Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 3, June 2026, pp. 774~785
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i3.pp774-785

Fine-tuned IndoBERT for stock market sentiment analysis: evidence from CNBC Indonesia news

Tri Agung Jiwandono[1], MS Hendriyawan Achmad[2], Suhirman[1]
[1]Master's Program, Information Technology Study Program, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
[2]Department of Electrical and Electronic Engineering, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

Article Info

Article history:

Received Nov 30, 2025
Revised Mar 27, 2026
Accepted May 26, 2026

Keywords:

Financial news classification
IndoBERT sentiment analysis
Stock market sentiment
Transformer models

ABSTRACT

Financial sentiment analysis in Indonesian markets faces significant accuracy challenges, with existing models achieving only 78-81% accuracy. We present a fine-tuned IndoBERT-Large model for classifying sentiment in Indonesian stock market news headlines, trained on 9,819 CNBC Indonesia headlines (January 2024-March 2025). Through systematic hyperparameter optimization and stratified vocabulary-balanced splitting, our model achieved 94.20% accuracy, surpassing previous baselines by 4-16 percentage points. These results demonstrate IndoBERT's effectiveness for Indonesian financial NLP and its potential for real-time market monitoring and investment decision support systems.

Corresponding Author:

Tri Agung Jiwandono
Master's Program, Information Technology Study Program, Universitas Teknologi Yogyakarta
Sleman, Special Region of Yogyakarta, Indonesia
Email: tj.triagungj@gmail.com

1. INTRODUCTION

Background and context: stock investments represent ownership claims in companies, where stock prices reflect investment value and market expectations [1]. Investors allocate financial resources to purchase securities with expectations of returns on capital. However, stock prices exhibit natural volatility, creating risk that necessitates sophisticated analytical tools. Financial news sentiment analysis has emerged as a critical component in predicting market movements and supporting investment decisions [2], [3].

Sentiment analysis, a fundamental branch of natural language processing (NLP), identifies emotions in text by classifying content as positive, negative, or neutral [4]. Recent advances in transformer-based models, particularly bidirectional encoder representations from transformers (BERT) and its variants, have achieved remarkable improvements in sentiment classification tasks, consistently outperforming traditional machine learning approaches [5], [6]. These pre-trained language models leverage transfer learning to capture contextual representations, enabling superior performance across diverse NLP applications [7].

Research gap and problem statement: despite significant advances in NLP, sentiment analysis of Indonesian financial news faces substantial performance limitations. Previous studies on Indonesian financial sentiment classification have achieved only 78% accuracy by Adhi *et al.* [8] and 81% accuracy by Tyas *et al.* [9], indicating a significant gap compared to general-domain sentiment analysis. In contrast, sentiment analysis on Indonesian product reviews achieves 96% accuracy, highlighting the unique challenges of financial text [10], [11].

The research gap: financial news exhibits distinct characteristics that complicate sentiment classification: i) complex linguistic structures with domain-specific terminology; ii) subtle sentiment

expressions requiring deep contextual understanding; iii) limited availability of large-scale labeled Indonesian financial datasets; and iv) the need for models that capture nuanced financial language patterns.

This study addresses these challenges by proposing a novel fine-tuning strategy for IndoBERT-Large, specifically optimized for the Indonesian financial domain. The theoretical contribution of this work lies in the systematic exploration of vocabulary-balanced splitting and hyperparameter optimization, providing a new methodological framework for Indonesian financial NLP that bridges the gap between general-domain and domain-specific performance. We explicitly demonstrate how domain-adapted transformer architectures can capture the unique semantics of Indonesian market news, offering significant academic insights into the transferability of large-scale language models to specialized, low-resource financial contexts.

Research objectives; this study addresses the identified gap through the following objectives:

a. Develop a large-scale Indonesian financial news sentiment dataset from CNBC Indonesia headlines.
b. Implement systematic hyperparameter tuning strategies for IndoBERT-Large fine-tuning.
c. Evaluate vocabulary-balanced data splitting approaches using Jaccard similarity metrics.
d. Compare model performance across multiple train-test configurations and learning rates.
e. Establish baseline benchmarks for Indonesian financial sentiment classification.

Contributions; this study contributes to the field of financial NLP and Indonesian language processing in several ways:

a. Enhanced dataset: We present a comprehensive dataset of 9,819 manually labeled Indonesian financial news headlines spanning 15 months (January 2024–March 2025), significantly larger than previous Indonesian financial sentiment datasets.
b. Systematic methodology: We introduce a rigorous approach combining stratified splitting with Jaccard similarity-based vocabulary balancing, ensuring representative train-test distributions for robust model evaluation.
c. Performance advancement: Our optimized IndoBERT-large model achieves 94.20% accuracy, representing a 4-16 percentage point improvement over existing Indonesian financial sentiment models, approaching the performance of general-domain sentiment classifiers.
d. Practical framework: We provide empirical evidence for optimal hyperparameter configurations (90:10 split ratio, 2e-6 learning rate) that can guide practitioners in developing production-grade financial sentiment monitoring systems.

Related work: financial sentiment analysis has evolved from dictionary-based approaches to sophisticated deep learning models. Early work by Schumaker and Chen [12] demonstrated the potential of textual analysis for stock market prediction using breaking financial news. Recent advances leverage transformer-based architectures for superior performance. Gu *et al*. [2] integrated FinBERT with LSTM networks for stock price prediction using news sentiment, achieving significant improvements over baseline models. Ruan and Jiang [13] combined FinBERT-enhanced sentiment with SHAP explainability and differential privacy for production-grade forecasting systems. The application of large language models to financial sentiment has gained momentum. Kirtac and Germano [14] benchmarked multiple LLMs including BERT and FinBERT against dictionary methods for trading strategy development, demonstrating the superiority of transformer-based approaches. Kumar [15] proposed a DistilBERT-GRU hybrid architecture for financial headline sentiment, achieving high accuracy through multimodal fusion. These studies emphasize the importance of domain-specific adaptation and hyperparameter optimization for financial text classification.

The transformer architecture, introduced by Vaswani *et al*. [16], revolutionized NLP through self-attention mechanisms that capture long-range dependencies without recurrence. BERT extended this foundation through bidirectional pre-training on masked language modeling and next sentence prediction tasks [5]. Subsequent research has focused on improving BERT's efficiency and effectiveness through architectural modifications and training strategies. Recent studies have investigated BERT's architectural components and optimization strategies. Kokab *et al*. [6] surveyed transformer-based models for sentiment analysis, highlighting the importance of fine-tuning strategies and evaluation metrics. Tang *et al*. [7] explored BERT fine-tuning for multi-label sentiment analysis in imbalanced datasets, demonstrating the effectiveness of data augmentation and ensemble methods. Lee *et al*. [17] introduced the StockEmotions dataset with 12 fine-grained emotion classes, establishing transformer baselines for financial sentiment classification.

Indonesian language processing presents unique challenges due to its agglutinative morphology, extensive affixation, and rich linguistic diversity. Wilie *et al*. [18] introduced IndoNLU, a comprehensive benchmark for Indonesian natural language understanding, demonstrating the need for language-specific models. IndoBERT, a monolingual BERT model pre-trained on Indonesian corpora, has shown superior performance compared to multilingual models for Indonesian tasks. Domain-specific adaptation of Indonesian language models has emerged as a critical research direction. Maharani *et al*. [19] demonstrated the benefits of post-training IndoBERT on Indonesian financial corpora for sentiment and topic classification

tasks. Their work showed that domain-specific post-training significantly improves performance on financial NLP tasks compared to general-domain models. However, systematic hyperparameter optimization and vocabulary-balanced splitting strategies for Indonesian financial sentiment remain underexplored.

Effective transfer learning requires careful hyperparameter tuning to balance pre-trained knowledge retention and task-specific adaptation. Batra *et al*. [3] demonstrated the effectiveness of Bayesian optimization (Optuna) for tuning transformer-based sentiment-driven stock prediction models. Their work showed that systematic hyperparameter search significantly improves downstream performance compared to default configurations. Learning rate selection represents a critical factor in BERT fine-tuning. Excessively high learning rates can cause catastrophic forgetting of pre-trained representations, while overly conservative rates may result in underfitting. Jiang and Zeng [20] compared various FinBERT fine-tuning strategies against BERT and classical models, emphasizing the importance of learning rate scheduling and early stopping mechanisms. These findings underscore the need for systematic experimentation across multiple hyperparameter configurations.

Previous work on Indonesian financial sentiment classification has achieved limited accuracy. Adhi *et al*. [8] applied machine learning approaches to Indonesian stock news sentiment, achieving 78% accuracy using traditional feature extraction methods. Tyas *et al*. [9] employed support vector machines for Indonesian stock market news sentiment analysis, reaching 81% accuracy. In contrast, Alifi *et al*. [21] reported that IndoBERT, when implemented within the IndoNLU framework and tested on a dataset of 7,685 stock news headlines, achieved an accuracy of approximately 90%. These studies highlight persistent challenges in Indonesian financial sentiment analysis: limited dataset sizes, lack of standardized annotation protocols, absence of systematic hyperparameter tuning, and insufficient comparative analysis across different model configurations. Our work addresses these gaps through a comprehensive methodology combining large-scale dataset construction, vocabulary-balanced splitting, and systematic optimization.

2. METHOD

2.1. Research design

This study employs a quantitative experimental design to evaluate IndoBERT-Large for Indonesian financial sentiment classification. Figure 1 illustrates our research workflow, encompassing data collection, preprocessing, model training, and evaluation phases. We systematically compare multiple configurations to identify optimal hyperparameters and splitting strategies for robust sentiment classification.

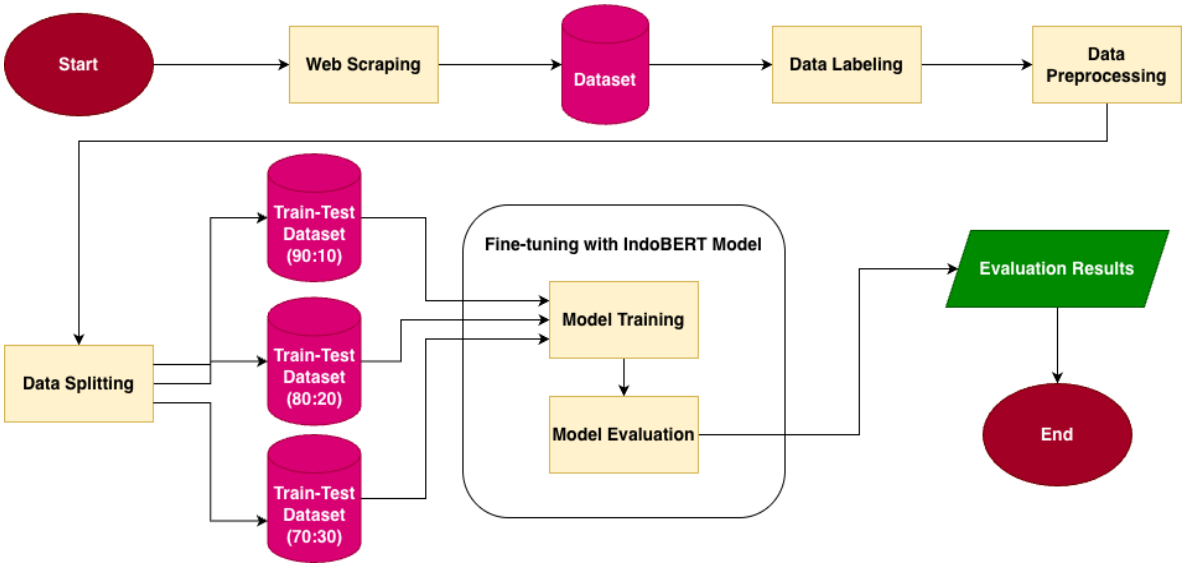


Figure 1. Research flow

2.2. Experimental setup

To facilitate reproducibility, the experimental work was conducted using a standardized computational environment. The primary development platform was Python 3.10 within a cloud-based environment (Google Colab Pro+ / Kaggle Kernels) equipped with an NVIDIA Tesla T4 GPU (16GB

VRAM) and 32GB of system RAM. We utilized the Hugging Face Transformers library for model implementation. All source code, including the web scraping scripts (BeautifulSoup, Selenium) and the fine-tuning pipelines, is maintained in a version-controlled repository to ensure that other scientists can replicate the findings with identical configurations.

2.3. Dataset construction

2.3.1. Data collection

We collected Indonesian financial news headlines from CNBC Indonesia (www.cnbcindonesia.com) covering the period from January 2024 to March 2025. CNBC Indonesia was selected as the primary data source due to its reputation as a leading financial news platform in Indonesia, comprehensive coverage of stock market developments, and consistent publication frequency. We employed web scraping techniques using Python libraries (BeautifulSoup, Selenium) to systematically extract headlines from the market news section. The data collection process yielded 9,819 headlines, representing a substantial increase over previous Indonesian financial sentiment datasets. Adhi *et al*. [8] used 1,200 headlines, while Tyas *et al*. [9] analyzed 2,500 headlines. Our larger dataset enables more robust model training and better generalization to diverse financial news patterns.

2.3.2. Annotation protocol

Each headline was manually annotated by two independent annotators with experience in stocks market and Indonesian language. The annotation guidelines defined three sentiment classes:

a. Positive: Headlines indicating favorable market conditions, company growth, positive economic indicators, or bullish investor sentiment (e.g., "*IHSG menguat 1.2% ditopang kenaikan saham perbankan*" / "IHSG strengthens 1.2% supported by banking stock increases").
b. Neutral: Headlines presenting information without clear positive or negative implications, including announcements, statistics, and balanced reports (e.g., "*Bank Indonesia pertahankan suku bunga acuan* 6%" / "Bank Indonesia maintains reference rate at 6%").
c. Negative: Headlines suggesting unfavorable conditions, company losses, economic concerns, or bearish sentiment (e.g., "*Rupiah melemah tertekan sentimen global*" / "Rupiah weakens pressured by global sentiment").

Inter-annotator agreement was measured using Cohen's kappa coefficient, yielding $\kappa = 0.87$, indicating strong agreement. Disagreements were resolved through discussion and consultation with a third expert annotator. The final dataset distribution was: 2,886 positive (29.4%), 4,360 neutral (44.4%), and 2,573 negative (26.2%) headlines, reflecting the natural distribution of financial news sentiment.

2.4. Data preprocessing

Text preprocessing followed established practices for BERT-based models while preserving Indonesian linguistic features:

a. Lowercase conversion: all text converted to lowercase to reduce vocabulary size and improve generalization.
b. Whitespace normalization: multiple consecutive spaces replaced with single spaces; leading and trailing whitespace removed.
c. Indonesian-specific processing: reduplicated words (e.g., "*besar-besar*" / very big) and affixes preserved as they carry semantic meaning in Indonesian.

We deliberately avoided aggressive preprocessing techniques such as stemming or lemmatization, as BERT's WordPiece tokenization effectively handles morphological variations. This approach aligns with best practices for transformer-based models Tang *et al*. [7], and preserves the rich morphological information inherent in Indonesian text. In summary, the preprocessing procedure consisted of letter lowercasing, removal of currency symbols and special characters, adjustment of tokenization using the IndoBERT-Large tokenizer.

2.5. Data splitting strategy

2.5.1. Stratified shuffle split

We employed stratified shuffle split to maintain class distribution balance across train-test partitions. Three split ratios were evaluated: 90:10, 80:20, and 70:30 (train:test). Stratification ensures that each partition contains approximately the same percentage of samples from each class as the complete dataset, preventing class imbalance issues that could bias model evaluation. The choice of multiple split ratios enables analysis of the trade-off between training data size and test set representativeness. Larger training sets (90:10) provide more examples for learning but smaller test sets for evaluation, while smaller training sets (70:30) offer larger test sets but potentially reduced model capacity due to limited training data.

2.5.2. Jaccard similarity for vocabulary balance

To ensure vocabulary distribution balance between training and test sets, we computed Jaccard similarity coefficients between the unique word vocabularies of each partition using formula in (1).

$$J(V_{train}, V_{test}) = \frac{|V_{train} \cap V_{test}|}{|V_{train} \cup V_{test}|} \quad (1)$$

V_{train} and V_{test} represent the unique vocabularies of training and test sets, respectively. Jaccard similarity ranges from 0 (no overlap) to 1 (identical vocabularies). Higher Jaccard scores indicate better vocabulary coverage in the test set, reducing the risk of out-of-vocabulary (OOV) issues during evaluation [22]. We performed 10 random splits for each ratio and selected the split with the highest Jaccard similarity while maintaining stratified class distribution. This approach ensures that the test set contains vocabulary representative of the training distribution, enabling fair evaluation of model generalization. Table 1 presents the Jaccard similarity scores for selected splits.

Table 1. Dataset splitting results using stratified shuffle split

Dataset ratio	Train data label distribution				Test data label distribution				Jaccard similarity
	Positive	Neutral	Negative	Total	Positive	Neutral	Negative	Total	
70:30	2021	3049	1803	6873	866	1307	773	2946	0.534
80:20	2309	3485	2061	7855	578	871	515	1964	0.478
90:10	2598	3920	2319	8837	289	436	257	982	0.367

The Jaccard similarity scores indicate good vocabulary overlap across all splits, with larger test sets naturally achieving higher scores due to increased vocabulary coverage. The stratified approach successfully maintains identical class distributions (train/test) across all configurations.

2.6. Model architecture

We employed IndoBERT-Large, a monolingual BERT model pre-trained on 23GB of Indonesian text from diverse sources including news articles, social media, and web documents [18]. IndoBERT-Large architecture specifications:

a. Layers: 24 transformer encoder layers
b. Hidden size: 1,024 dimensions
c. Attention heads: 16 multi-head attention mechanisms per layer
d. Parameters: ~340 million trainable parameters
e. Vocabulary: 40,000 WordPiece tokens optimized for Indonesian morphology
f. Maximum sequence length: 512 tokens (we used 128 for efficiency)

IndoBERT-large was chosen over multilingual models (mBERT, XLM-RoBERTa) based on prior research, as its monolingual pre-training enables it to better capture Indonesian linguistic nuances compared to models that share vocabulary and parameters across multiple languages [18].

2.7. Evaluation

Model performance was assessed using standard classification metrics. Several evaluation metrics were calculated to assess the performance of the IndoBERT model in classifying stock news sentiment. These include,

a. Accuracy: This metrics measures the proportion of correct predictions over the total number of predictions made. Accuracy can be calculated using (2).

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

b. Precision: This metrics calculates the proportion of correct for a class out of all instances predicted. Precision can be calculated using (3).

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

c. Recall: This metrics also known as sensitivity or true positive rate, measures how well the model identifies all actual positive instances of a class. Recall can be calculated using (4).

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

d. F1-Score: This metric is the harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives, and can be calculated using (5).

$$F1 - Score = 2\ x\ \frac{Precission\ x\ Recall}{Precission + Recall} \quad (5)$$

3. RESULTS AND DISCUSSION

The IndoBERT model was fine-tuned using four learning rate values (2e-6, 3e-6, 2e-5, and 3e-5) across three train-test split ratios: 90:10, 80:20, and 70:30. Each configuration was evaluated using four performance metrics: accuracy, precision, recall, and F1-score. The training process employed an early stopping mechanism with a patience of 3 to prevent overfitting and ensure optimal generalization.

3.1. Model training and evaluation results

3.1.1. Dataset ratio 70:30

Testing with a 70:30 dataset ratio allows for a more rigorous and realistic assessment of the model's performance on unseen data. The training and evaluation model result for dataset with this ratio are illustrated in Figure 2 for each learning rate configuration.

The model with a 2e-6 learning rate reached its best validation accuracy of 92.40% at epoch 8 with a validation loss of 20.53 as shown in Figure 2(a). Training accuracy was 97.90% with a loss of 14.86, and later epochs showed only a slight accuracy drop and minimal loss increase, indicating stable learning with mild overfitting. The 3e-6 model achieved the highest score for this split, reaching 92.80% at epoch 6 (validation loss 20.55) as shown in Figure 2(b), and maintained the same accuracy until the final epoch with only a small loss rise, demonstrating excellent stability. The 2e-5 model peaked at 91.96% at epoch 2 (loss 21.36) but later declined to 89.75% with a higher loss, indicating early overfitting as shown in Figure 2(c). The 3e-5 model achieved 90.63% at epoch 4 (loss 32.54) but showed no meaningful improvement afterward, with high validation loss and stagnant accuracy as shown in Figure 2(d), reflecting poor generalization.

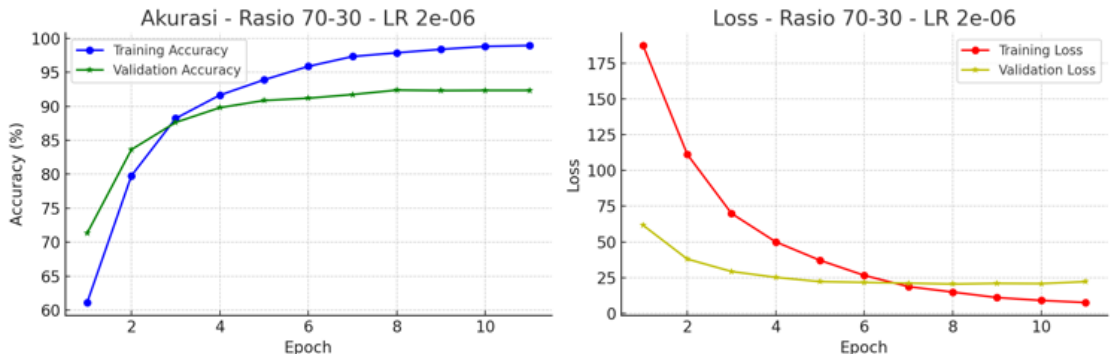
3.1.2. Dataset ratio: 80:20

The 80:20 dataset ratio represents a balanced configuration between the amount of training and testing data. The training and evaluation model result for dataset with this ratio are illustrated in Figure 3. For the 2e-6 learning rate, the model showed the most stable and optimal performance, reaching 93.38% validation accuracy at epoch 11 with a validation loss of 14.44 Figure 3(a). Training accuracy was 98.99% with a loss of 8.06, and both accuracy and loss trends indicated minimal overfitting. This configuration was the most effective for the 80:20 split. The 3e-6 model achieved 91.80% accuracy at epoch 5 (loss 15.22), as shown in Figure 3(b). Although slightly lower than 2e-6, it remained stable, ending with 91.50% accuracy and only a small loss increase. The 2e-5 learning rate peaked at 90.63% at epoch 3 (loss 22.58), but accuracy and loss worsened in the following epoch, indicating early overfitting Figure 3(c). The highest learning rate, 3e-5, reached 90.84% at epoch 2 (loss 17.65) as seen in Figure 3(d). Training accuracy increased afterward, but validation accuracy stagnated, and loss did not improve, showing rapid overfitting and weak generalization.

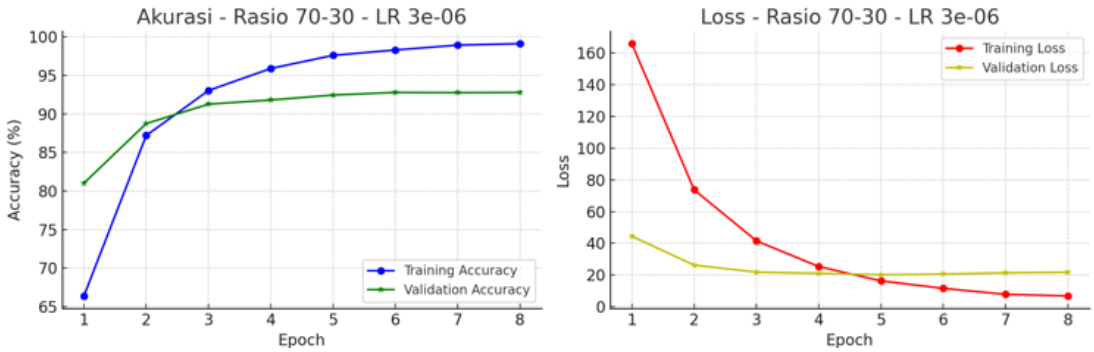
3.1.3. Dataset ratio 90:10

Testing with a 90:10 training-to-testing data ratio allocates a larger portion of data for model learning, which is expected to enhance its ability to accurately capture sentiment patterns within the training set. The training and evaluation model result for dataset with this ratio are illustrated in Figure 4 for each learning rate configuration.

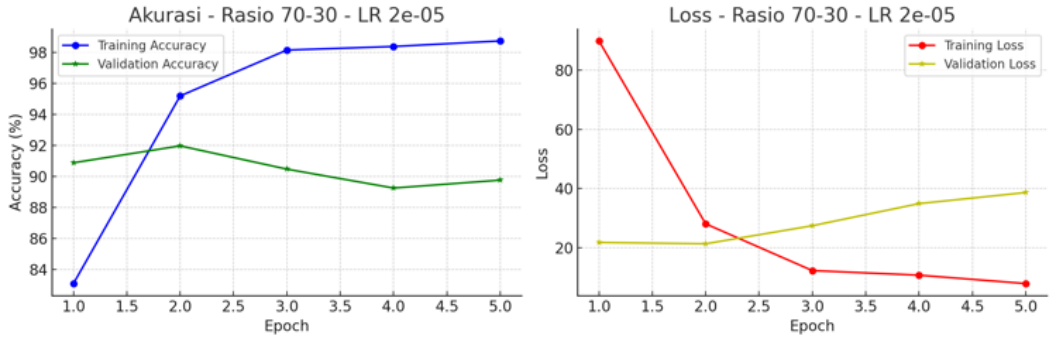
For the 2e-6 learning rate, the model achieved the highest validation accuracy of 94.20% at epoch 10 with a validation loss of 6.52 as shown in Figure 4(a). Training accuracy rose steadily from 68.77% to 99.45%, and loss dropped sharply, indicating a consistent learning pattern that plateaued after epoch 10. The 3e-6 model reached 93.69% accuracy at epoch 7 (loss 6.57), with training accuracy increasing from 68.76% to 98.95% and loss decreasing accordingly as shown in Figure 4(b). This reflects stable and effective learning. The 2e-5 model peaked at 93.38% at epoch 3 (loss 6.65), but validation performance declined afterward despite rising training accuracy, showing early overfitting as shown in Figure 4(c). The 3e-5 model recorded 92.46% at epoch 1 (loss 6.35), yet validation accuracy stagnated in later epochs, and loss did not improve, indicating rapid overfitting and poor generalization as shown in Figure 4(d).



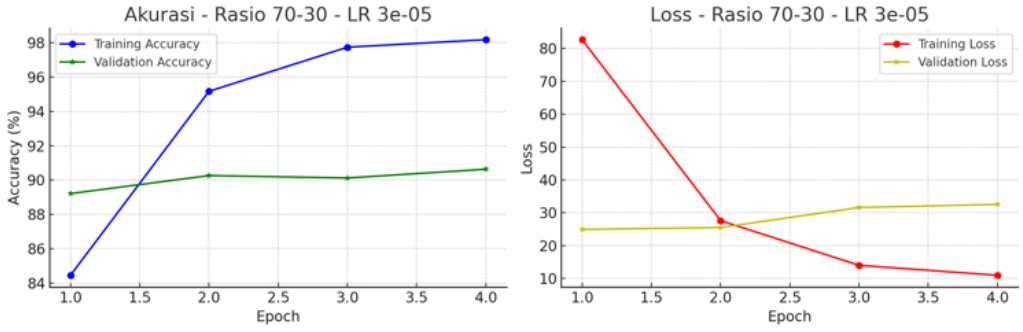
(a)



(b)

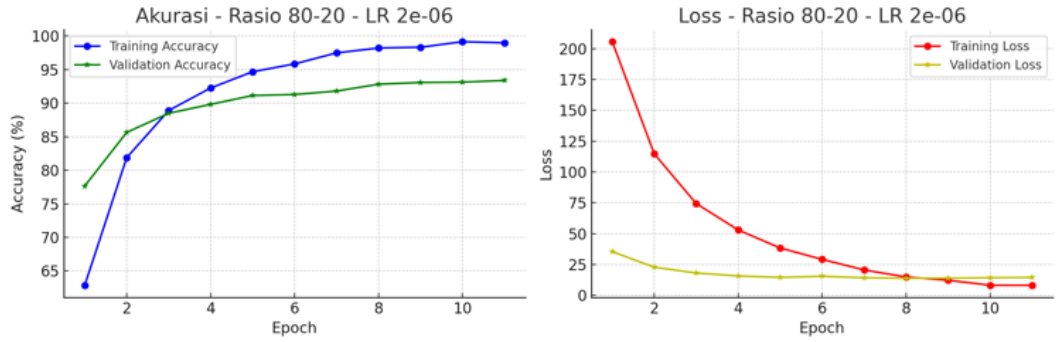


(c)

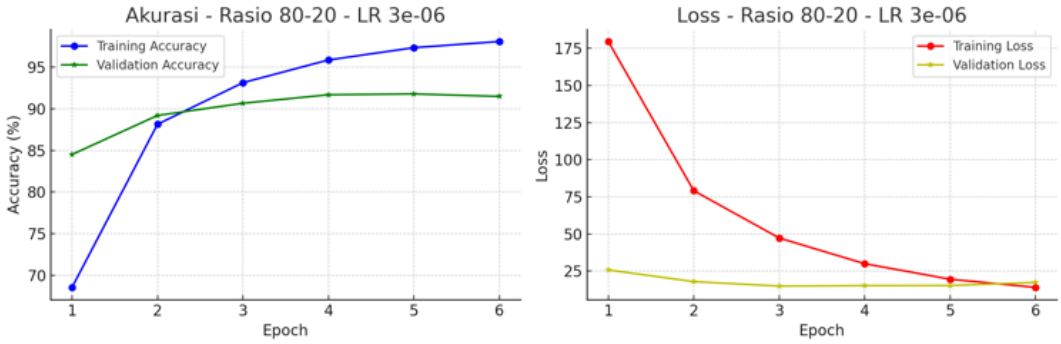


(d)

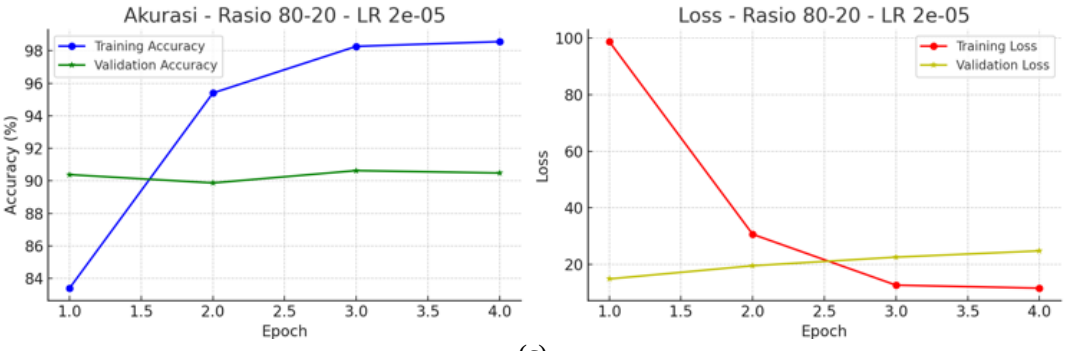
Figure 2. Training and evaluation model result on dataset with ratio 70:30; (a) 2e−6, best val acc 92.40% at epoch 8, mild overfitting, (b) 3e−6, best val acc 92.80% at epoch 6, excellent stability, (c) 2e−5, peak 91.96% at epoch 2, early overfitting, and (d) 3e−5, peak 90.63% at epoch 4, poor generalization



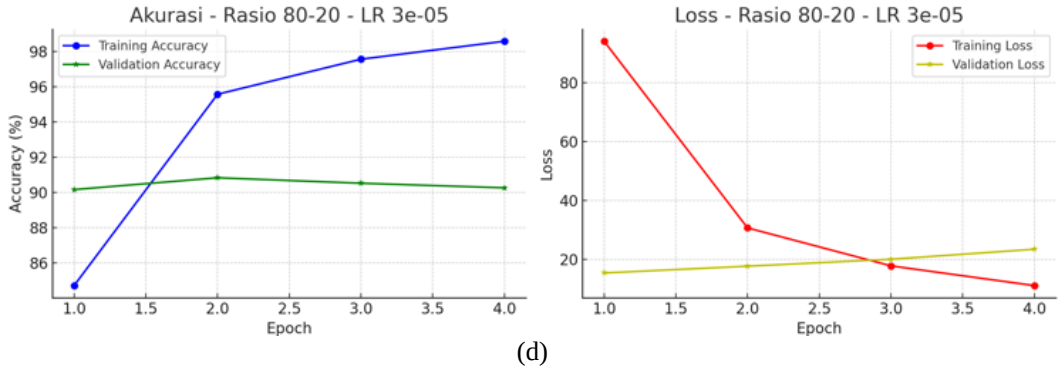
(a)



(b)



(c)



(d)

Figure 3. Training and evaluation model result on dataset with ratio 80:20; (a) 2e−6, best val acc 93.38% at epoch 11, most stable and optimal, (b) 3e−6, best val acc 91.80% at epoch 5, stable with slight drop, (c) 2e−5, peak 90.63% at epoch 3, early overfitting, and (d) 3e−5, peak 90.84% at epoch 2, rapid overfitting and weak generalization

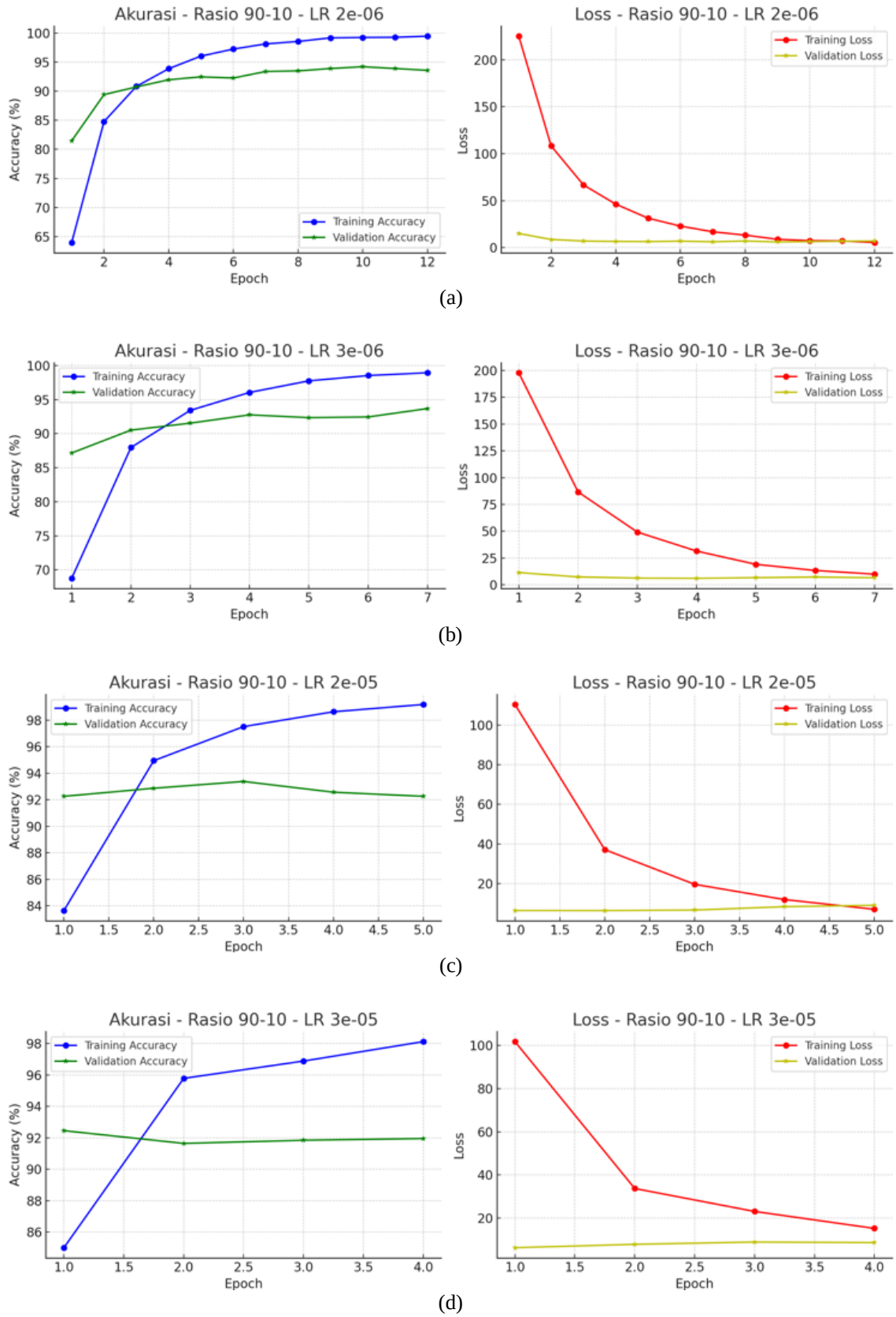


Figure 4. Training and evaluation model result on dataset with ratio 90:10; (a) 2e−6, best val acc 94.20% at epoch 10, consistent learning, (b) 3e−6, best val acc 93.69% at epoch 7, stable and effective, (c) 2e−5, peak 93.38% at epoch 3, early overfitting, (d) 3e−5, peak 92.46% at epoch 1, rapid overfitting and poor generalization

3.2. Comparison with existing benchmarks

Our fine-tuned IndoBERT-Large model demonstrates superior performance compared to several recent benchmarks in Indonesian financial sentiment analysis. As summarized in Table 2, our optimal configuration reached 94.20% accuracy, significantly outperforming the 78% accuracy achieved by Adhi *et al*. [8] using traditional machine learning and the 81% reported by Tyas *et al*. [9] with SVM-based classifiers. Furthermore, our results exceed the 90% accuracy benchmark established by Alifi *et al*. [21], who utilized a smaller dataset of 7,685 headlines. These improvements can be attributed to several factors: (1) the larger and more representative dataset of 9,819 CNBC Indonesia headlines; (2) the systematic hyperparameter search that identified 2e-6 as the optimal learning rate; and (3) the vocabulary-balanced splitting strategy using Jaccard similarity, which ensures consistent feature distribution between training and testing sets. The goal of this evaluation was to identify the parameter combination that offers the best generalization performance on unseen data, as summarized in Table 2.

Table 2. Model evaluation results comparation

Dataset Ratio	Learning Rate	Epoch	Accuracy	Precision	Recall	F1-Score	Notes
70:30	2e-6	8	92,40%	92,40%	92,40%	92,40%	Stable, balanced
	3e-6	6	92,80%	92,81%	92,80%	92,80%	
	2e-5	2	91,96%	91,99%	91,96%	91,96%	
	3e-5	4	90,63%	90,66%	90,63%	90,63%	
80:20	2e-6	11	93,38%	93,45%	93,38%	93,38%	Best for mid split
	3e-6	5	91,80%	92,20%	91,80%	91,80%	
	2e-5	3	90,63%	91,25%	90,63%	90,60%	
	3e-5	2	90,84%	91,25%	90,84%	90,82%	
90:10	2e-6	10	94,20%	94,25%	94,20%	94,19%	Best overall
	3e-6	7	93,69%	93,75%	93,69%	93,69%	
	2e-5	3	93,38%	93,43%	93,38%	93,38%	
	3e-5	1	92,46%	92,60%	92,46%	92,48%	

Among all configurations tested, the smallest learning rate (2e-6) consistently provided the best overall performance across almost all dataset ratios. This consistent performance demonstrates that smaller learning rates facilitate deeper and more gradual learning, enabling the model to better generalize to unseen data. Similarly, the model with a learning rate of 3e-6 showed strong and stable performance, particularly notable for the 70:30 split, where it achieved the highest accuracy of 92.80% with balanced metrics overall. In contrast, larger learning rates (2e-5 and 3e-5) led to faster convergence but resulted in instability, earlier overfitting, and lower validation accuracy. Thus, a balance between learning rate size and the number of epochs is crucial for optimal model performance, with smaller learning rates generally benefiting from extended training durations. Recent studies show that integrating public sentiment with historical data using deep learning methods can significantly improve stock return and price prediction in emerging markets [23, 24]. Furthermore, comparisons between lexicon-based methods and BERT models in the government debt market indicate the robustness of transformer architectures [25].

4. CONCLUSION

This investigation established a new performance benchmark for IndoBERT-based sentiment classification of Indonesian financial news: 94.20% accuracy achieved through 90:10 data partitioning and 2e-6 learning rate configuration. This result substantially surpasses prior research performance baselines. Two principal conclusions emerge from this analysis: i) hyperparameter optimization, particularly learning rate selection, critically influences classification performance; and ii) training data volume combined with appropriate train-test partitioning substantially enhances model generalization capacity.

Future work can expand this study by adding named entity recognition to identify key financial entities and aspect-based sentiment analysis to capture more detailed, entity-specific sentiment. Development of real-time classification pipelines would extend applicability to analysts and investment professionals requiring immediate sentiment assessment. Systematic investigation of transfer learning approaches to financial domains represents an additional avenue for methodological advancement. With continued development, IndoBERT-based systems demonstrate substantial potential as foundational infrastructure for financial sentiment monitoring applications.

ACKNOWLEDGMENTS

The authors would like to express sincere gratitude to the developers of the IndoNLU framework and IndoBERT models, which served as the foundation of this research. The dataset used in this study, consisting of manually labeled Indonesian stock market news headlines, was sourced from publicly available

articles on CNBC Indonesia. The content was used strictly for academic research. To promote transparency and reproducibility, the dataset has been made publicly accessible.

FUNDING INFORMATION

The authors declare that this research received no specific research grant or contract from any funding agency in the public, commercial, or not-for-profit sectors.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the contributor roles taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Tri Agung Jiwandono	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
MS Hendriyawan Achmad	✓	✓		✓	✓	✓	✓	✓		✓				✓
Suhirman										✓				✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available in the Kaggle repository at https://www.kaggle.com/datasets/triagungj/cnbc-indonesia-stock-news-sentiment-dataset.

REFERENCES

[1] R. A. Brealey, S. C. Myers, and F. Allen, *Principles of corporate finance*. McGraw-Hill Education, 2011.

[2] W. Jun Gu, Y. Hao Zhong, S. Zun Li, C. Song Wei, L. Ting Dong, and Z. Yue Wang, "Predicting Stock Prices with FinBERT-LSTM: Integrating News Sentiment Analysis," in *Proceedings of the 2024 8th International Conference on Cloud and Big Data Computing*, in ICCBDC '24. New York, NY, USA: Association for Computing Machinery, 2024, pp. 67–72. doi: 10.1145/3694860.3694870.

[3] V. Batra, Suman, and R. K. Tipu, "Deep Learning and Optimization Techniques for Sentiment-Driven Stock Market Prediction Using Transformer Models," in *2025 2nd International Conference on Multidisciplinary Research and Innovations in Engineering (MRIE)*, 2025, pp. 433–438. doi: 10.1109/MRIE66930.2025.11156446.

[4] B. Liu, *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, 2nd ed. Cambridge: Cambridge University Press, 2020. doi: 10.1017/9781108639286.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, 2019, doi: 10.18653/v1/N19-1423.

[6] S. Tabinda Kokab, S. Asghar, and S. Naz, "Transformer-based deep learning models for the sentiment analysis of social media data," *Array*, vol. 14, p. 100157, 2022, doi: 10.1016/j.array.2022.100157.

[7] T. Tang, X. Tang, and T. Yuan, "Fine-Tuning BERT for Multi-Label Sentiment Analysis in Unbalanced Code-Switching Text," *IEEE Access*, vol. 8, pp. 193248–193256, 2020, doi: 10.1109/ACCESS.2020.3030468.

[8] A. P. Adhi, K. Umuri, and G. Triyono, "Sentiment Analysis and Entity Detection on News Headlines to Support Investment Decisions," *Jurnal Teknik Informatika (Jutif)*, vol. 5, no. 6, Dec. 2024, doi: 10.52436/1.jutif.2024.5.6.3434.

[9] S. M. P. Tyas, R. Sarno, and B. S. Rintyarna, "Comparative analysis of stock news sentiment classification methods: A machine learning, deep learning, transfer learning, and graph approach," (in Indonesian) *Jurnal Penelitian Ipteks*, vol. 9, no. 1, pp. 58–64, 2024. doi: 10.32528/penelitianipteks.v9i1.1479.

[10] E. Yulianti and N. K. Nissa, "ABSA of Indonesian customer reviews using IndoBERT: single-sentence and sentence-pair classification approaches," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 5, pp. 3579–3589, Oct. 2024, doi: 10.11591/eei.v13i5.8032.

[11] M. B. Nugroho, A. Khanif Zyen, and A. Widiastuti, "Multiclass Sentiment Analysis of Electric Vehicle Incentive Policies Using IndoBERT and DeBERTa Algorithms," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, pp. 2548–6861, 2025, doi: 10.30871/jaic.v9i3.9511.

[12] R. P. Schumaker and H. Chen, "Textual analysis of stock market prediction using breaking financial news: The AZFin text system," *ACM Trans. Inf. Syst.*, vol. 27, no. 2, Mar. 2009, doi: 10.1145/1462198.1462204.

[13] L. Ruan and H. Jiang, "Stock price prediction using FinBERT-Enhanced sentiment with SHAP explainability and differential privacy," *Mathematics*, vol. 13, no. 17, p. 2747, 2025, doi: 10.3390/math13172747.

[14] K. Kirtac and G. Germano, "Enhanced Financial Sentiment Analysis and Trading Strategy Development Using Large Language Models," in *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, O. De Clercq, V. Barriere, J. Barnes, R. Klinger, J. Sedoc, and S. Tafreshi, Eds., Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1–10. doi: 10.18653/v1/2024.wassa-1.1.
[15] R. Kumar, "Financial Sentiment Analysis on Stock Using NLTK, Transformer, and Deep Learning," *Authorea Preprints*, 2025, doi: 10.36227/techrxiv.175624603.30439185/v1.
[16] A. Vaswani *et al.*, "Attention is all you need. InAdvances in Neural Information Processing Systems," 2017, doi: 10.48550/arXiv.1706.03762.
[17] J. Lee, H. L. Youn, J. Poon, and S. C. Han, "StockEmotions: Discover Investor Emotions for Financial Sentiment Analysis and Multivariate Time Series," *arXiv preprint arXiv:2306.02136*, 2023, doi: 10.48550/arXiv.2301.09279.
[18] B. Wilie *et al.*, "IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding," in *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, K.-F. Wong, K. Knight, and H. Wu, Eds., Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. doi: 10.18653/v1/2020.aacl-main.85.
[19] N. P. I. Maharani, A. Purwarianti, Y. Yustiawan, and F. C. Rochim, "Domain-Specific Language Model Post-Training for Indonesian Financial NLP," in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2023, pp. 1–6. doi: 10.1109/ICEEI59426.2023.10346625.
[20] T. Jiang and A. Zeng, "Financial sentiment analysis using FinBERT with application in predicting stock movement," *arXiv preprint arXiv:2306.02136*, 2023, doi: 10.48550/arXiv.2306.02136.
[21] M. R. Alifi, D. C. U. Lieharyani, B. P. Sudimulya, and M. R. Maulidhan, "Implementation of IndoNLU Pre-Trained Model for Aspect-Based Sentiment Analysis of Indonesian Stock News," *Jurnal Teknik Informatika*, vol. 16, no. 2, 2023, doi: 10.15408/jti.v16i2.33791.
[22] F. Farias, T. Ludermir, and C. Bastos-Filho, "Similarity Based Stratified Splitting: an approach to train better classifiers," Oct. 2020, doi: 10.48550/arXiv.2010.06099.
[23] A. R. Nugroho, R. W. Sholikah, H. T. Ciptaningtyas, M. Husni, and A. S. Indrawanti, "Integrating Public Sentiment on Stock Return Percentage Prediction in Emerging Markets with a Deep Learning Approach," in *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, 2025, pp. 1–7. doi: 10.1109/SIML65326.2025.11081119.
[24] Ardisurya and M. Rizkinia, "Implementation of Diffusion Variational Autoencoder for Stock Price Prediction with the Integration of Historical and Market Sentiment Data," *International Journal of Electrical, Computer, and Biomedical Engineering*, vol. 2, no. 2, Jun. 2024, doi: 10.62146/ijecbe.v2i2.55.
[25] F. Rachmawati, U. Azmi, and R. Azwarini, "Comparison of Lexicon-Based Methods and Bidirectional Encoder Representations for Transformers Models in Sentiment Analysis of Government Debt Market Movements," *International Journal of Engineering and Computer Science Applications (IJECSA)*, vol. 4, no. 1, 2025, doi: 10.30812/ijecsa.v4i1.4832.

BIOGRAPHIES OF AUTHORS

Tri Agung Jiwandono was born in Banyumas, Central Java, Indonesia, in 2000. He received the Bachelor's degree in Informatics from Universitas Teknologi Yogyakarta in 2023, and the Master's degree in Information Technology from Universitas Teknologi Yogyakarta in November 2025. His research interests include natural language processing (NLP), particularly sentiment analysis. He is currently a self-employed freelance software engineer. He can be contacted at email: tj.triagungj@gmail.com.

MS Hendriyawan Achmad received his Engineer degree in Electrical Engineering from Institute of Science & Technology AKPRIND in 2006, and the Master's degree in Electrical Engineering from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2013. He earned his PhD in Electrical and Electronics Engineering from Universiti Malaysia Pahang, Malaysia, in 2020. Currently, he is a Senior Lecturer at the University of Technology Yogyakarta, Indonesia. His research interests include human-robot interaction, robotic vision, Internet of Things (IoT), and biomedical signal processing. He can be contacted at email: hendriyawanachmad@uty.ac.id.

Prof. Suhirman, S.Kom., M.Kom., Ph.D. received his Bachelor's degree in Information Technology from STMIK Akakom Yogyakarta, Indonesia, in 1998, and the Master's degree in Computer Science from Universitas Gadjah Mada, Yogyakarta, Indonesia, in 2004. He earned his PhD in Computer Science from Universiti Malaysia Pahang, Malaysia, in 2016. Currently, he is a Professor and Head of the Master's Program in Information Technology at Universitas Teknologi Yogyakarta, Indonesia. His research interests include Information Systems, Data Warehouse and Data Mining, and Artificial Intelligence. He can be contacted at email: suhirman@uty.ac.id.