JURNAL TEKNOLOGI DAN OPEN SOURCE
Vol. 9, No. 1, June 2026, pp. 164~176
e-ISSN: 2622-1659, accredited Four Grade by Kemenristekdikti, Decree No: 152/E/KPT/2023
DOI: https://doi.org/10.36378/jtos.v9i1.5583

Benchmarking IndoBERT and Multilingual BERT for Indonesian Financial News Sentiment Classification

Matius Ivan Bimasena[1], I Kadek Yogi Prayoga[2], I Gusti Agung Putu Mahendra[3], Purnama Sidik[4]

[1,2,4]Department of Information Technology, Institut Teknologi dan Bisnis STIKOM Bali, Indonesia
[3]Department of Informatic Engineering, Politeknik Negeri Bengkalis, Indonesia

Article Info

Article history:

Received 04 20, 2026
Revised 05 16, 2026
Accepted 06 10, 2026

Keywords:

Financial news;
Classification;
Transformer-based models;
IndoBERT;
Multilingual BERT

ABSTRACT

Financial news sentiment classification was important for understanding market narratives and investor perception, but Indonesian financial news remained challenging because it contained domain-specific terminology, numerical expressions, and imbalanced sentiment categories. This study benchmarked two transformer-based models, IndoBERT and Multilingual BERT, for classifying Indonesian financial news sentiment into negative, neutral, and positive classes. The dataset consisted of economic and financial news articles from Kontan, CNBC, and Bisnis.com during the first quarter of 2026. After preprocessing, 3,366 articles were used, consisting of 3,070 neutral, 184 negative, and 112 positive articles. The dataset was divided into training, validation, and testing sets using stratified splitting. Class weighting was applied to reduce the effect of class imbalance. The results showed that IndoBERT achieved the best overall performance, with 0.94 accuracy and 0.71 macro F1-score, while Multilingual BERT achieved 0.93 accuracy and 0.70 macro F1-score. These findings indicated that IndoBERT was more suitable for Indonesian financial news sentiment classification, although Multilingual BERT remained competitive, especially in detecting positive sentiment.

Corresponding Author:
Matius Ivan Bimasena
Department of Information Technology,
Institut Teknologi dan Bisnis STIKOM Bali, Indonesia
Email: ivan_bimasena@stikom-bali.ac.id

1. Introduction

Media sentiment has become an important component in financial market analysis because news narratives can influence investor perception, market confidence, and trading behavior[1]. In the digital financial ecosystem, investors are continuously exposed to large volumes of online news related to stock movements, macroeconomic conditions, corporate actions, government policies, investment risks, and market expectations. [2] These narratives may contain positive, negative, or neutral sentiment that can shape public interpretation of economic and financial conditions[3]. Therefore, sentiment classification of financial news has become an important natural language processing task, especially for understanding how media narratives represent stock market and economic issues [4].

In Indonesia, financial and economic news is widely disseminated through online media platforms such as Kontan, CNBC Indonesia, and Bisnis.com. These media outlets provide continuous information related to investment, capital markets, listed companies, stock exchange activity, and economic policy. However, the large volume of financial news makes manual sentiment analysis inefficient, time-consuming, and difficult to scale. In addition, financial sentiment is often context-dependent. Words or phrases that appear neutral in general language may carry positive or negative implications in a financial context,

especially when associated with stock prices, investor actions, interest rates, exchange rates, commodity movements, or corporate performance [5].

The main problem addressed in this study is the need for a reliable benchmarking framework for Indonesian financial news sentiment classification using transformer-based natural language processing models. Although sentiment analysis has been widely studied, financial news sentiment classification remains challenging because financial texts contain domain-specific terminology, numerical expressions, implicit sentiment, and market-specific narratives. In addition, Indonesian financial sentiment datasets are still limited, particularly those focusing on stock market-related news from major financial media sources. These challenges require models that can capture Indonesian linguistic structures and contextual financial meaning. [6], [7].

Transformer-based language models have significantly improved text classification tasks because they can learn contextual word representations from large-scale text data. In this study, two transformer-based models were evaluated: IndoBERT and Multilingual BERT. IndoBERT represents an Indonesian-specific pre-trained language model, while Multilingual BERT represents a multilingual transformer model that supports Indonesian as one of its languages. Comparing these two models is important because it provides insight into whether a language-specific transformer model is more effective than a multilingual model for Indonesian financial news sentiment classification [8], [9].

In addition to IndoBERT, Multilingual BERT is also relevant for this study because it was pre-trained on multiple languages, including Indonesian [10]. Multilingual BERT provides a broader language representation and can be used as a comparison model to evaluate whether an Indonesian-specific model performs better than a multilingual transformer model [11]. Comparing these two models is important because Indonesian financial news classification requires both language understanding and contextual interpretation. IndoBERT may have advantages in Indonesian linguistic representation, while Multilingual BERT may benefit from multilingual pretraining and broader cross-lingual knowledge[12].

However, there is still a research gap in benchmarking Indonesian-specific and multilingual transformer models for Indonesian financial news sentiment classification. Many sentiment analysis studies focus on general-domain text, such as product reviews, social media comments, or public opinion data[13]. Financial news has different linguistic characteristics because it contains formal journalistic language, financial terminology, numerical expressions, and market-specific narratives. Moreover, Indonesian financial sentiment datasets are still limited, especially datasets that focus on stock market-related news from major Indonesian financial media outlets [14], [15], [16].

The dataset used in this study consisted of Indonesian economic and financial news from the first quarter of 2026 collected from Kontan, CNBC Indonesia, and Bisnis.com. The dataset included article titles, article content, media sources, sentiment labels, polarity scores, and confidence scores. The sentiment labels were generated automatically by a machine learning model; therefore, they were treated as pseudo-labels rather than manually verified ground truth. This condition makes the study relevant for weakly supervised sentiment classification, where machine-generated labels are used to train and evaluate classification models.

To address this problem, this study proposes a benchmarking experiment using two transformer-based models: IndoBERT and Multilingual BERT. The models are trained and evaluated on the same Indonesian financial news dataset to classify sentiment into three categories: negative, neutral, and positive. The experiment uses a supervised text classification framework with stratified data splitting, class weighting to handle class imbalance, and evaluation metrics such as accuracy, precision, recall, macro F1-score, and weighted F1-score. Macro F1-score is emphasized because the dataset is highly imbalanced, with neutral sentiment dominating the class distribution.

The novelty of this research lies in the comparative evaluation of Indonesian-specific and multilingual transformer models for financial news sentiment classification in the Indonesian language. Unlike general sentiment analysis studies, this research focuses on financial and stock market news, where sentiment interpretation is more domain-sensitive. Furthermore, this study explicitly considers the use of machine-generated sentiment labels as pseudo-labels, making it relevant for weakly supervised sentiment classification. By comparing IndoBERT and Multilingual BERT under the same experimental setting, this research provides insight into whether an Indonesian-specific transformer model is more suitable than a multilingual transformer model for Indonesian financial news sentiment classification.

The expected contribution of this study is twofold. First, it provides an experimental benchmark for Indonesian financial news sentiment classification using two transformer-based models. Second, it offers methodological insight into the use of pseudo-labeled financial news data for natural language processing experiments. The results of this study are expected to support future research in financial sentiment analysis, media narrative analysis, and the development of sentiment-based indicators for Indonesian stock market analysis.

2. Literature Review

Sentiment analysis has become an important task in natural language processing because it enables automatic identification of opinions, attitudes, and emotional polarity in text. In the financial domain, sentiment analysis is more complex than general-domain sentiment classification because financial text often contains technical terms, numerical indicators, implicit meanings, and market-specific expressions. For example, terms related to price correction, foreign net sell, interest rate changes, or commodity fluctuations may contain sentiment implications that require contextual interpretation.

The development of transformer-based models has improved the performance of text classification tasks. BERT introduced bidirectional contextual representation, allowing a model to understand text by considering both left and right contexts simultaneously. This architecture has been widely used for downstream tasks such as sentiment classification, question answering, and natural language inference [17]. For Indonesian natural language processing, IndoBERT was developed as part of Indonesian language understanding resources and has been widely used in Indonesian text classification tasks. Since IndoBERT was pre-trained on Indonesian corpora, it is expected to capture Indonesian linguistic characteristics more effectively than multilingual models[18].

Multilingual BERT is also relevant for Indonesian text classification because it was trained on multiple languages and can process Indonesian text without requiring a monolingual model. However, multilingual models may distribute their representational capacity across many languages, which can reduce their sensitivity to language-specific patterns. Therefore, comparing IndoBERT and Multilingual BERT is important to determine whether an Indonesian-specific model provides better performance in financial news sentiment classification [19].

Previous studies in financial sentiment analysis have shown that domain-specific and language-specific models can provide advantages when dealing with specialized vocabulary and contextual expressions. However, studies focusing on Indonesian financial news sentiment classification are still limited. Most sentiment analysis research in Indonesian focuses on social media, public opinion, product reviews, or general news. This study addresses this gap by benchmarking IndoBERT and Multilingual BERT using Indonesian financial news data and by analyzing model performance under class imbalance conditions.

3. Research Method

This study used a quantitative experimental research design with a benchmarking approach. The objective was to compare two transformer-based models, namely IndoBERT and Multilingual BERT, for Indonesian financial news sentiment classification. The classification task was formulated as a supervised text classification problem, where the input consisted of Indonesian financial news text and the output consisted of three sentiment classes: negative, neutral, and positive.

The use of transformer-based models was selected because these models have demonstrated strong performance across a wide range of natural language processing tasks, particularly in text classification, sentiment analysis, and sequence understanding. Transformer architectures are capable of capturing long-range dependencies within textual data more effectively than many traditional machine learning and recurrent neural network approaches. Their self-attention mechanism allows the model to identify relationships between words and phrases regardless of their positions in a sentence, making them highly suitable for processing complex textual information. In addition, transformer-based models benefit from large-scale pre-training, enabling them to learn rich linguistic representations that can later be adapted to specific downstream tasks through fine-tuning.

BERT introduced bidirectional contextual representation, which enables a language model to understand a word based on both left and right contexts and to be fine-tuned for downstream classification tasks with minimal task-specific architectural changes [20]. This capability allows the model to generate more context-aware representations of text, improving its ability to capture semantic meaning and linguistic nuances. As a result, BERT has become one of the most influential pre-trained language models in modern natural language processing research.

IndoBERT was used because it was developed as part of IndoNLU, a benchmark and resource collection for Indonesian natural language understanding tasks, including Indonesian pre-trained language models [21]. Since IndoBERT was specifically trained on Indonesian-language corpora, it is expected to better understand Indonesian vocabulary, grammar, sentence structures, and contextual expressions compared with models trained primarily on other languages. This makes IndoBERT particularly suitable for research involving Indonesian textual data.

Multilingual BERT was used as a comparison model because it was pre-trained on the top 104 languages using Wikipedia and masked language modeling[22]. The inclusion of Multilingual BERT enables

an evaluation of whether a language-specific model such as IndoBERT provides advantages over a multilingual model when handling Indonesian-language classification tasks.

The research workflow consisted of data acquisition, data inspection, text preprocessing, label encoding, dataset splitting, class imbalance handling, tokenization, model fine-tuning, model testing, and performance evaluation can be see in Figure 1. Each stage was conducted systematically to ensure data quality, improve model learning effectiveness, and provide a reliable assessment of model performance throughout the experimental process.

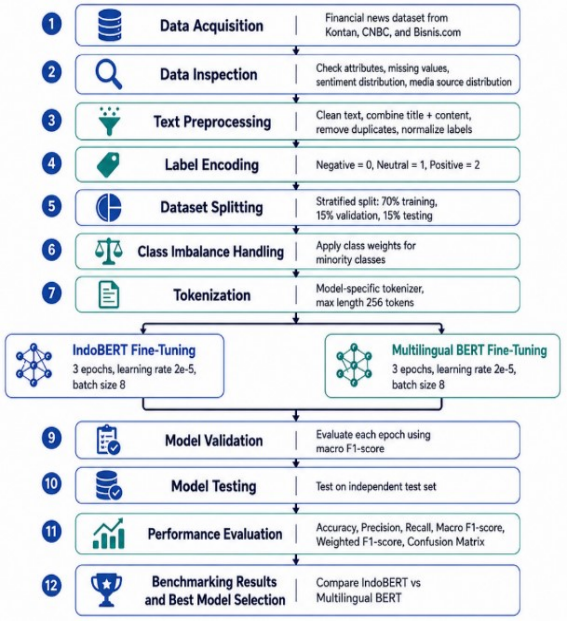

Figure 1. Research Flow

2.1 Data Acquisition

The dataset used in this study was Indonesian economic and financial news sentiment data from the first quarter of 2026. The dataset contained news articles from three Indonesian financial news sources, namely Kontan, CNBC, and Bisnis.com. The raw dataset consisted of 3,376 records and 8 attributes, namely date_publication, media_source, title_article, content_article, url_article, sentiment_label, polarity_score, and ai_confidence summarized in the Table 1.

Table 1. Dataset attributes

Attribute	Description
date_publication	Publication date of the financial news article
media_source	Source of the news article
title_article	News article title
content_article	Article content
url_article	Article URL
sentiment_label	Sentiment class generated by the sentiment model
polarity_score	Sentiment polarity score
ai_confidence	Confidence score of the machine-generated label

The sentiment labels used in this study were not manually assigned by human annotators. Instead, the labels were generated automatically using a machine learning-based sentiment classification process. Each article was assigned one of three sentiment labels: negative, neutral, or positive. In addition, the dataset included polarity scores and confidence scores that represented the model-generated sentiment output.

Because the labels were generated automatically, they were treated as pseudo-labels rather than human-verified ground truth. This distinction is important because machine-generated labels may contain noise or misclassification, especially in domain-specific financial texts. Therefore, the objective of this study was not to claim absolute sentiment correctness, but to benchmark how well IndoBERT and Multilingual BERT learned and generalized sentiment patterns from pseudo-labeled Indonesian financial news data.

This labeling condition also influenced the interpretation of the results. High model performance indicates consistency with the machine-generated labeling patterns, while lower performance in minority classes may reflect both model limitations and potential uncertainty in the pseudo-labeling process.

2.2 Data Inspection

The initial inspection showed that there were no missing values in the main dataset attributes summarized in the Table 2 and 3. The initial sentiment distribution consisted of 3,080 neutral articles, 184 negative articles, and 112 positive articles. The media source distribution consisted of 1,711 articles from Kontan, 883 articles from CNBC, and 782 articles from Bisnis.com.

Table 2. Initial sentiment distribution

Sentiment Class	Number of Articles
Neutral	3,080
Negative	184
Positive	112

Table 3. Media source distribution

Media Source	Number of Articles
Kontan	1,711
CNBC	883
Bisnis.com	782

The class distribution showed that the dataset was highly imbalanced, with the neutral class dominating the dataset. This condition required the use of class weights and macro-averaged metrics during model evaluation.

2.3 Text Preprocessing

Text preprocessing was performed to prepare the dataset for transformer-based classification. The preprocessing steps included:

1. **Text cleaning**
   The first step was text cleaning, which removed unnecessary elements such as HTML tags, URLs, excessive whitespace, and non-informative formatting. This step was important because raw news content may contain reporter information, editor information, extraction artifacts, or web formatting that does not directly contribute to sentiment classification.
2. **Text construction**
   The article title and article content were combined into one text input because the title often contains the main news signal, while the content provides additional contextual information. The combined text was used as the input for both IndoBERT and Multilingual BERT.
   $$text = title\_article + content\_article \quad (1)$$
3. **Duplicate removal**
   Duplicate articles were removed based on the article URL and the combination of title and content to avoid data redundancy and potential bias during training and testing.
4. **Label normalization**
   Sentiment labels were standardized into three classes: negative, neutral, and positive. These labels were then encoded into numerical form, where negative was encoded as 0, neutral as 1, and positive as 2.
5. **Invalid data removal**
   Since transformer models require tokenized input, each text was tokenized using the tokenizer associated with each model. IndoBERT used the IndoBERT tokenizer, while Multilingual BERT used the multilingual BERT tokenizer. The maximum sequence length was set to 256 tokens. Longer articles were truncated to fit this maximum length, while shorter articles were padded dynamically during training.

After preprocessing, the final dataset contained 3,366 records. The final sentiment distribution consisted of 3,070 neutral articles, 184 negative articles, and 112 positive articles summarized in the Table 4.

Table 4. Final sentiment distribution after preprocessing

Sentiment Class	Number of Articles
Neutral	3,070
Negative	184
Positive	112
Total	3,366

2.4 Label Encoding

The sentiment labels were converted into numerical values to be used in the classification models. This encoding was used consistently for both IndoBERT and Multilingual BERT summarized in the Table 5.

Table 5. Label encoding

Sentiment Label	Encoded Value
Negative	0
Neutral	1
Positive	2

2.5 Dataset Splitting

The dataset was divided into training, validation, and testing subsets using stratified splitting**.** Stratified splitting was used to preserve the proportion of each sentiment class in all subsets. The training set consisted of 2,149 neutral**,** 129 negative, and 78 positive articles. The validation set consisted of 461 neutral**,** 27 negative**,** and 17 positive articles. The test set consisted of 460 neutral**,** 28 negative, and 17 positive articles summarized in the Table 6.

Table 6. Class distribution in each data subset

Subset	Negative	Neutral	Positive	Total
Training	129	2,149	78	2,356
Validation	27	461	17	505
Testing	28	460	17	505

2.6 Handling Imbalanced Data

Because the dataset was dominated by the neutral class, class weighting was applied during model training. The computed class weights were summarized in the Table 7.

Table 7. Class distribution in each data subset

Sentiment Class	Class Weight
Negative	6.0879
Neutral	0.3654
Positive	10.0684

The positive class received the highest weight because it had the smallest number of samples. The weighted cross-entropy loss was used to reduce the tendency of the model to overfit the majority class [23].The weighted cross-entropy loss is defined as:

$$L = -\sum_{i=1}^{C} w_i y_i \log(\hat{y}_i) \quad (2)$$

where C is the number of classes, w_i is the class weight, y_i is the true label, and $\hat{y}_i$ is the predicted probability for class i.

2.7 Transformer-Based Models

IndoBERT was used to represent an Indonesian-specific transformer model. Multilingual BERT was used as a multilingual comparison model. Both models were fine-tuned for the same three-class sentiment classification task.This study compared two transformer-based models summarized in the Table 8.

Table 8. Models used in the experiment

Model	Checkpoint	Description
IndoBERT	indobenchmark/indobert-base-p1	Indonesian pre-trained BERT model
Multilingual BERT	bert-base-multilingual-cased	Multilingual BERT model trained on multiple languages

2.8 Tokenization

Each model used its own tokenizer. The text input was tokenized using a maximum sequence length of 256 tokens. Tokenization was performed with truncation to ensure that long articles could be processed within the input length limit [24]. The tokenization process produced the following inputs summarized in the Table 9.

Table 9. Training configuration

Input	Description
input_ids	Token identifiers generated by the tokenizer
attention_mask	Mask indicating valid tokens and padding tokens
labels	Encoded sentiment labels

2.9 Model Training Procedure

Both IndoBERT and Multilingual BERT were fine-tuned using the same training configuration to ensure a fair comparison summarized in the Table 10.

Table 10. Training configuration

Parameter	Value
Number of epochs	3
Learning rate	2e-5
Batch size	8
Maximum sequence length	256 tokens
Loss function	Weighted cross-entropy loss
Validation strategy	Evaluation every epoch
Model selection metric	Macro F1-score

Macro F1-score was used as the main model selection metric because the dataset was highly imbalanced. Unlike accuracy, macro F1-score gives equal importance to each class, including minority classes.

2.10 Evaluation Metrics

The models were evaluated using accuracy, precision, recall, macro F1-score, and weighted F1-score[25].

Accuracy

Accuracy measures the overall proportion of correctly classified articles[26].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

Precision

Precision measures how many predicted sentiment labels were correct.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

Recall

Recall measures how many actual sentiment labels were successfully identified.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F1-Score

F1-score balances precision and recall.

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

Macro F1-Score

Macro F1-score calculates the average F1-score across all classes with equal weight for each class. This metric was emphasized because the dataset contained a severe class imbalance.

$$Macro\ F1 = \frac{1}{C}\sum_{i=1}^{C} F1_i \tag{7}$$

Weighted F1-Score

Weighted F1-score calculates the average F1-score by considering the number of samples in each class[27].

$$Weighted\ F1 = \sum_{i=1}^{C} \frac{n_i}{N} F1_i \tag{8}$$

4. Result and Discussion

This section presents the experimental results and discussion of transformer-based sentiment classification models for Indonesian financial news. The models evaluated in this study were IndoBERT and Multilingual BERT. The evaluation was conducted using accuracy, precision, recall, macro F1-score, and weighted F1-score. The discussion is divided into dataset analysis, training results, test performance, and comparative interpretation.

4.1 Dataset Overview

The dataset used in this study consisted of Indonesian economic and financial news articles from the first quarter of 2026. The original dataset contained 3,376 records and 8 attributes, namely date_publication, media_source, title_article, content_article, url_article, sentiment_label, polarity_score, and ai_confidence. No missing values were found in the main attributes. After preprocessing and duplicate removal, the dataset contained 3,366 records summarized in the Table 11.

Table 11. Dataset summary

Component	Value
Initial data	3,376 records
Final data after preprocessing	3,366 records
Number of attributes	8 attributes
Sentiment classes	Negative, Neutral, Positive
Media sources	Kontan, CNBC, Bisnis.com

The media source distribution showed that most articles came from Kontan, followed by CNBC and Bisnis.com. Kontan contributed 1,711 articles, CNBC contributed 883 articles, and Bisnis.com contributed 782 articles. This distribution indicates that the dataset was dominated by Kontan articles summarized in the Table 12.

Table 12. Media source distribution

Media Source	Number of Articles
Kontan	1,711
CNBC	883
Bisnis.com	782

4.2 Sentiment Class Distribution

The sentiment class distribution showed a strong imbalance. After preprocessing, the dataset consisted of 3,070 neutral articles, 184 negative articles, and 112 positive articles. This means that the neutral class dominated the dataset, while positive and negative classes were minority classes summarized in the Table 13.

Table 13. Sentiment class distribution after preprocessing

Sentiment Class	Number of Articles
Neutral	3,070
Negative	184
Positive	112
Total	3,366

This imbalance is important because a model can achieve high accuracy by mainly predicting the majority class. Therefore, macro F1-score was used as an important evaluation metric because it gives equal importance to each sentiment class. In addition, class weights were applied during training to reduce the bias toward the neutral class. The computed class weights were 6.0879 for negative, 0.3654 for neutral, and 10.0684 for positive. These values indicate that the positive and negative classes received higher weights during training.

4.3 Data Splitting

The dataset was divided into training, validation, and testing sets using stratified splitting. The training set contained 2,356 articles, the validation set contained 505 articles, and the test set contained 505 articles. Stratified splitting was applied to preserve the proportion of sentiment classes in each subset summarized in the Table 14.

Table 14. Dataset split

Data Subset	Number of Articles	Negative	Neutral	Positive
Training	2,356	129	2,149	78
Validation	505	27	461	17

Testing	505	28	460	17

The class distribution in the test set shows that neutral articles remained dominant. Therefore, test results must be interpreted carefully. Accuracy and weighted F1-score may appear high because the majority class is large, while macro F1-score provides a more balanced view of model performance across all sentiment classes.

4.4 Validation Performance

The two transformer models were fine-tuned for three epochs. The best validation performance of IndoBERT occurred in the second epoch, where it achieved an accuracy of 0.9584 and a macro F1-score of 0.8050. In the third epoch, the macro F1-score slightly decreased to 0.7968, although the validation recall increased. This indicates that IndoBERT reached its strongest validation performance before the final epoch summarized in the Table 15.

Table 15. IndoBERT validation performance

Epoch	Training Loss	Validation Loss	Accuracy	Precision Macro	Recall Macro	F1 Macro	F1 Weighted
1	0.6589	1.0362	0.9287	0.7901	0.5498	0.6053	0.9209
2	0.7265	0.8745	0.9584	0.9036	0.7393	0.8050	0.9543
3	0.2875	0.7252	0.9485	0.8047	0.7894	0.7968	0.9483

Multilingual BERT showed gradual improvement across epochs. Its macro F1-score increased from 0.5044 in the first epoch to 0.7084 in the third epoch. However, its best macro F1-score was still lower than the best validation macro F1-score achieved by IndoBERT summarized in the Table 16.

Table 16. Multilingual BERT validation performance

Epoch	Training Loss	Validation Loss	Accuracy	Precision Macro	Recall Macro	F1 Macro	F1 Weighted
1	0.8475	1.2098	0.9248	0.7764	0.4684	0.5044	0.9097
2	0.9510	0.8869	0.9366	0.7347	0.5933	0.6448	0.9268
3	0.5238	0.7734	0.9347	0.7142	0.7074	0.7084	0.9329

Based on the validation results, IndoBERT showed stronger performance than Multilingual BERT, especially in macro F1-score. This suggests that an Indonesian-specific pre-trained transformer model was more suitable for capturing sentiment patterns in Indonesian financial news.

4.5 Test Performance Comparison

The final evaluation was conducted using the independent test set containing 505 articles. Table 17 presents the test performance of IndoBERT and Multilingual BERT.

Table 17. Test performance comparison

Model	Accuracy	Precision Macro	Recall Macro	F1 Macro	F1 Weighted
IndoBERT	0.94	0.75	0.69	0.71	0.94
Multilingual BERT	0.93	0.69	0.72	0.70	0.93

The results show that IndoBERT achieved the best overall performance based on accuracy, macro precision, macro F1-score, and weighted F1-score. IndoBERT obtained an accuracy of 0.94, macro precision of 0.75, macro recall of 0.69, macro F1-score of 0.71, and weighted F1-score of 0.94.

Multilingual BERT achieved slightly lower accuracy and macro F1-score, with an accuracy of 0.93 and macro F1-score of 0.70. However, Multilingual BERT achieved higher macro recall, namely 0.72, compared with IndoBERT's macro recall of 0.69. This means that Multilingual BERT was slightly better in identifying sentiment classes from a recall perspective, although its overall balance of precision and recall was slightly lower.

4.6 Class-Level Performance Analysis

A class-level analysis was conducted to evaluate how each model performed on negative, neutral, and positive sentiment classes summarized in the Table 18 and 19.

Table 18. IndoBERT class-level performance

Class	Precision	Recall	F1-Score	Support
Negative	0.78	0.50	0.61	28
Neutral	0.97	0.98	0.97	460
Positive	0.50	0.59	0.54	17
Accuracy	-	-	0.94	505
Macro Average	0.75	0.69	0.71	505
Weighted Average	0.94	0.94	0.94	505

Table 19. Multilingual BERT class-level performance

Class	Precision	Recall	F1-Score	Support
Negative	0.50	0.43	0.46	28
Neutral	0.97	0.97	0.97	460
Positive	0.59	0.76	0.67	17
Accuracy	-	-	0.93	505
Macro Average	0.69	0.72	0.70	505
Weighted Average	0.93	0.93	0.93	505

The class-level results show that both models performed very well on the neutral class. This was expected because the neutral class dominated the dataset. IndoBERT achieved a neutral F1-score of 0.97, while Multilingual BERT also achieved a neutral F1-score of 0.97. For the negative class, IndoBERT performed better than Multilingual BERT. IndoBERT achieved a negative F1-score of 0.61, while Multilingual BERT achieved 0.46. This indicates that IndoBERT was more effective in recognizing negative sentiment patterns in Indonesian financial news. For the positive class, Multilingual BERT performed better than IndoBERT. Multilingual BERT achieved a positive F1-score of 0.67, while IndoBERT achieved 0.54. This suggests that Multilingual BERT was more sensitive in detecting positive sentiment, as shown by its higher positive recall of 0.76.

The class-level evaluation showed that both IndoBERT and Multilingual BERT performed very well in classifying the neutral class. IndoBERT achieved a neutral F1-score of 0.97, while Multilingual BERT also achieved a neutral F1-score of 0.97. This strong performance was expected because neutral sentiment dominated the dataset. However, both models showed lower performance on minority classes. For the negative class, IndoBERT achieved better performance than Multilingual BERT, with an F1-score of 0.61 compared with 0.46. This indicates that IndoBERT was more effective in identifying negative sentiment patterns in Indonesian financial news. The better performance of IndoBERT may be related to its Indonesian-specific pretraining, which allows it to better capture local linguistic structures and contextual expressions.

For the positive class, Multilingual BERT performed better than IndoBERT. Multilingual BERT achieved a positive F1-score of 0.67, while IndoBERT achieved 0.54. This result indicates that Multilingual BERT was more sensitive in identifying positive sentiment in the test set. However, the number of positive samples was very limited, with only 17 positive articles in the test set. Therefore, the performance on this class should be interpreted carefully. The confusion matrix analysis confirmed that most classification errors occurred in the minority classes. Negative and positive articles were more likely to be misclassified as neutral. This occurred because the neutral class dominated the training data, making the models more exposed to neutral sentiment patterns. Although class weighting was applied, the limited number of positive and negative examples still affected the models' ability to learn minority-class patterns.

4.7 Discussion

The experimental results showed that IndoBERT achieved slightly better overall performance than Multilingual BERT. IndoBERT obtained an accuracy of 0.94 and a macro F1-score of 0.71, while Multilingual BERT obtained an accuracy of 0.93 and a macro F1-score of 0.70. Although the difference was small, this result suggests that an Indonesian-specific pre-trained model was more suitable for Indonesian financial news sentiment classification. One possible reason for IndoBERT's better performance is its language-specific pretraining. Since IndoBERT was trained on Indonesian text, it may better capture Indonesian grammar, vocabulary, word formation, and contextual usage. This advantage is important for financial news, where sentiment is often expressed through formal journalistic language and domain-specific terms. In contrast, Multilingual BERT was trained on many languages. Although it supports Indonesian, its representational capacity is distributed across multiple languages. As a result, it may be less specialized in capturing Indonesian-specific expressions compared with IndoBERT. Nevertheless, Multilingual BERT remained competitive and even showed better performance in detecting positive sentiment. This indicates that multilingual representation can still provide useful contextual understanding for Indonesian financial text. The results also show that model performance was strongly affected by class imbalance. The high accuracy and weighted F1-score were influenced by the dominant neutral class. Therefore, macro F1-score provides a more reliable comparison because it evaluates the average performance across all classes equally. Based on macro F1-score, IndoBERT was the best model in this study.[28], [29].

This study has several limitations. First, the sentiment labels were machine-generated and were not manually verified by human annotators. Therefore, the labels should be interpreted as pseudo-labels rather than definitive ground truth. This may affect the reliability of the evaluation, especially for sentiment expressions that require deeper financial domain interpretation.

Second, the dataset was highly imbalanced, with the neutral class dominating the distribution. Although class weighting was applied, the small number of positive and negative samples limited the models' ability to learn minority-class patterns effectively.

Third, this study only compared two transformer-based models, namely IndoBERT and Multilingual BERT. Although this comparison was useful for evaluating Indonesian-specific and multilingual transformer models, future research should include additional baselines such as Logistic Regression, Support Vector Machine, Long Short-Term Memory, Convolutional Neural Network, IndoBERTweet, Indonesian RoBERTa, or financial-domain transformer models.

Fourth, the experiment used a maximum sequence length of 256 tokens. Since financial news articles may contain important information beyond this limit, truncation may have removed useful context from longer articles.

5. Conclusion

This study successfully addressed the objective stated in the Introduction, namely to benchmark two transformer-based models, IndoBERT and Multilingual BERT, for Indonesian financial news sentiment classification. The experiment used Indonesian financial news data from Kontan, CNBC, and Bisnis.com. After preprocessing, the dataset contained 3,366 articles consisting of 3,070 neutral, 184 negative, and 112 positive articles. The results showed that IndoBERT achieved the best overall performance, with an accuracy of 0.94, macro precision of 0.75, macro recall of 0.69, macro F1-score of 0.71, and weighted F1-score of 0.94. Multilingual BERT achieved competitive performance, with an accuracy of 0.93, macro precision of 0.69, macro recall of 0.72, macro F1-score of 0.70, and weighted F1-score of 0.93. These results indicate that IndoBERT was slightly more suitable for Indonesian financial news sentiment classification, particularly based on macro F1-score.

The class-level analysis showed that both models performed very well on the neutral class but still faced challenges in classifying minority classes. IndoBERT performed better on negative sentiment, while Multilingual BERT performed better on positive sentiment. This finding confirms that class imbalance affected model performance and that minority-class classification remains an important challenge.

The findings of this study suggest that Indonesian-specific transformer models can provide advantages in Indonesian financial text classification. However, the use of machine-generated pseudo-labels and the imbalance of sentiment classes should be considered when interpreting the results. Future research should include manually annotated data, more balanced sentiment distributions, additional baseline models, and more advanced transformer architectures. In addition, future studies may integrate financial sentiment classification results with stock market indicators to analyze the relationship between media sentiment and market behavior..

References

[1] G. Anese, M. Corazza, M. Costola, and L. Pelizzon, "Impact of public news sentiment on stock market index return and volatility," *Computational Management Science*, vol. 20, no. 1, p. 20, Dec. 2023, doi: 10.1007/s10287-023-00454-2.

[2] M. Bask, L. Forsberg, and A. Östling, "Media sentiment and stock returns," *The Quarterly Review of Economics and Finance*, vol. 94, pp. 303–311, Apr. 2024, doi: 10.1016/j.qref.2024.02.008.

[3] L. Yacoubian, "The Predictive Power of Social Media Sentiment on Stock Market Returns," *International Journal For Multidisciplinary Research*, vol. 7, no. 3, May 2025, doi: 10.36948/ijfmr.2025.v07i03.46689.

[4] S. D. Sriasih, F. A. Razak, and H. al I. Ikhsan, "AI-Driven Sentiment Analysis of Retail Investor Behavior during Market Volatility: A Study of Twitter Data in Southeast Asia," *Journal of Management and Informatics*, vol. 4, no. 1, pp. 741–756, Apr. 2025, doi: 10.51903/jmi.v4i1.179.

[5] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial Sentiment Analysis: Techniques and Applications," *ACM Comput. Surv.*, vol. 56, no. 9, pp. 1–42, Oct. 2024, doi: 10.1145/3649451.

[6] M. R. Agam, N. Setyawan, C.-C. Sun, H.-K. Su, and J.-W. Hsieh, "Classification of Indonesian Language News Documents Using RNN and Transformers," in *2025 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, IEEE, Jul. 2025, pp. 407–408. doi: 10.1109/ICCE-Taiwan66881.2025.11207956.

[7] A. Karentia, F. Winaya, and D. Suhartono, "Hybrid approach sentiment analysis using Transformer-LSTM in the Indonesian language," in *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, IEEE, Dec. 2024, pp. 754–758. doi: 10.1109/ISRITI64779.2024.10963365.

[8] Z. Zhu, "BERT and Its Applications in Natural Language Understanding," *Applied and Computational Engineering*, vol. 175, no. 1, pp. 99–105, Aug. 2025, doi: 10.54254/2755-2721/2025.AST26090.

[9] J. Wang *et al.*, "Utilizing BERT for Information Retrieval: Survey, Applications, Resources, and Challenges," *ACM Comput. Surv.*, vol. 56, no. 7, pp. 1–33, Jul. 2024, doi: 10.1145/3648471.

[10] N. E. Aliyah, R. W. Sholikah, H. Firdausi, H. T. Ciptaningtyas, and I. A. Sabilla, "Enhancing Automated Essay Scoring in Bahasa Indonesia with IndoBERT and IndoSBERT," in *2025 International Conference on Smart Computing, IoT and Machine Learning (SIML)*, IEEE, Jun. 2025, pp. 1–7. doi: 10.1109/SIML65326.2025.11080721.

[11] L. Afuan, N. Hidayat, H. Hamdani, H. Ismanto, B. C. Purnama, and D. I. Ramdhani, "Optimizing BERT Models with Fine-Tuning for Indonesian Twitter Sentiment Analysis," *J. Wirel. Mob. Netw. Ubiquitous Comput. Dependable Appl.*, vol. 16, no. 2, pp. 248–267, Jun. 2025, doi: 10.58346/JOWUA.2025.I2.016.

[12] Taufiq Dwi Purnomo and Joko Sutopo, "COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA," *International Journal Science and Technology*, vol. 3, no. 3, pp. 11–21, Nov. 2024, doi: 10.56127/ijst.v3i3.1739.

[13] E. Z. Rahardjo and T. Mauritsius, "Predicting Bank Share Prices in Indonesia using News Sentiment Analysis," in *2025 International Conference on Information Management and Technology (ICIMTech)*, IEEE, Aug. 2025, pp. 746–751. doi: 10.1109/ICIMTech67074.2025.11265540.

[14] J. Delgadillo, J. Kinyua, and C. Mutigwe, "FinSoSent: Advancing Financial Market Sentiment Analysis through Pretrained Large Language Models," *Big Data and Cognitive Computing*, vol. 8, no. 8, p. 87, Aug. 2024, doi: 10.3390/bdcc8080087.

[15] S. Adhikari, S. Thapa, U. Naseem, H. Y. Lu, G. Bharathy, and M. Prasad, "Explainable hybrid word representations for sentiment analysis of financial news," *Neural Networks*, vol. 164, pp. 115–123, Jul. 2023, doi: 10.1016/j.neunet.2023.04.011.

[16] A. A. Wijaya, B. A. Jabar, G. M. Sutarman, and A. Wijaya, "Analyzing the Short Term Impact of News Sentiment on Indonesian Stock Prices Using Deep Learning Models," in *2025 4th International Conference on Digital Transformation and Applications (ICDXA)*, IEEE, Oct. 2025, pp. 268–272. doi: 10.1109/ICDXA69105.2025.11329903.

[17] R. Gupta, "Bidirectional encoders to state-of-the-art: a review of BERT and its transformative impact on natural language processing," *Информатика. Экономика. Управление - Informatics. Economics. Management*, vol. 3, no. 1, pp. 0311–0320, Mar. 2024, doi: 10.47813/2782-5280-2024-3-1-0311-0320.

[18] 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.

[19] L. S. Yuan and L. T. Ming, "Sentiment Prediction Using Multilingual Bidirectional Encoder Representations And Cross-Lingual Language Model Robustly Optimized Bert Approach From Transformers on Code-mixed Text," in *2025 IEEE International Conference on Computation, Big-Data and Engineering (ICCBE)*, IEEE, Jun. 2025, pp. 901–905. doi: 10.1109/ICCBE65177.2025.11256168.

[20] M. I. Salih, S. M. Mohammed, A. Kh. Ibrahim, O. M. Ahmed, and L. M. Haji, "Fine-Tuning BERT for Automated News Classification," *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 22953–22959, Jun. 2025, doi: 10.48084/etasr.10625.

[21] 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)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICEEI59426.2023.10346625.

[22] J. Acs, E. Hamerlik, R. Schwartz, N. A. Smith, and A. Kornai, "Morphosyntactic probing of multilingual BERT models," *Nat. Lang. Eng.*, vol. 30, no. 4, pp. 753–792, Jul. 2024, doi: 10.1017/S1351324923000190.

[23] V. Ganganwar and R. Rajalakshmi, "Employing synthetic data for addressing the class imbalance in aspect-based sentiment classification," *Journal of Information and Telecommunication*, vol. 8, no. 2, pp. 167–188, Apr. 2024, doi: 10.1080/24751839.2023.2270824.

[24] J. Yang, F. Wei, N. Huber-Fliflet, A. Dabrowski, Q. Mao, and H. Qin, "An Empirical Analysis of Text Segmentation for BERT Classification in Extended Documents," in *2023 IEEE International Conference on Big Data (BigData)*, IEEE, Dec. 2023, pp. 2793–2797. doi: 10.1109/BigData59044.2023.10386783.

[25] I. Markoulidakis and G. Markoulidakis, "Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis," *Technologies (Basel).*, vol. 12, no. 7, p. 113, Jul. 2024, doi: 10.3390/technologies12070113.

[26] M. Heydarian, T. E. Doyle, and R. Samavi, "MLCM: Multi-Label Confusion Matrix," *IEEE Access*, vol. 10, pp. 19083–19095, 2022, doi: 10.1109/ACCESS.2022.3151048.

[27] M. C. Hinojosa Lee, J. Braet, and J. Springael, "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores," *Applied Sciences*, vol. 14, no. 21, p. 9863, Oct. 2024, doi: 10.3390/app14219863.

[28] 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)*, IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICEEI59426.2023.10346625.

[29] Taufiq Dwi Purnomo and Joko Sutopo, "COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA," *International Journal Science and Technology*, vol. 3, no. 3, pp. 11–21, Nov. 2024, doi: 10.56127/ijst.v3i3.1739.