Recursive Journal of Informatics
Vol. XX, No. XX, Month 20XX

p-ISSN 2963-5551 https://journal.unnes.ac.id/sju/index.php/rji e-ISSN 2986-6588

Application of Ensemble Method for Indonesian-Language Hoax News Classification Using a Combination of Logistic Regression and Long Short-Term Memory

Achmad Dylan Alfaris[1]*, Endang Sugiharti[2]
[1,2]Computer Science Department, Faculty of Mathematics and Natural Sciences,
Universitas Negeri Semarang, Indonesia

Abstract.
Purpose: The rapid spread of Indonesian-language hoax news through online media threatens public trust and social stability. Existing detection systems often rely solely on either machine learning or deep learning, which limits their ability to capture both statistical and contextual aspects of language. This study aims to build an accurate and efficient hoax classification system by integrating Logistic Regression (LR) and Long Short-Term Memory (LSTM) within an ensemble soft voting framework.
Methods/Study design/approach: A quantitative approach was applied using the Indonesian Fact and Hoax News dataset from Kaggle, consisting of 12,405 news articles evenly divided into 5,842 hoax and 6,563 valid items. The dataset underwent preprocessing including text cleaning, tokenization, stopword removal, and stemming. Logistic Regression was implemented with TF-IDF features and L2 regularization, while LSTM employed 100-dimensional Word2Vec embeddings, the Adam optimizer, and early stopping. The probabilistic outputs of both models were averaged through soft voting to generate the final prediction.
Result/Findings: The ensemble of Logistic Regression and LSTM outperformed the individual models used in this study, achieving an accuracy of 97.42%, precision of 96.87%, recall of 98.02%, and F1-score of 97.44%. These results demonstrate that combining machine learning and deep learning methods in an ensemble framework significantly enhances the model's capability to detect Indonesian-language hoax news with higher precision and recall compared to single-model approaches.
Novelty/Originality/Value: This study presents a novel ensemble approach that integrates a statistical machine learning model (Logistic Regression with TF-IDF) and a deep learning architecture (LSTM with Word2vec) for Indonesian-language hoax detection. The proposed method leverages the interpretability and efficiency of Logistic Regression alongside the contextual understanding of LSTM, resulting in a more robust and accurate classification system. This framework can be adapted for multilingual misinformation detection and serves as a practical solution for media platforms and regulators in combating hoax dissemination.

Keywords: Hoax, Ensemble Learning, TF-IDF, Word2vec, Logistic Regression, LSTM
Received Month 20xx / **Revised** Month 20xx / **Accepted** Month 20xx

INTRODUCTION

The growth of online media in Indonesia has accelerated the spread of hoaxes, with over 12,500 cases recorded by the Ministry of Communication and Informatics between August 2018 and December 2023, dominated by health and political issues [1]. Technology's ease of access shapes how people consume and respond to information [2]. In Indonesia, spreading false and misleading news is punishable under the Electronic Information and Transactions Law (UU ITE) with up to six years' imprisonment and/or a fine of one billion rupiah [3]. These conditions underscore the urgency of developing robust and context-aware automated hoax detection systems, especially considering the morphological complexity of the Indonesian language.

Machine learning (ML) offers efficiency in processing textual patterns [4], while deep learning (DL) provides superior contextual understanding [5]. Previous research has explored these approaches

[1]*Corresponding author.
Email addresses: achmaddyl@students.unnes.ac.id (Alfaris), endangsugiharti@mail.unnes.ac.id (Sugiharti)
DOI: 10.15294/sji.v8i1.25356

separately. Ramadhan et al. applied Random Forest and Logistic Regression with TF-IDF features, achieving 84% accuracy but lacking semantic depth [6]. Yusuf and Suyanto implemented LSTM with Word2vec, improving contextual modeling to 89.42% accuracy but at higher computational costs [7]. Adrian et al. achieved 95% with LSTM–Word2vec, yet still relied solely on DL [8]. Hanum et al. used BERT-based models with strong language representation but reached only 76% accuracy, while Rachmawati and Darmawan found GRU to be lighter than LSTM but still computationally demanding at 90% accuracy [9], [10].

In broader text classification studies, hybrid models combining ML and DL have shown potential to enhance performance and reduce detection errors [11], [12]. TF-IDF with Logistic Regression effectively captures simple textual features such as word frequency and sentence structure [13], while LSTM excels in modeling sequential dependencies for more context-aware predictions [14]. However, this hybrid approach remains underexplored in Indonesian-language hoax detection. To address this gap, this study proposes an ensemble of TF-IDF-based Logistic Regression and Word2vec-based LSTM using soft voting, combining ML's efficiency with DL's contextual depth to improve accuracy and efficiency in Indonesian hoax detection over single models.

METHODS

This study employed a quantitative research design utilizing an ensemble soft voting approach that integrates TF-IDF-based Logistic Regression and Word2vec-based Long Short-Term Memory (LSTM) to detect hoax news. The research procedure comprised six main stages: data labelling, preprocessing, data splitting, feature extraction for model training, ensemble soft voting, and performance evaluation. A comprehensive representation of these stages is illustrated in Figure 1.

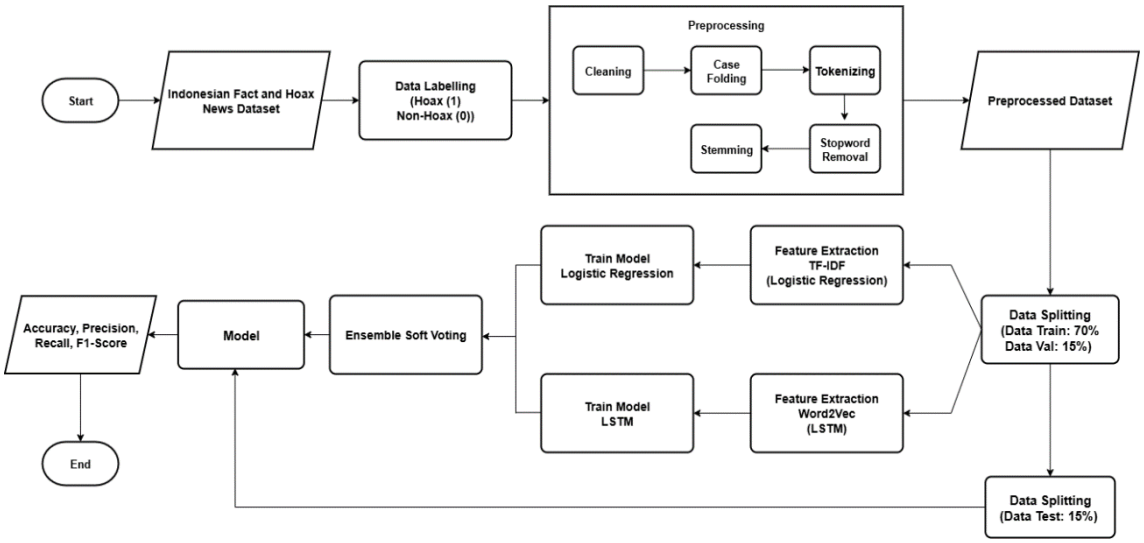

Figure 1. Flowchart Model

Dataset

The dataset used in this study was the Indonesian Fact and Hoax News dataset obtained from Kaggle, sourced from news sites such as TurnBackHoax.id, Kompas, CNN, and Tempo. The final dataset contained a balanced distribution of 5.050 hoax news texts and 5.042 non-hoax news texts. The dataset is organized into two columns: text, containing the news content, and label, a binary value where 0 represents non-hoax news and 1 represents hoax news.

Preprocessing

The preprocessing stage aimed to clean and normalize the text data. This process included text cleaning, case folding, tokenization, stopword removal, and stemming. The main objective of this stage was to

improve the efficiency and accuracy of natural language processing by reducing data dimensionality and focusing on more meaningful words [15].

Data Splitting

In this stage, a total of 12,405 preprocessed Indonesian news texts were used. The dataset was split into three subsets: 70% for training, 15% for validation, and 15% for testing, as shown in Table 1. This division aimed to optimize the model training process while ensuring fair and unbiased performance evaluation on unseen data.

Table 1. Data Splitting

Dataset	Total
Training	8.683
Validation	1.861
Testing	1.861

Model Architecture

The model architecture employed in this study integrates two distinct approaches, namely Logistic Regression (LR) and Long Short-Term Memory (LSTM), which are subsequently combined using the Ensemble Soft Voting method. In the first branch of the architecture, Logistic Regression functions as a conventional, statistically based classification model that effectively processes binary text data after feature extraction through the Term Frequency–Inverse Document Frequency (TF-IDF) technique. This model is capable of capturing explicit statistical patterns present in textual data, thereby aiding in the identification of word characteristics frequently found in both hoax and valid news. The processing and training pipeline for the Logistic Regression model is depicted in the flowchart presented in Figure 2.

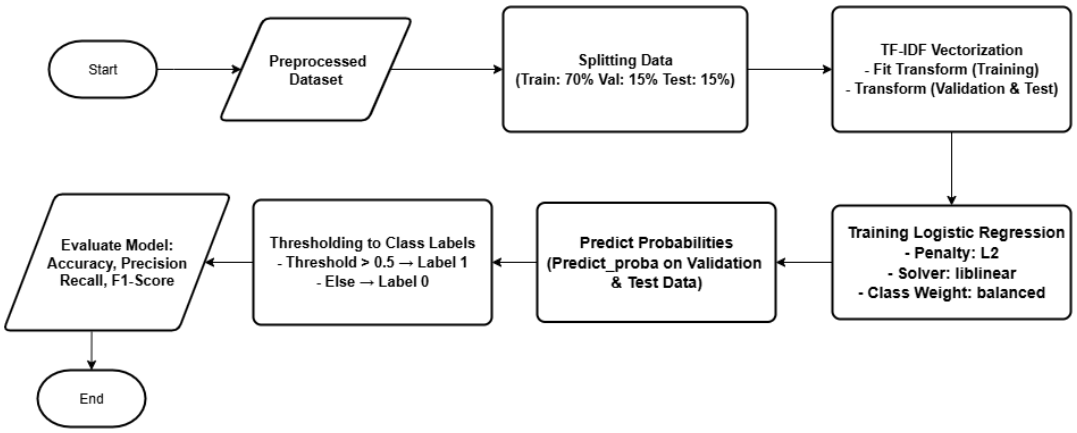

Figure 2. Flowchart Of Logistic Regression

In parallel, the second branch of the architecture employs LSTM, a deep learning model with the capacity to capture word sequences and contextual relationships within the text. The LSTM architecture implemented in this study comprises an embedding layer initialized with a pre-trained Word2Vec embedding matrix, followed by an LSTM layer with a specified number of units to process sequential information, and a dropout layer to mitigate overfitting. The architecture concludes with a dense layer equipped with a sigmoid activation function to produce classification probability values. The process of constructing and training the LSTM model is illustrated in the flowchart shown in Figure 3. Both models are trained independently to achieve optimal performance, after which their probability outputs are fused during the ensemble stage to yield final predictions that are both more accurate and more stable.

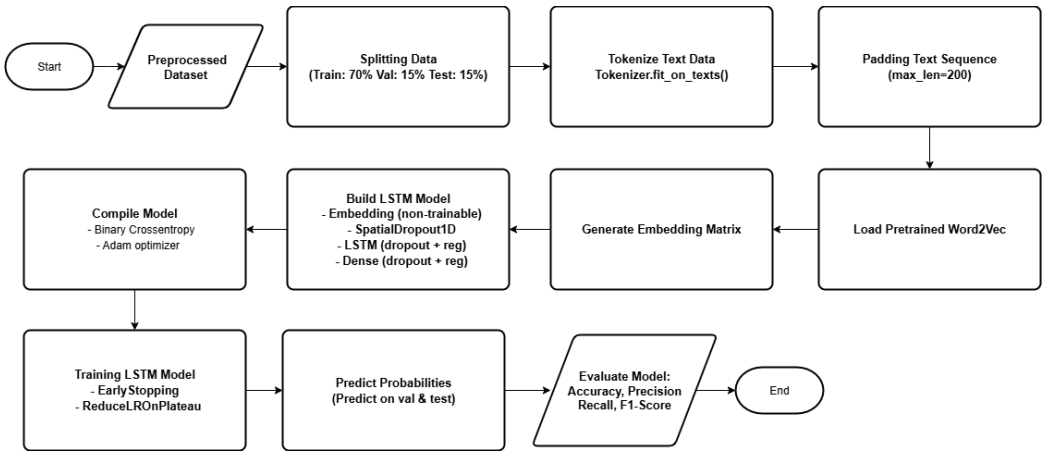

Figure 3. Flowchart Of Long Short-Term Memory

Feature Extraction

In this stage, feature extraction transforms preprocessed text data into numerical representations suitable for machine learning and deep learning models. Two primary methods were employed: Term Frequency-Inverse Document Frequency (TF-IDF) for the Logistic Regression model and Word2Vec embeddings for the Long Short-Term Memory (LSTM) model. TF-IDF vectors capture the importance of words within each document, while Word2Vec embeddings represent semantic relationships between words. This process aims to optimally preserve textual information, thereby enhancing the models' performance in classifying hoax and non-hoax news.

Ensemble Soft Voting

This study employs a soft voting ensemble method to combine Logistic Regression (LR) and LSTM models using a late fusion approach, where each model independently produces probabilistic predictions after training [16]. Soft voting considers the confidence level of each model's prediction, leading to a more balanced decision. The averaged probabilities are then thresholded at 0.5, predictions above this value are classified as hoax news (class 1), while those below are valid (class 0). The comprehensive soft voting ensemble process is illustrated in Figure 4.

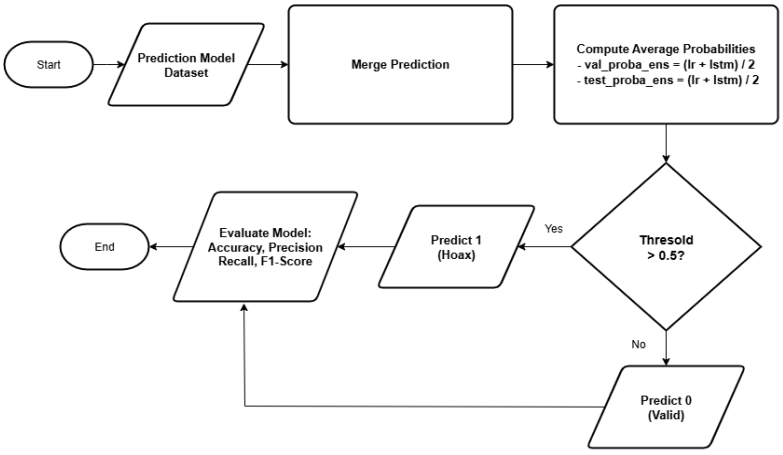

Figure 4. Flowchart Of Ensemble Model

Evaluation Metrics

Model performance was evaluated using several classification metrics: accuracy, precision, recall, and F1-score. These metrics were derived from the confusion matrix, which provides a breakdown of true positives, false positives, true negatives, and false negatives. The use of these metrics is common in binary classification to provide a comprehensive overview of the predictive capability of the developed model [17].

Table 2. Confusion Matrix

Confusion Matrix		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

RESULT AND DISCUSSION

Preprocessing

Preprocessing is an essential step in this research to ensure the quality of textual data used for model training. The process consisted of several stages, including checking for missing values, case folding, tokenization, removal of symbols and numbers, stopword removal, and stemming.

The initial inspection showed that the dataset contained no missing values, allowing all data to be processed directly. Case folding was applied to standardize the text by converting all characters to lowercase. Tokenization was then performed to split news articles into smaller units of words (tokens). Subsequently, symbols, numbers, and special characters were removed to retain only the words relevant to the context of the news. Stopword removal was carried out to eliminate common words such as “yang,” “dan,” or “di,” which do not provide significant contribution to the classification process. As a result, the total number of words in the dataset was reduced from 1,648,798 to 1,451,822. Finally, stemming was applied to convert inflected words into their root forms so that words with the same meaning were treated as identical.

In addition to text cleaning, document length statistics (measured in the number of words per news article) were also analyzed to better understand the dataset characteristics. As shown in Figure 5, the shortest document contained 10 words, while the longest reached 2,824 words. The average document length was 115.08 words, with a median of 68 words, indicating that most news articles were relatively short. Furthermore, the 90th percentile showed a document length of 260 words, meaning that the vast majority of news texts did not exceed 260 words.

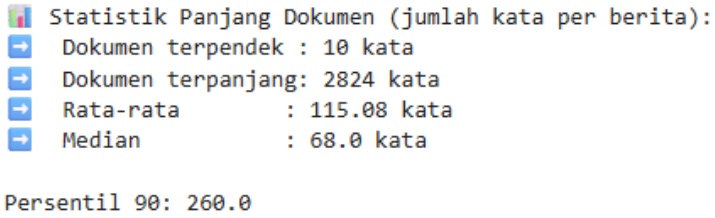

Figure 5. Document Length Statistics

Overall, preprocessing successfully removed irrelevant elements while normalizing the text into a cleaner and more consistent form. The document length statistics provide additional insights into the diversity of text lengths in the dataset, which is particularly important when determining parameters such as the maximum sequence length for the LSTM model. With these preprocessing results, the dataset becomes more concise, representative, and ready for optimal feature extraction using TF-IDF and Word2Vec.

Logistic Regression Model Performance

The Logistic Regression model trained using TF-IDF features with L2 regularization and the liblinear solver achieved an accuracy of 95.90%, precision of 94.60%, recall of 97.35%, and an F1-score of 95.96%. These results indicate that Logistic Regression is quite effective for classifying Indonesian hoax news, despite being a relatively simple statistical method. The model successfully captures word distribution patterns and the frequency relationships commonly found in hoax news. However, since TF-IDF does not account for the sequential context of words, the model still has limitations in handling complex sentence structures and nuanced semantic meanings.

Logistic Regression Model Performance

As shown in Figure 6, confusion matrix shows a high number of True Positives (TP), indicating strong capability in detecting hoax news. The False Negatives (FN) are relatively low, meaning very few hoax news items were misclassified as valid. Nevertheless, there are some False Positives (FP), indicating that a

small portion of valid news was incorrectly labeled as hoax. This suggests that while the model's precision is high, it could benefit from being more selective to reduce the misclassification of valid news as hoax.

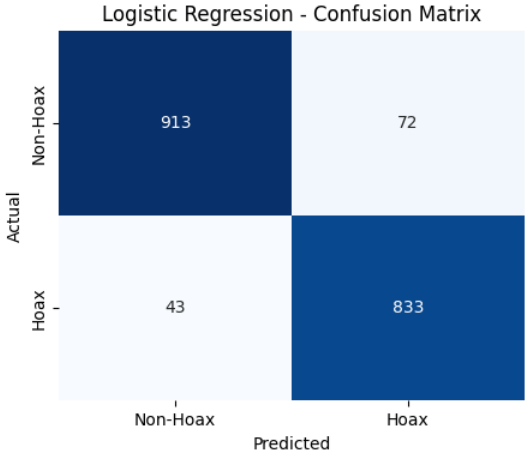

Figure 6. Logistic Regression Confusion Matrix

Long Short-Term Memory (LSTM) Model Performance

The LSTM model, utilizing Word2Vec embeddings, the Adam optimizer, and an early stopping strategy, demonstrated superior performance compared to Logistic Regression, with an accuracy of 97.22%, precision of 96.73%, recall of 97.75%, and an F1-score of 97.24%. This improvement is attributed to LSTM's ability to understand word context based on sequences, which TF-IDF cannot capture. By leveraging Word2Vec embeddings, the model gains richer semantic representations of words, allowing it to better handle the linguistic variations often found in hoax news.

LSTM Confusion Matrix Analysis

The confusion matrix for LSTM shows a more balanced number of True Positives (TP) and True Negatives (TN) compared to Logistic Regression, with fewer False Positives (FP) and False Negatives (FN). The high recall indicates that LSTM excels at correctly identifying hoax news (minimizing FN), thereby reducing the risk of hoax content being misclassified as valid news. Its confusion matrix is displayed in Figure 7.

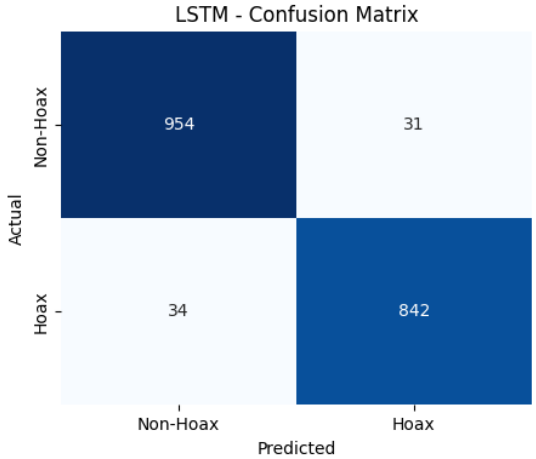

Figure 7. Long Short-Term Memory Confusion Matrix

LSTM Training and Validation Curves

The graphs in Figure 8 illustrate the training and validation loss (left) and accuracy (right) of the LSTM model across 17 epochs. The training loss (blue line) shows a steady decline from the first epoch, indicating that the model is successfully learning patterns from the training data. The validation loss (red dashed line) follows a similar downward trend and eventually stabilizes around a low value, suggesting that the model is generalizing well to unseen data without overfitting.

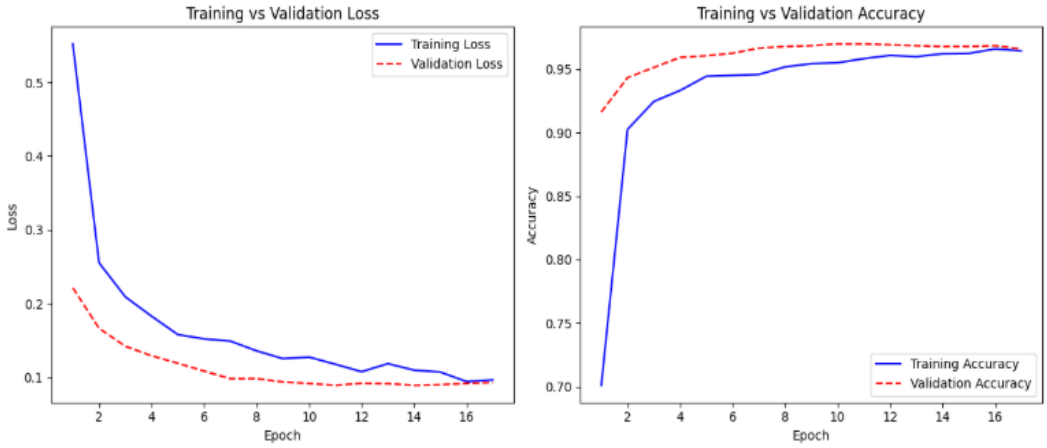

Figure 8. LSTM Training and Validation Curves

The accuracy curves show a complementary pattern: both training and validation accuracy increase rapidly during the initial epochs, with the validation accuracy (red dashed line) consistently higher than training accuracy in the early stages, which may indicate effective regularization and early stopping. After approximately epoch 10, both training and validation accuracy converge near 97%, demonstrating stable and high performance of the LSTM model. Overall, these curves indicate that the LSTM model learns efficiently, achieves good generalization, and avoids significant overfitting throughout the training process.

Ensemble Soft Voting Model Performance

The soft voting ensemble approach, which combines probabilistic predictions from Logistic Regression and LSTM, achieved the highest performance with an accuracy of 97.42%, precision of 96.87%, recall of 98.02%, and an F1-score of 97.44%. These results suggest that combining statistical-based (LR) and context-based (LSTM) models allows their strengths to complement each other. Logistic Regression contributes by effectively handling explicit features like word frequency, while LSTM enhances contextual understanding. By integrating both through soft voting, a better balance between precision and recall is achieved.

Ensemble Confusion Matrix Analysis

The confusion matrix of the ensemble method shows the highest number of True Positives among all models and the lowest False Negatives, meaning nearly all hoax news was correctly classified. The False Positives decreased compared to Logistic Regression, though they remain slightly higher than LSTM. This demonstrates that the ensemble not only improves overall accuracy but also stabilizes predictions by leveraging the advantages of both models. Its confusion matrix is displayed in Figure 9.

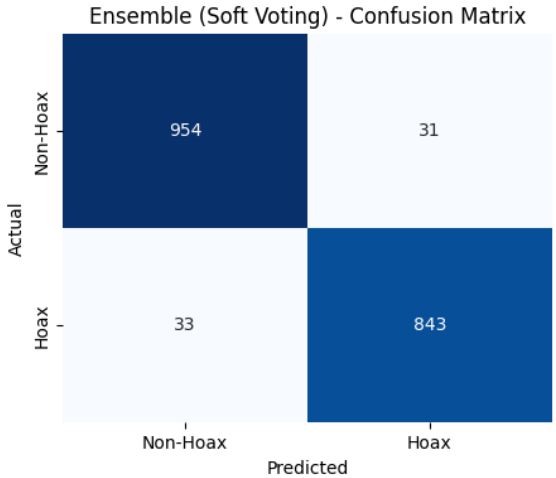

Figure 9. Ensemble Soft Voting Confusion Matrix

Training Result

Table 3 presents the complete performance metrics for all models, including accuracy, precision, recall, and F1-score on the test set. The results confirm that the Ensemble Soft Voting method consistently achieved the highest performance across all metrics.

Table 3. Training Result Model

No	Model	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	95.90%	94.60%	97.35%	95.96%
2	Long Short-Term Memory	97.22%	96.73%	97.75%	97.24%
3	Ensemble Soft Voting	97.42%	96.86%	98%	97.43%

Performance Comparison Analysis on Previous Studies

This study presents a comparison with several previous works that addressed the detection of Indonesian-language hoax news. The comparison of accuracy results between this study and prior research is shown in Table 4.

Table 4. Comparison on Previous Studies

No	Author	Method	Dataset	Result
1	Hanum [9]	BERT	Kaggle	76%
2	Rachmawati & Darmawan [10]	LSTM	Kaggle	90%
3	Ramadhan [6]	Logistic Regression	Kaggle	77%
4	Adrian [8]	LSTM	Kaggle	95%
4	Purpose Method	Logistic Regression + LSTM	Kaggle	97.42%

The comparison results indicate that selecting an appropriate ensemble model, such as the combination of Logistic Regression and LSTM through the soft voting method, can achieve higher accuracy compared to single-model approaches used in earlier studies. The main challenges in classifying Indonesian-language hoax news lie in the diversity of vocabulary, the complexity of sentence structures, and the varying contexts, which require a method capable of capturing both statistical patterns and contextual meaning. The approach proposed in this study still holds potential for further improvement through hyperparameter optimization, the use of more representative embeddings, or the application of next-generation deep learning architectures such as transformers.

CONCLUSION

This study successfully developed a hoax news classification model using an ensemble soft voting method that combines the machine learning Logistic Regression model with the deep learning LSTM model. The proposed method achieved 97.42% accuracy, 96.86% precision, 98.01% recall, and a 97.44% F1-score, outperforming individual models. These results demonstrate that integrating machine learning and deep learning yields more accurate and stable predictions by leveraging both TF-IDF-based features and semantic-rich embeddings, effectively mitigating the limitations of each individual model. This research not only contributes to improving hoax detection accuracy but also provides a foundation for future studies exploring other ensemble techniques such as stacking or boosting, incorporating more than two models, extending to multilingual datasets, real-time classification, and integrating advanced pre-trained language models to enhance robustness and adaptability.

REFERENCES

[1] Tim Cek Fakta, "INFOGRAFIK: Kominfo Temukan 12.547 Konten Hoaks, Simak Datanya," *Kompas*, Jan. 04, 2024. [Online]. Available: https://www.kompas.com/cekfakta/read/2024/01/04/192000682/infografik--kominfo-temukan-12.547-konten-hoaks-simak-datanya

[2] C. Juditha, "Hoax Communication Interactivity in Social Media and Anticipation," *J. Pekommas*, vol. 3, no. 1, p. 31, 2018, doi: 10.30818/jpkm.2018.2030104.

[3] Y. S. Laowo, "Analisis Hukum Tentang Penyebararan Berita Bohong (Hoax) Menurut Uu No. 11 Tahun 2008 Jo Uu No. 19 Tahun 2016," *J. Educ.*, vol. 8, no. 1, pp. 440–448, 2020, [Online]. Available: http://journal.ipts.ac.id/index.php/ED/article/view/1650

[4] I. I. Sholikhah, A. T. J. Harjanta, and K. Latifah, “Machine Learning Untuk Deteksi Berita Hoax Menggunakan BERT,” *Pros. Semin. Nas. Inform.*, vol. 1, no. 1, pp. 524–531, 2023, [Online]. Available: https://conference.upgris.ac.id/index.php/infest/article/view/3818

[5] Winda Kurnia Sari, D. P. Rini, Reza Firsandaya Malik, and Iman Saladin B. Azhar, “Multilabel Text Classification in News Articles Using Long-Term Memory with Word2Vec,” *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 2, pp. 276–285, 2020, doi: 10.29207/resti.v4i2.1655.

[6] N. G. Ramadhan, F. D. Adhinata, A. J. T. Segara, and D. P. Rakhmadani, “Deteksi Berita Palsu Menggunakan Metode Random Forest dan Logistic Regression,” *JURIKOM (Jurnal Ris. Komputer)*, vol. 9, no. 2, p. 251, 2022, doi: 10.30865/jurikom.v9i2.3979.

[7] R. Yusuf and S. Suyanto, “Hoax Detection on Indonesian Text using Long Short-Term Memory,” *ICOIACT 2022 - 5th Int. Conf. Inf. Commun. Technol. A New W. to Make AI Useful Everyone New Norm. Era, Proceeding*, pp. 268–271, 2022, doi: 10.1109/ICOIACT55506.2022.9972086.

[8] M. G. Adrian, S. S. Prasetyowati, and Y. Sibaroni, “Effectiveness of Word Embedding GloVe and Word2Vec within News Detection of Indonesian uUsing LSTM,” *J. Media Inform. Budidarma*, vol. 7, no. 3, p. 1180, 2023, doi: 10.30865/mib.v7i3.6411.

[9] A. R. Hanum *et al.*, “Analisis Kinerja Algoritma Klasifikasi Teks Bert dalam Mendeteksi Berita Hoaks,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 11, no. 3, pp. 537–546, 2024, doi: 10.25126/jtiik.938093.

[10] O. C. R. Rachmawati and Z. M. E. Darmawan, “The Comparison of Deep Learning Models for Indonesian Political Hoax News Detection,” *CommIT J.*, vol. 18, no. 2, pp. 123–135, 2024, doi: 10.21512/commit.v18i2.10929.

[11] A. Bondielli and F. Marcelloni, “A survey on fake news and rumour detection techniques,” *Inf. Sci. (Ny).*, vol. 497, pp. 38–55, Sep. 2019, doi: 10.1016/j.ins.2019.05.035.

[12] A.-A. Tanvir, E. M. Mahir, S. Akhter, and M. R. Huq, *Detecting Fake News using Machine Learning and Deep Learning Algorithms*. IEEE, 2019.

[13] R. A. Zahra and E. B. Setiawan, “Hoax Identification on Social Media Using Recurrent Neural Network (RNN) and Long Short-term Memory (LSTM) Methods,” *2023 11th Int. Conf. Inf. Commun. Technol. ICoICT 2023*, vol. 2023-Augus, pp. 448–451, 2023, doi: 10.1109/ICoICT58202.2023.10262687.

[14] Nurkholis, M. Z. Negara, G. F. Shidik, A. Z. Fanani, Muljono, and E. Noersasongko, “Sentiment Analysis of Indonesian News Using Deep Learning,” *Int. Semin. Appl. Technol. Inf. Commun.*, pp. 261–265, 2018.

[15] E. Alhenawi, R. A. Khurma, P. A. Castillo, M. G. Arenas, and A. M. Al-Hinawi, “Effects of term weighting approach with and without stop words removing on Arabic text classification,” *2023 9th Int. Conf. Optim. Appl. ICOA 2023 - Proc.*, 2023, doi: 10.1109/ICOA58279.2023.10308816.

[16] L. Hasimi and A. Poniszewska-Maranda, “Ensemble Learning-based Fake News and Disinformation Detection System,” *Proc. - 2021 IEEE Int. Conf. Serv. Comput. SCC 2021*, pp. 145–153, 2021, doi: 10.1109/SCC53864.2021.00027.

[17] M. Fahmuddin, M. K. Aidid, and M. J. Taslim, “Implementasi Analisis Regresi Logistik Dengan Metode Machine Learning Untuk Mengklasifikasi Berita Di Indonesia,” *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 5, no. 03, pp. 155–162, 2023, doi: 10.35580/variansiunm116