Jurnal Informatika dan Tekonologi Komputer

E-ISSN: 2809-9230
P-ISSN: 2809-9249

Comparison of Deep Learning Models for Sentiment Analysis of IPOT Financial App Reviews Using Convolutional Neural Network (CNN) and IndoBERT

I Gusti Ngurah Agung Pawana [1]*, Made Widya Jayantari [2], and I Gusti Ngurah Agung Bagus Aryawana [3]

[1] Computer Engineering Department, Warmadewa University; Terompong Street, Bali, Indonesia; e-mail: agungpawana@warmadewa.ac.id
[2] Civil Engineering Department, Udayana University; Bukit Jimbaran Street; e-mail : widyajayantari13@gmail.com
3 Information System Department, Institute of Technology and Business STIKOM; Renon Street, Bali; e-mail: agoesgung@gmail.com
* Corresponding Author: I Gusti Ngurah Agung Pawana

Abstract: *The rapid expansion of mobile-based financial applications has generated a large volume of user reviews that contain valuable insights into user satisfaction and system performance. IPOT (Indo Premier Online Trading) is a widely used financial application in Indonesia, making sentiment analysis of its user reviews essential for evaluating its service and improving it. This study applies an experimental methodology to compare the performance of two deep learning architectures, Convolutional Neural Network (CNN) and IndoBERT, for sentiment analysis of financial application reviews. User review data were collected from the Google Play Store. Sentiment labels were automatically assigned based on user ratings, and the dataset was balanced using stratified sampling to obtain 15,000 reviews. Text preprocessing included case folding, removal of punctuation and special characters, tokenization, stopword removal, and stemming. The dataset was then split into training, validation, and testing sets, with oversampling applied only to the training data to prevent data leakage. The comparison between Convolutional Neural Networks (CNNs) and IndoBERT for sentiment analysis of IPOT financial application reviews shows that both models perform sentiment classification effectively, with different strengths across sentiment categories. The CNN model achieved higher overall accuracy (0.8113) compared to IndoBERT (0.7880), indicating strong performance in detecting dominant sentiment patterns, particularly positive sentiment. Meanwhile, IndoBERT achieved superior performance in negative and neutral sentiment classification, as evidenced by higher recall and F1 scores. The confusion matrix and error analysis results further indicate that IndoBERT is more effective at understanding contextual and nuanced language, whereas CNN is more sensitive to explicit lexical sentiment indicators.*

Keywords: *Sentiment Analysis; Deep Learning; CNN; IndoBERT; Financial App Reviews*

Received: February 13, 2026
Revised: February 20, 2026
Accepted: March 7, 2026
Published: March 14, 2026
Curr. Ver.: March 14, 2026

1. Introduction

The rapid advancement of digital financial services has significantly transformed investment activities, particularly through mobile-based trading applications [1], [2]. IPOT (Indo Premier Online Trading) is one of the most widely used financial applications in Indonesia, providing retail investors with access to stock trading, mutual funds, and other investment instruments [3], [4]. As user adoption increases, a large volume of user-generated reviews emerges on application marketplaces such as Google Play Store, reflecting user satisfaction, complaints, expectations, and overall experiences [5], [6]. Despite the availability of abundant user feedback, manually analyzing thousands of reviews is inefficient and prone to subjectivity. Sentiment analysis has become an essential approach for automatically identifying users' opinions and emotional tendencies in text. This technique, also known as opinion mining, leverages natural language processing (NLP) and machine learning to extract and analyze sentiments expressed in text, providing valuable insights into public opinion, customer satisfaction, and brand perception. The exponential growth of user-generated content on social

media platforms, product reviews, and other online sources has significantly increased the importance of sentiment analysis, enabling businesses and researchers to make data-driven decisions and improve customer experiences [7], [8], [9], [10], [11]. In the financial application domain, accurate sentiment classification is particularly important because user reviews often contain critical information about system reliability, transaction security, and service quality.

Recent studies have shown that deep learning approaches outperform traditional machine learning methods in sentiment analysis, particularly for unstructured text data. Convolutional Neural Networks (CNNs) are widely used for their ability to capture local semantic patterns via convolutional filters. In contrast, transformer-based models such as BERT have shown superior performance in modeling contextual relationships. IndoBERT, a pre-trained language model designed for Indonesian text, is effective across various natural language processing tasks [12].

Wang [13] evaluated various sentiment analysis methods, comparing traditional machine learning techniques with advanced deep learning models, with a focus on BERT-derived architectures, CNNs, Long Short-Term Memory (LSTM) models, Support Vector Machines (SVMs), and Naive Bayes classifiers. This study shows that BERT achieves the best results, with 86.51% accuracy and an F1 score of 0.8653, but requires the longest training time (≈10,980 seconds / ~3 hours). LSTM achieves strong performance with 83.18% accuracy and an F1-score of 0.8326, while significantly reducing training time to 1,080 seconds (~18 minutes), making it a good balance between accuracy and efficiency. SVM achieves moderate performance with 79.12% accuracy and an F1-score of 0.7930, but it trains much faster (60 seconds) and is suitable for limited-resource environments. The lexicon-based method has the lowest performance (78.45% accuracy, 0.7868 F1-score) but is the fastest (30 seconds) and easily interpretable.

CNNs are effective at extracting localized semantic patterns because their convolutional layers are designed to capture local dependencies, such as specific word patterns in text or fine visual details in images. This makes CNNs particularly useful for tasks involving smaller datasets, where capturing detailed local features is important, such as in medical image classification and segmentation. Additionally, when combined with pretraining strategies, CNNs can achieve competitive performance in certain tasks, indicating their continued relevance even in areas increasingly dominated by Transformer-based models [14].

In contrast, Transformer models, especially BERT, are highly effective at capturing global semantic relationships through self-attention mechanisms that model long-range dependencies in the data. BERT benefits from pretraining on large-scale text corpora, allowing strong generalization across various natural language processing tasks, including semantic matching and question answering. Its bidirectional architecture further enhances its ability to understand complex linguistic structures and contextual meaning, making it particularly powerful for tasks requiring deep language comprehension [15].

However, comparative studies focusing on sentiment analysis of Indonesian financial application reviews remain limited. In particular, the performance differences between CNN and IndoBERT in handling positive, negative, and neutral sentiments within financial contexts have not been thoroughly investigated. This study addresses the research gap by comparing CNN and IndoBERT models on user reviews of the IPOT financial application collected from the Google Play Store. The objective of this research is to analyze and compare the effectiveness of CNN and IndoBERT models in sentiment classification of IPOT application reviews. The results of this study are expected to provide insights into the strengths and limitations of each model and to support the development of more effective sentiment analysis systems for financial applications.

2. Literature Review

Sentiment analysis has become an important task in natural language processing (NLP), particularly for analyzing user-generated content such as application reviews. With the rapid growth of financial technology applications, understanding user sentiment is essential to

improve service quality and user experience. Deep learning models, especially Convolutional Neural Networks (CNN) and transformer-based models such as BERT and IndoBERT, have demonstrated strong performance in sentiment classification tasks.

2.1. Deep Learning Approaches in Sentiment Analysis

Deep learning techniques have significantly improved sentiment classification performance compared to traditional machine learning approaches. Balakrishnan et al. [16] showed that neural network-based models achieved up to 20% higher accuracy than traditional methods, with hybrid CNN–RNN–BiLSTM architectures achieving up to 96%. This finding confirms that deep learning architectures are well-suited to complex text classification tasks, such as sentiment analysis.

Similarly, Lal and Nasir [17] reported that BERT-based models generally outperform traditional RNN architectures because they capture contextual embeddings. However, transformer models require more computational resources and longer training times, which are important considerations when selecting models for real-world implementation.

2.2 CNN in Sentiment Analysis of Application and Financial Reviews

CNNs are widely used for text classification because they efficiently capture local semantic patterns. Subowo [18] evaluated CNN, LSTM, and BERT models for application review sentiment classification and found that CNN achieved 75% accuracy, lower than BERT but still demonstrating reliable performance. The study highlighted CNN's efficiency and simpler architecture as advantages.

In financial and fintech contexts, Achmad et al. [19] applied CNN to analyze sentiment on digital wallet services and achieved 81% accuracy. Their findings showed that CNNs can effectively classify sentiment polarity but may be limited in their ability to understand contextual and complex linguistic structures.

Mahendra and Kusrini [20] further explored hybrid CNN architectures for investment application reviews and found that CNN-Gated Recurrent Unit (GRU) achieved 87.60% accuracy and a strong F1 score performance.

2.3 Transformer-Based Models and IndoBERT for Indonesian Sentiment Analysis

Transformer-based models, particularly BERT and its Indonesian adaptation, IndoBERT, have shown superior performance in sentiment analysis due to their contextual word-representation capabilities. Ramadhan et al. [21] compared CNN and IndoBERT for sentiment analysis of Indonesian political news and found that IndoBERT achieved 92.93% accuracy, outperforming CNN (89.13%). IndoBERT was particularly strong at classifying neutral sentiment, which is typically more difficult than classifying positive or negative sentiment.

IndoBERT has also demonstrated strong performance in aspect-based sentiment analysis tasks. Yulianti and Nissa [22] showed that fine-tuned IndoBERT models significantly outperformed CNN-based baselines, improving F1 scores by up to 23.6%. This indicates IndoBERT's strong ability to capture contextual meaning in Indonesian language datasets.

2.4 State of The Art and Gap Analysis

Although many prior studies have compared deep learning models for sentiment analysis, several research gaps remain. First, there is a domain gap, with most existing studies focusing on political news, e-commerce platforms, restaurant reviews, or digital wallet services. Research specifically analyzing sentiment in financial investment application reviews, such as the IPOT application, remains limited. Second, there is a language-specific gap. Although IndoBERT has demonstrated strong performance in Indonesian natural language processing tasks, comparative studies between IndoBERT and CNN, specifically in the fintech sentiment analysis domain, remain scarce. Third, there is a gap in model comparisons. Some studies compare CNNs with BERT or hybrid deep learning models, but direct comparisons between CNNs and IndoBERT on financial application review datasets are still rare. Additionally,

there is a gap in dataset quality. Several prior studies highlight the lack of spam or fake review filtering, which can affect model performance and reliability. This indicates the importance of robust data preprocessing techniques in future research.

Furthermore, several common limitations have been identified across previous studies. Many studies provide limited discussion regarding dataset bias and model generalizability. There is also a lack of focus on multilingual datasets or Indonesian-specific datasets in the financial domain. In addition, only a few studies evaluate the trade-off between computational cost and model performance, which is an important factor in real-world implementation. Moreover, only a limited number of studies analyze sentiment using real-world fintech application review datasets.

3. Proposed Method

This study employs an experimental research methodology to evaluate and compare the performance of two deep learning architectures, a Convolutional Neural Network (CNN) and IndoBERT, for sentiment analysis of financial application reviews. The methodological framework consists of sequential stages: data acquisition, sentiment labeling, text preprocessing, dataset partitioning and balancing, model construction, training, and evaluation. The overall research workflow is illustrated in Figure 1.

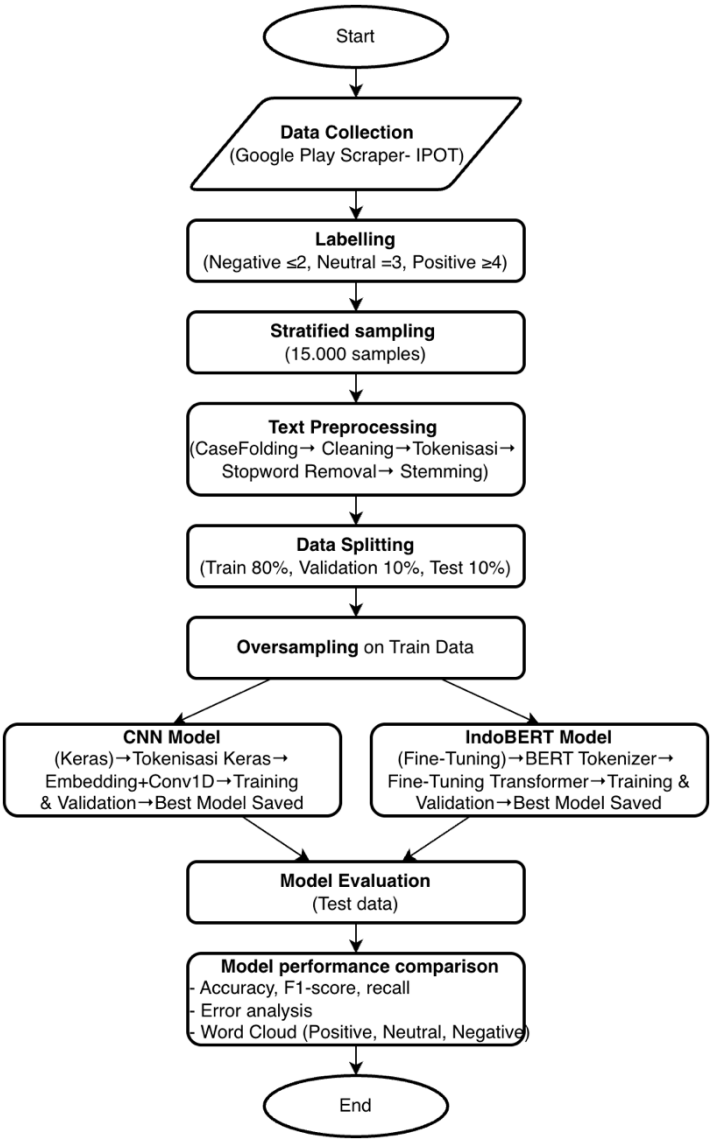

Figure 1. Research Framework

3.1. Data Collection

User reviews of the IPOT (Indo Premier Online Trading) financial application were collected from the Google Play Store using the Google Play Scraper library. The collected data include review text, numerical ratings, timestamps, and application version information. Reviews were filtered to retain only Indonesian-language content to ensure linguistic consistency and relevance. This process produced a large corpus of authentic user-generated text representing diverse user experiences with the IPOT application.

3.2. Sentiment Labeling and Sampling

Sentiment labels were automatically assigned based on user ratings to ensure objectivity and scalability. Reviews with ratings of 4–5 were labeled as positive, 3 as neutral, and 1–2 as negative. To mitigate the impact of class imbalance commonly found in user review datasets, stratified sampling was applied to construct a balanced dataset consisting of 15,000 reviews across the three sentiment categories.

3.3. Text Preprocessing

Text preprocessing was conducted to improve data quality and enhance model learning capability. The preprocessing pipeline included case folding, removal of punctuation, numerical values, and special characters, tokenization, stopword removal, and stemming using the Sastrawi library for Indonesian language processing. These steps aimed to reduce noise while preserving semantic meaning, resulting in a normalized textual representation suitable for deep learning models.

Pseudo-code for the preprocessing procedure is presented in Table 1.

Table 1. Pseudocode of text Preprocessing Pipeline

Pseudocode

for each review in the dataset:
  text = lowercase(review)
  text = remove_punctuation(text)
  tokens = tokenize(text)
  tokens = remove_stopwords(tokens)
  stemmed_tokens = stemming(tokens)
  final_text = join(stemmed_tokens)

3.4. Data Splitting and Balancing

The preprocessed dataset was divided into training, validation, and testing subsets to ensure unbiased performance evaluation. Oversampling techniques were applied exclusively to the training data to address class imbalance while preventing data leakage. This separation enables reliable evaluation of model generalization on unseen data.

3.5. Model Architecture

3.5.1 IndoBERT Model

IndoBERT is a transformer-based pre-trained language model optimized for Indonesian text. In this study, IndoBERT was fine-tuned for sentiment classification by adding a dense classification layer on top of the contextualized [CLS] token representation. This approach allows the model to capture long-range dependencies and contextual nuances within user reviews, which are particularly important for identifying neutral and ambiguous sentiments.

3.5.2 Convolutional Neural Network (CNN) Model

Both CNN and IndoBERT models were trained using the training dataset and validated on the validation dataset to optimize hyperparameters and prevent overfitting. The categorical cross-entropy loss function was employed as the optimization objective. Final model performance was evaluated on the test dataset using standard classification metrics, including accuracy, precision, recall, and F1-score. Comparative analysis was conducted to assess the strengths and limitations of each model across sentiment classes.

3.6. Model Training and Evaluation

Both CNN and IndoBERT models were trained using the training dataset and validated on the validation dataset to optimize hyperparameters and prevent overfitting. The categorical cross-entropy loss function was employed as the optimization objective. Final model performance was evaluated on the test dataset using standard classification metrics, including accuracy, precision, recall, and F1-score. Comparative analysis was conducted to assess the strengths and limitations of each model across sentiment classes.

3.7. Error Analysis and Visualization

Error analysis was performed by examining misclassified reviews to identify recurring linguistic patterns and sources of ambiguity. Additionally, word cloud visualizations were generated for positive, negative, and neutral sentiment classes to provide qualitative insights into dominant lexical features and to support the quantitative evaluation results.

4. Results and Discussion

4.1. Data Preprocessing

The dataset used in this study comprises 15,000 user reviews collected from the Google Play Store via stratified sampling. Text preprocessing was conducted to reduce noise and improve the quality of input data for deep learning models. The preprocessing pipeline included case folding, removal of punctuation, numbers, and special characters, tokenization, stopword removal, and stemming using the Sastrawi library for Indonesian language processing.

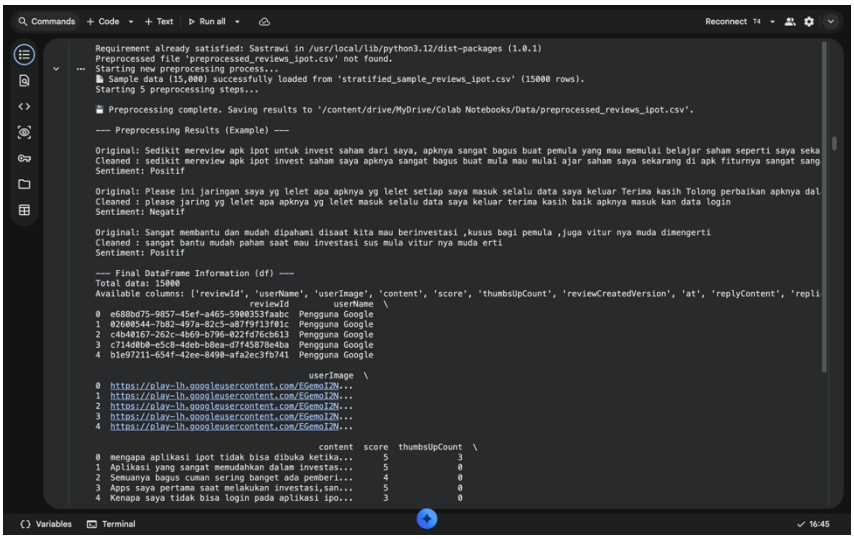

Figure 2. Preprocessing Flow

Figure 2 shows that informal, lengthy, and noisy user reviews were successfully transformed into cleaner, more structured textual representations without altering core sentiment information. Additional preprocessing outputs, such as tokens, stemmed tokens, and final cleaned text, were stored in the dataset to support model training and analysis. This preprocessing stage plays a critical role in improving the stability and convergence of both CNN and IndoBERT models.

4.2. Training and Validation Performance

Both the CNN and IndoBERT models were trained on the oversampled training dataset to address class imbalance. The CNN model was trained for up to four epochs with early stopping, achieving its best validation accuracy of 0.8153 at the first epoch and a validation loss of 0.7136. Subsequent epochs did not improve validation performance, indicating the onset of overfitting, which was effectively mitigated by early stopping. In contrast, the IndoBERT model was trained for three epochs and exhibited a consistent improvement in validation accuracy, reaching a maximum of 0.8300 in the third epoch. Although the validation loss increased slightly, the improvement in accuracy suggests that IndoBERT benefits from

its contextual representation capabilities, enabling it to capture complex semantic patterns in user reviews better.

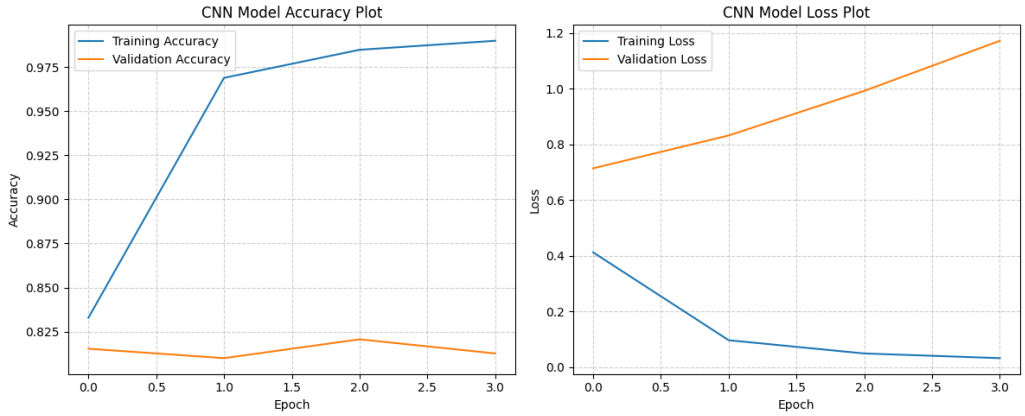

Figure 3. CNN Training and Validation Accuracy and Loss Curves

In the CNN Plot stage shown in Figure 3, the accuracy and loss curves reveal training dynamics that indicate overfitting. The training accuracy increases rapidly, reaching nearly 100%, while the validation accuracy improves only during the early epochs and then tends to stagnate or slightly decline. This pattern suggests that the CNN model effectively learns patterns from the training data but generalizes poorly to unseen validation data. This observation is further supported by the loss curves, where the training loss decreases sharply, whereas the validation loss begins to increase after a certain number of epochs. The divergence between decreasing training loss and increasing validation loss is a strong indicator of overfitting in the CNN model.

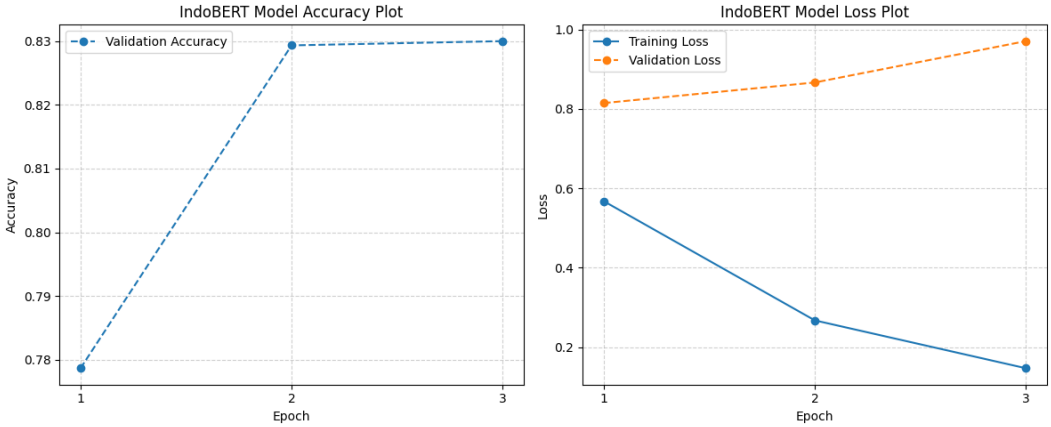

Figure 4. IndoBERT Training and Validation Accuracy and Loss Curves

In contrast, during the IndoBERT Plot stage, the model demonstrates more stable training behavior and superior generalization performance (Figure 4). The validation accuracy increases steadily across epochs and peaks at the third epoch, indicating that IndoBERT requires several epochs to adapt its contextual representations to the dataset effectively. Both training and validation loss curves generally decrease, with only minor fluctuations in the final epoch. Overall, these patterns reflect a more controlled learning process and effective error minimization on validation data.

Based on a comparison of the training curves of the two models, IndoBERT exhibits greater stability and better generalization performance than CNN when trained for the same number of epochs. This finding highlights the advantage of transformer-based models in capturing semantic context in Indonesian-language text reviews, particularly when early stopping halts CNN training prematurely due to overfitting.

4.3. Evaluation on Test Data

The final evaluation was conducted on a separate test dataset to assess the generalization capability of both models. The CNN model achieved an overall accuracy of 0.8113, slightly outperforming IndoBERT, which achieved 0.7880. However, a class-wise analysis reveals more nuanced performance differences. As shown in Table 2.

Table 2. Classification Report

Class	Model	Precision	Recall	F1-Score
Negative	CNN	0.6582	0.7143	0.6851
	IndoBERT	0.7180	0.6786	0.6977
Neutral	CNN	0.3125	0.0935	0.1439
	IndoBERT	0.2222	0.5047	0.3086
Positive	CNN	0.8826	0.9203	0.9010
	IndoBERT	0.9650	0.8562	0.9073
Accuracy	CNN			0.8113
	IndoBERT			0.7880
Macro Avg	CNN	0.6178	0.5760	0.5767
	IndoBERT	0.6351	0.6798	0.6379
Weighted Avg	CNN	0.7875	0.8113	0.7946
	IndoBERT	0.8520	0.7880	0.8137

IndoBERT demonstrated superior performance in handling negative and neutral sentiment classes, achieving higher recall and F1-score for the negative class and substantially higher recall for the neutral class than the CNN. Meanwhile, CNN performed better in the positive sentiment class, particularly in recall and F1-score, indicating its effectiveness in capturing dominant sentiment patterns through local feature extraction.

4.4. Confusion Matrix Analysis

The confusion matrix analysis provides deeper insight into misclassification patterns. The CNN model correctly classified the majority of positive reviews but struggled significantly with neutral reviews, frequently misclassifying them as positive or negative. This limitation suggests that CNNs rely heavily on local lexical cues, which may not adequately capture subtle or context-dependent sentiment expressions. Figure 3 shows the confusion matrices for the CNN and IndoBERT models on the test dataset.

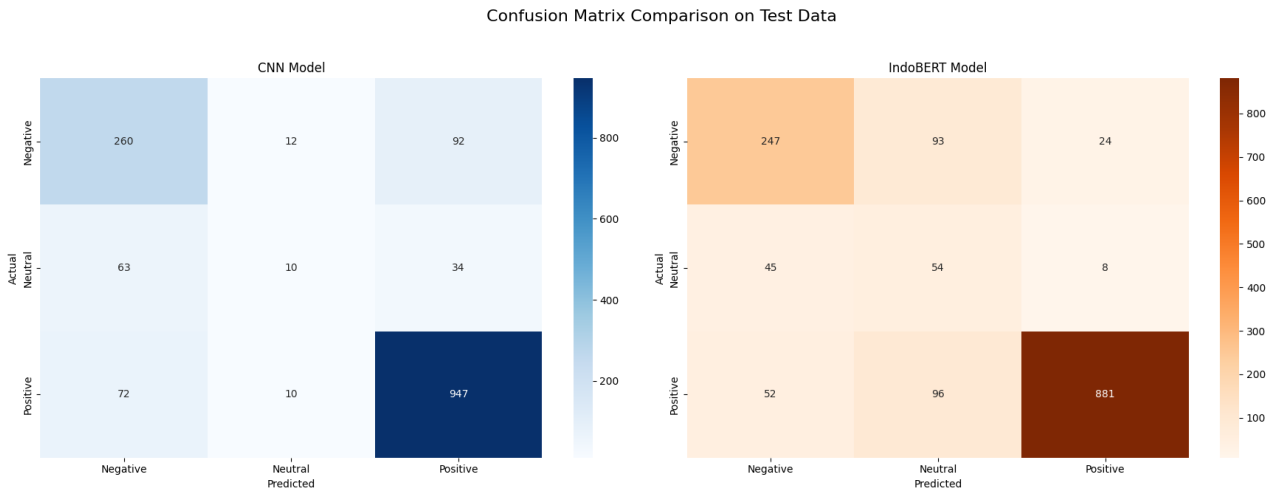

Figure 5. Confusion Matrices of CNN and IndoBERT on the Test Dataset

The confusion matrices of CNN and IndoBERT on the test dataset provide a comprehensive visualization of classification performance for each sentiment category. Each cell in the matrix represents the number of samples predicted for a specific class relative to their true class labels. The diagonal elements indicate correctly classified instances, while the off-diagonal elements reveal misclassification patterns. This representation enables a more granular evaluation of model behavior, going beyond overall accuracy by highlighting how each model distributes prediction errors across negative, neutral, and positive sentiments.

The analysis shows that the CNN model performs strongly in classifying positive reviews, with a high concentration of correct predictions along the positive diagonal. However, it struggles significantly with neutral reviews, frequently misclassifying them as either positive or negative. This pattern suggests that CNN relies heavily on explicit lexical cues, making it less capable of distinguishing sentiment when emotional indicators are subtle or context-dependent. In contrast, IndoBERT demonstrates a more balanced classification pattern across classes. It achieves better recognition of neutral reviews, as reflected in a higher number of correct neutral predictions and fewer misclassifications than CNN. Although some neutral instances are still incorrectly assigned to the negative class, IndoBERT produces fewer false positives overall. This improvement highlights the advantage of its transformer-based architecture, which leverages self-attention mechanisms and contextualized word representations learned from large-scale Indonesian corpora. Consequently, IndoBERT is better equipped to interpret nuanced or implicitly expressed sentiment, particularly in informational or procedural financial reviews.

4.5. Error Analysis

An in-depth error analysis was conducted on 1,500 test samples to compare individual prediction outcomes. A total of 1,082 reviews (72.1%) were correctly classified by both models, indicating a high level of agreement. However, 183 reviews (12.2%) were misclassified by both models, suggesting inherent ambiguity or inconsistencies in the data. As shown in Table 3.

Table 3. Error Analysis Comparison

Category	Amount	Percentage
Both Correct	1082	72,1%
Both Incorrect	183	12,2%
Only CNN Correct	135	9,0%
Only IndoBERT Correct	100	6,7%

IndoBERT correctly classified 100 reviews (6.7%) that CNN misclassified, particularly in cases involving neutral or context-dependent expressions. Conversely, CNN correctly classified 135 reviews (9.0%) that IndoBERT misclassified, primarily those containing explicit sentiment keywords related to technical issues. These findings indicate that CNN is more sensitive to strong lexical cues, whereas IndoBERT excels at contextual comprehension.

4.6. Word Cloud Insights

Word clouds were generated based on term frequency after applying the same preprocessing pipeline used for model training, including stopword removal and stemming. Only terms with a minimum frequency threshold were visualized to reduce noise and highlight dominant lexical patterns.

Word cloud visualizations were generated to qualitatively examine the dominant terms within each sentiment class. Figure 5 illustrates that positive sentiment reviews were dominated by terms such as “bagus”, “mudah”, “aplikasi”, and “investasi”, reflecting user satisfaction with usability and features. Negative sentiment (Figure 6) reviews frequently contained terms such as “error”, “login”, “tidak bisa”, and “lama”, highlighting technical issues and service delays. Neutral sentiment (Figure 7) reviews mainly focused on administrative or

informational terms such as “akun”, “verifikasi”, and “transfer”. These visual patterns support the quantitative results and help explain the difficulty models, particularly CNNs, face in distinguishing neutral sentiment due to its lexical overlap with the positive and negative classes.

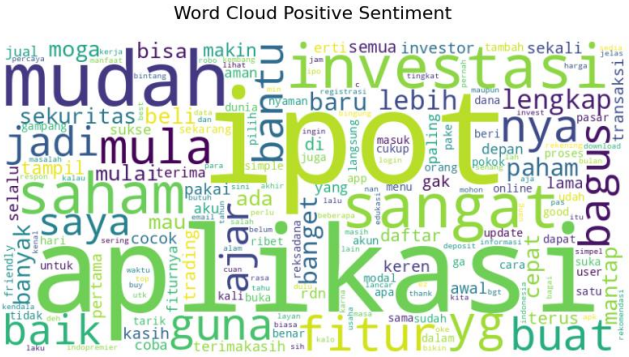

Figure 5. Word Cloud Visualization for Positive Sentiment

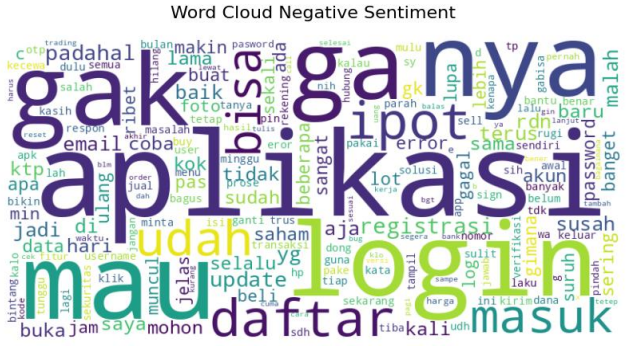

Figure 6. Word Cloud Visualization for Negative Sentiment

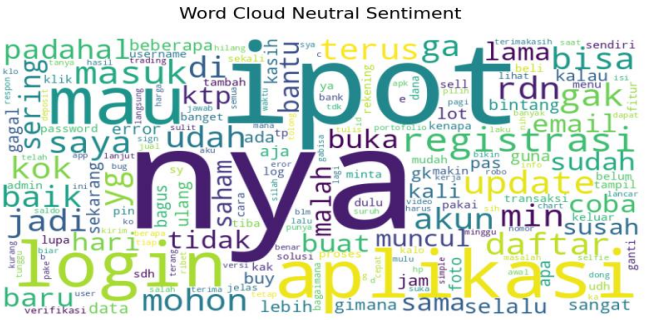

Figure 7. Word Cloud Visualization for Neutral Sentiment

5. Comparison

A comparative evaluation of CNN and IndoBERT shows that model performance in sentiment analysis of financial application reviews cannot be adequately assessed by overall accuracy alone. Although the CNN model achieved slightly higher accuracy on the test dataset, a more detailed analysis shows that IndoBERT captures contextual and semantically nuanced information more effectively, particularly for neutral sentiment classification. Neutral sentiment reviews in financial applications often contain informational or procedural content rather than explicit emotional expressions, such as inquiries about account verification, fund transfers, or transaction processes. These characteristics make neutral reviews inherently more challenging to classify because they often lack strong sentiment-bearing keywords. The

significantly higher recall achieved by IndoBERT for the neutral class indicates its advantage in modeling contextual dependencies and implicit semantic cues, which convolution-based architectures capture less effectively. These results indicate that sentiment analysis for financial applications should not rely solely on accuracy metrics, as performance on neutral sentiment is critical for understanding informational user feedback. In practical fintech applications, misclassifying neutral reviews can lead to misinterpretation of user needs. It may overlook operational issues or support-related inquiries that do not explicitly convey sentiment. Therefore, models with stronger contextual understanding, such as IndoBERT, offer a strategic advantage for comprehensive sentiment monitoring in financial platforms. CNN is effective at detecting dominant lexical patterns associated with strong positive or negative sentiment, while IndoBERT provides deeper semantic interpretation, which is essential for nuanced sentiment categories. A hybrid approach that integrates CNN's local feature extraction with IndoBERT's contextual representation may further enhance sentiment analysis performance in future financial application research.

Similar results from other journals show that CNN models are widely recognized for capturing local textual features through convolutional layers, making them effective for text classification tasks, including sentiment analysis. However, CNNs often struggle to capture long-range dependencies and deeper contextual meaning, which are important for accurately understanding sentiment. Previous studies show that CNNs achieve around 75% accuracy in app review sentiment analysis, which is still lower than that of BERT-based models [18]. Another study found that CNN struggled particularly with neutral sentiment, achieving 89.13% accuracy in political news sentiment analysis, which remained below IndoBERT's performance [21]. In contrast, IndoBERT, a BERT variant pre-trained on Indonesian-language corpora, demonstrates a strong ability to capture contextual meaning and complex linguistic patterns, resulting in superior sentiment analysis performance. IndoBERT achieved 92.93% accuracy in political news sentiment classification and showed significant advantages over CNN, particularly in neutral sentiment detection [21]. Furthermore, IndoBERT has proven highly effective in aspect-based sentiment analysis, outperforming other models, including CNN, in F1 score [22]. Comparative studies consistently indicate that BERT-based architectures outperform CNN and even LSTM in app review sentiment analysis tasks due to their superior contextual understanding [18]. Although hybrid approaches combining CNN with models such as LSTM or GRU have been developed to improve performance, IndoBERT generally remains superior in handling complex language structures and achieving higher accuracy [20]. Nevertheless, CNNs still have practical value in scenarios that prioritize computational efficiency and simplicity, as their architecture enables faster training and inference, making them suitable for real-time applications or environments with limited computational resources. Hybrid models that combine CNNs with sequential architectures such as LSTM or GRU can also provide a balanced solution by leveraging complementary strengths to improve overall sentiment analysis performance. In addition to classification performance, computational considerations are relevant for real-world implementation. Although detailed runtime analysis was not conducted in this study, transformer-based models such as IndoBERT generally require more computational resources and longer training times than CNN architectures. Future studies should incorporate systematic evaluation of training time, inference speed, and hardware requirements to provide a more comprehensive cost-performance comparison.

From a practical standpoint, the findings provide actionable insights for IPOT developers and other fintech platforms. For sentiment monitoring systems that require deep contextual understanding, particularly for identifying neutral or implicitly expressed feedback, IndoBERT is recommended due to its superior contextual modeling capabilities. Conversely, for systems prioritizing faster computation and lower resource consumption, especially in detecting strongly polarized positive or negative sentiment, CNN offers a more efficient alternative. Therefore, model selection should align with operational priorities, computational resources, and monitoring objectives within fintech environments.

6. Conclusions

The comparison between Convolutional Neural Networks (CNNs) and IndoBERT for sentiment analysis of IPOT financial application reviews shows that both models perform sentiment classification effectively, with different strengths across sentiment categories. The CNN model achieved slightly higher overall accuracy (0.8113) compared to IndoBERT (0.7880), indicating strong performance in detecting dominant sentiment patterns, particularly positive sentiment. Meanwhile, IndoBERT achieved superior performance in negative and neutral sentiment classification, as evidenced by higher recall and F1 scores. The confusion matrix and error analysis results further indicate that IndoBERT is more effective at understanding contextual and nuanced language, whereas CNN is more sensitive to explicit lexical sentiment indicators.

The results align with the research objective of comparing deep learning architectures for sentiment classification in Indonesian fintech application reviews. The findings support the argument that transformer-based models such as IndoBERT provide stronger contextual understanding, which is especially important for identifying neutral sentiment that often lacks explicit emotional keywords. At the same time, CNN remains competitive due to its efficiency and strong performance in classifying clearly expressed sentiment. Therefore, model selection should consider not only accuracy but also sentiment class distribution and real-world application requirements.

From a practical perspective, the findings contribute to the development of automated sentiment monitoring systems for financial applications. The results provide insights for fintech companies in selecting appropriate AI models to analyze user feedback. IndoBERT is better suited to comprehensive sentiment monitoring, where contextual understanding is critical, while CNN is better suited to systems that require faster computation and lower resource consumption. Academically, the findings contribute to the limited literature on Indonesian-language sentiment analysis in the financial investment application domain and provide a benchmark comparison between CNN and IndoBERT using real-world fintech datasets.

Several limitations remain. Sentiment labeling was based on rating scores, which may not always reflect the true sentiment expressed in text. The dataset was limited to a single financial application (IPOT), which may limit generalizability to other fintech platforms. Spam or fake-review filtering was not explored in depth, which could affect model performance. Additionally, the computational cost of the models was not evaluated in detail.

Future research can consider manual or hybrid sentiment labeling approaches to improve label accuracy. Expanding datasets to include multiple fintech applications would improve generalizability. Future studies can also explore spam detection, multilingual sentiment analysis, and cost-performance trade-off analysis. In addition, hybrid architectures that combine CNN feature extraction with IndoBERT contextual representation may improve sentiment classification performance and offer promising directions for further investigation.

Author Contributions: Conceptualization: I.G.N.A.P. and W.J.; Methodology: I.G.N.A.P.; Software: A.B.; Validation: I.G.N.A.P., W.J. and A.B.; Formal analysis: I.G.N.A.P.; Investigation: W.J.; Resources: W.J.; Data curation: A.B.; Writing original draft preparation: I.G.N.A.P.; Writing review and editing: I.G.N.A.P. and W.J.; Visualization: A.B.; Supervision: I.G.N.A.P.; Project administration: W.J; Funding acquisition: I.G.N.A.P

Funding: This research was funded by Warmadewa University through an internal research grant (Warmadewa University Grant).

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to thank Warmadewa University for its support of this research. The authors also acknowledge the institutional and technical support provided by the affiliated universities. This research utilized AI-assisted tools for language

refinement and editing, while the authors performed all scientific content, analysis, and interpretation.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

[1] M. Singh, "The Future of Investing : Analysing the Influence of Trading Applications on Investors Trends in Demat Accounts in 2022 (in mn)," vol. 07, no. 04, pp. 1755–1762, 2024, doi: 10.47191/ijmra/v7-i04-38.

[2] D. Das, R. Shenoy, and L. D, "The Role of Fintech in Increasing Retail Participation in the Indian Stock Market," *INTERNATIONAL J. Sci. Res. Eng. Manag.*, vol. 08, pp. 1–7, Nov. 2024, doi: 10.55041/IJSREM38135.

[3] N. Kurniasari and W. Wibowo, *Sentiment analysis of IPOT application reviews using naïve Bayes method*, vol. 2668. 2022. doi: 10.1063/5.0116704.

[4] R. Fuad Armansyah, "A Study Of Investor Financial Behavior on Online Trading System in Indonesian Stock Exchange: E-Satisfaction, E-Loyalty, And E-Trust," *J. Econ. Business, Account. Ventur.*, vol. 23, pp. 69–84, Jul. 2020, doi: 10.14414/jebav.v23i1.2176.

[5] A. Yasin, R. Fatima, A. N. Ghazi, and Z. Wei, "Python data odyssey: Mining user feedback from Google Play Store," *Data Br.*, vol. 54, p. 110499, 2024, doi: https://doi.org/10.1016/j.dib.2024.110499.

[6] S. F. Fahim, S. A. Sounok, N. Shaeed, M. H. Orpa, and N. T. Niloy, "Analyzing User Sentiment in Google Play Store Reviews: A Natural Language Processing Approach," in *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, 2024, pp. 1–5. doi: 10.1109/ICCCNT61001.2024.10725684.

[7] C. Zucco, B. Calabrese, G. Agapito, P. Guzzi, and M. Cannataro, "Sentiment analysis for mining texts and social networks data: Methods and tools," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 10, Aug. 2019, doi: 10.1002/widm.1333.

[8] F. Khanum, P. S. Lakshmi, and H. V. R. K, "Sentiment Analysis Using Natural Language Processing, Machine Learning and Deep Learning," in *2024 5th International Conference on Circuits, Control, Communication and Computing (I4C)*, 2024, pp. 113–118. doi: 10.1109/I4C62240.2024.10748425.

[9] K. Taghandiki and E. Ehsan, *Types of Approaches, Applications, and Challenges in the Development of Sentiment Analysis Systems*. 2023. doi: 10.48550/arXiv.2303.11176.

[10] S. Awasthi, "A Comprehensive Analysis of the Current Methods, Challenges and Innovations in Sentiment Analysis," *J. Inf. Syst. Eng. Manag.*, vol. 10, pp. 382–401, Mar. 2025, doi: 10.52783/jisem.v10i23s.3712.

[11] S. Redhu, "Sentiment Analysis Using Text Mining: A Review," *Int. J. Data Sci. Technol.*, vol. 4, p. 49, Jun. 2018, doi: 10.11648/j.ijdst.20180402.12.

[12] T. Purnomo and J. Sutopo, "Comparison of Pre-Trained BERT-Based Transformer Models for Regional Language Text Sentiment Analysis in Indonesia," *Int. J. Sci. Technol.*, vol. 3, pp. 11–21, Nov. 2024, doi: 10.56127/ijst.v3i3.1739.

[13] F. Wang, "Comparative Evaluation of Sentiment Analysis Methods : From Traditional Techniques to Advanced Deep Learning Models," vol. 0, pp. 23–29, 2024, doi: 10.54254/2755-2721/105/2024TJ0056.

[14] P. Sharma, P. Poonam, and A. Yadav, "TRANSFORMER SVS.CNN SIN MEDICAL IMAGING: A COMPARATIVE REVIEW," *Int. J. Adv. Res.*, vol. 13, pp. 8–11, Jul. 2025, doi: 10.21474/IJAR01/21300.

[15] Z. Zhu, "BERT and Its Applications in Natural Language Understanding," *Appl. Comput. Eng.*, vol. 175, pp. 99–105, Aug. 2025, doi: 10.54254/2755-2721/2025.AST26090.

[16] V. Balakrishnan, Z. Shi, C. L. Law, R. Lim, L. L. Teh, and Y. Fan, "A deep learning approach in predicting products' sentiment ratings: a comparative analysis," *J. Supercomput.*, vol. 78, no. 5, pp. 7206–7226, 2022, doi: 10.1007/s11227-021-04169-6.

[17] C. Lal and Z. Nasir, "Comparative Analysis of Deep Learning Methods in the Realm of Sentiment Analysis," in *2023 International Multi-disciplinary Conference in Emerging Research Trends (IMCERT)*, 2023, pp. 1–3. doi: 10.1109/IMCERT57083.2023.10075107.

[18] E. Subowo, "Implementasi Pembelajaran Mendalam dalam Klasifikasi Sentimen Ulasan Aplikasi: Evaluasi Model BERT, LSTM, dan CNN," *SURYA Inform.*, vol. 14, no. 2, pp. 66–70, 2024.

[19] A. A. Alfin, I. Kurniasari, and I. Yanuartanti, "Analisis Klasifikasi Sentimen Berbasis Topik pada Ulasan Layanan Dana dan Sakuku dengan Convolutional Neural Network," *Inf. (Journal Inform. dan Sist. Informasi)*, vol. 15, no. 2, pp. 225–236, 2023.

[20] I. P. A. A. Mahendra and K. Kusrini, "ANALISIS KOMPARATIF KLASIFIKASI SENTIMEN PENGGUNA APLIKASI INVESTASI MENGGUNAKAN ALGORITMA HYBRID CNN-LSTM, CNN-GRU DENGAN IMPLEMENTASI SMOTE (Comparative Analysis of Investment Application User Sentiment Classification Using Hybrid CNN- LSTM, CNN-GRU Alg," *TEKNIMEDIA*, vol. 6, no. 1, pp. 112–118, 2025.

[21] C. Ramadhan, V. Atina, and H. Permatasari, "Analisis Perbandingan Model CNN dan IndoBERT Dalam Sentimen Berita Politik Indonesia," in *Prosiding Seminar Nasional Teknologi Informasi dan Bisnis*, Jul. 2025, pp. 110–118. doi: 10.47701/v1r9ka69.

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