence of premature convergence or instability.

3.7 Evaluation on Test Set

Final model evaluation was performed on the 900 held-out test samples. The confusion matrix (Table 5) summarizes the classification outcomes across both categories.

Table 5. Confusion Matrix

	Predicted: Hoax	Predicted: Non-Hoax
Actual: Hoax	415 (TP)	35 (FN)
Actual: Non-Hoax	33 (FP)	417 (TN)

3.8 Performance Metrics

From the confusion matrix values (TP = 415, TN = 417, FP = 33, FN = 35), the four standard evaluation metrics were computed:

6. Accuracy = (TP + TN) / (TP + TN + FP + FN) = (415 + 417) / 900 = 832/900 = 92.44%
7. Precision = TP / (TP + FP) = 415 / 448 = 92.63%
8. Recall = TP / (TP + FN) = 415 / 450 = 92.22%
9. F1-Score = 2 × (Precision × Recall) / (Precision + Recall) = 92.42%

3.9 Discussion

The evaluation results confirm that the LSTM-based classifier achieves strong and balanced performance across all four metrics. The 92.44% accuracy reflects the model's overall classification reliability, while the near-symmetric

The IJICS (International Journal of Informatics and Computer Science)
Vol 10 No 2, July 2026, Page 161-166
ISSN 2548-8384 (online), ISSN 2548-8449 (print)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/ijics/index
DOI 10.30865/ijics.v10i2.9899

precision (92.63%) and recall (92.22%) values indicate that the model does not exhibit disproportionate bias toward either classan important characteristic for real-world deployment where both false positives (flagging genuine content) and false negatives (missing hoaxes) carry significant consequences. The effectiveness of the preprocessing pipeline was a critical contributing factor. By standardizing and compressing the textual representation, preprocessing enabled the LSTM layers to focus computational resources on semantically meaningful features rather than surface-level noise. The model's proficiency in capturing long-range word dependencies a limitation of traditional bag-of-words approachesdirectly underpins its ability to contextualize hoax indicators within sentence-level and paragraph-level structures. These results establish a strong empirical baseline for LSTM-based hoax classification in the Indonesian language context and support the viability of developing production-ready misinformation screening tools.

Table 6. presents a comparison between the proposed LSTM model and several previous studies on Indonesian hoax detection.

Reference	Method	Reported Accurascy (%)
Sudrajat *et al* [13]	Naive Bayes	82.00
Mambunan [14]	Naïve Bayes	66.52
Mambunan *et al* [14]	SVM	70.87
This study	LSTM	92.44

As shown in Table 6, the proposed LSTM model achieved an accuracy of 92.44%, exceeding the reported performance of several conventional machine learning methods. However, because each study employed different datasets and evaluation procedures, the comparison should be interpreted only as a general performance reference.

4. CONCLUSION

This study designed and validated an LSTM-based deep learning pipeline for the automated classification of Indonesian hoax content on Facebook. The complete workflowspanning data acquisition, multi-stage text preprocessing, word vectorization, model training, and rigorous evaluationyielded a classifier with 92.44% accuracy, 92.63% precision, 92.22% recall, and a 92.42% F1-score on a balanced test set. These results demonstrate the LSTM architecture's suitability for capturing the sequential semantic patterns that distinguish credible information from fabricated narratives in Indonesian text. The system presents a functional prototype for real-world application as a content credibility screening tool for social media users and platform administrators. Future research directions include scaling the training corpus with more linguistically diverse samples, incorporating contextual word embeddings such as IndoBERT or multilingual BERT to leverage transfer learning, and extending the system into a deployable real-time web or mobile application. Investigating ensemble methods that combine LSTM with transformer-based models may further push classification performance beyond current benchmarks.

REFERENCES

[1] P. W. Cahyo and U. S. Aesyi, "Perbandingan LSTM dengan Support Vector Machine dan Multinomial Naïve Bayes pada Klasifikasi Kategori Hoax," vol. 20, no. 2, pp. 23–29, 2023.

[2] A. Wenang, K. A. Widianto, and M. E. Nugraha, "Implementasi Model Deep Learning IndoBERT dengan Antarmuka Aplikasi Mobile untuk Deteksi Berita Hoaks Berbahasa Indonesia," vol. 5, no. 2, pp. 99–113, 2025.

[3] P. Ramadani *et al.*, "Klasifikasi Berita Hoaks Menggunakan Algoritma Support Vector Machine," no. November, 2025.

[4] R. Yunanto, A. P. Purfini, and A. Prabuwisesa, "Survei Literatur : Deteksi Berita Palsu Menggunakan Pendekatan Deep Learning," vol. xx, pp. 118–130, 2021, doi: 10.34010/jamika.v11i2.493.

[5] S. Herdiyani, L. Auliana, and I. Sukoco, "PERANAN MEDIA SOSIAL DALAM MENGEMBANGKAN SUATU BISNIS :," vol. 18, no. 2, pp. 103–121, 2022.

[6] H. Kopi, J. Pendek, and D. A. N. Jangka, "Penerapan lstm dalam deep learning untuk prediksi harga kopi jangka pendek dan jangka panjang," vol. 10, no. 1, pp. 554–564, 2025.

[7] F. Santoso and F. Lazim, "G-Tech : Jurnal Teknologi Terapan," vol. 8, no. 3, pp. 1749–1758, 2024.

[8] A. Hanafiah, Y. Arta, H. O. Nasution, and Y. D. Lestari, "BULLETIN OF COMPUTER SCIENCE RESEARCH Penerapan Metode Recurrent Neural Network dengan Pendekatan Long Short-Term Memory (LSTM) Untuk Prediksi Harga Saham," vol. 4, no. 1, pp. 27–33, 2023, doi: 10.47065/bulletincsr.v4i1.321.

[9] W. Afandi *et al.*, "Klasifikasi Judul Berita Clickbait menggunakan RNN-," vol. 7, no. 2, pp. 85–89, 2022.

[10] F. Salim *et al.*, "Jurnal Computer Science and Information Technology (CoSciTech) Klasifikasi Berita Palsu Menggunakan

The IJICS (International Journal of Informatics and Computer Science)
Vol 10 No 2, July 2026, Page 161-166
ISSN 2548-8384 (online), ISSN 2548-8449 (print)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/ijics/index
DOI 10.30865/ijics.v10i2.9899

Pendekatan Hybrid CNN-LSTM Fake News Classification Using Hybrid CNN-LSTM Approach," vol. 6, no. 1, pp. 55–59, 2025.

[11] S. Merinda, C. Ciksadan, and M. Fadhli, "Perbandingan Algoritma Support Vector Machine dan Bi-Directional Long Short Term Memory Dalam Mengklasifikasi Berita Hoaks," vol. 7, no. 1, pp. 367–378, 2025, doi: 10.47065/bits.v7i1.7391.

[12] V. No, J. Hal, M. D. Desriansyah, and I. Utna, "Analisis Efektivitas Algoritma Machine Learning dalam Deteksi Hoaks : Pada Berita Digital Berbahasa Indonesia," vol. 3, no. 2, pp. 63–69, 2025.

[13] A. Sudrajat and R. R. Wulandari, "Indonesian Language Hoax News Classification Based on Naïve Bayes," vol. 7, no. 1, pp. 70–79, 2022.

[14] K. A. Mumbunan *et al.*, "Perbandingan Algoritma Naive Bayes Dan Support Vector Machine Dalam Deteksi Berita Hoax Berbahasa Indonesia," vol. 4, no. 1, pp. 52–64, 2025.