Hybrid IndoBERT–BiLSTM with Attention Mechanism for Sentiment Analysis in Imbalanced Digital Society Data

Aris Budiasono[1,*], Muhamad Khafid[2]

[1,2]Master of Computer Science, Amikom Purwokerto University, Indonesia

ABSTRACT

Sentiment analysis plays an important role in understanding public opinion within digital society, particularly from social media data that is often noisy and imbalanced. This study proposes a hybrid deep learning model that integrates IndoBERT, Bidirectional Long Short-Term Memory (BiLSTM), and an attention mechanism to improve sentiment classification performance. IndoBERT is utilized to capture contextual semantic representations, while BiLSTM models sequential dependencies in text data. The attention mechanism further enhances the model by focusing on important features within sequences. To address class imbalance, a weighted cross-entropy loss is applied during training. The model is evaluated using accuracy, precision, recall, and F1-score, along with confusion matrix and visualization analysis. Experimental results show that the proposed model achieves high overall performance, with an accuracy of 91.16% and a weighted F1-score of 0.9005. However, class-wise evaluation reveals limitations in handling minority classes, particularly the negative class, due to severe data imbalance. These findings highlight that while hybrid architectures can significantly improve performance, additional strategies are required to ensure balanced classification. This study contributes to the development of robust sentiment analysis models for digital society applications.

Keywords Sentiment Analysis, Digital Society, IndoBERT, Attention-Based Deep Learning, Imbalanced Data

INTRODUCTION

The rapid growth of digital society has led to an exponential increase in user-generated content on social media platforms. This massive volume of textual data provides valuable insights into public opinion, behavior, and sentiment across various domains such as politics, business, and social issues. Sentiment analysis, as a subfield of natural language processing (NLP), plays a crucial role in extracting and understanding these opinions automatically [1]. However, sentiment classification in social media data remains a challenging task due to the presence of informal language, noise, ambiguity, and highly imbalanced class distributions [2].

Recent advancements in deep learning have significantly improved sentiment analysis performance. Transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) have demonstrated strong capability in capturing contextual semantic information [3]. In particular, IndoBERT has been widely adopted for Indonesian language processing tasks due to its pretrained contextual representations. In addition, sequential models such as Bidirectional Long Short-Term Memory (BiLSTM) are effective in modeling word order and long-term dependencies in text. Furthermore, attention mechanisms have been introduced to enhance model performance by allowing the network to focus on the most relevant parts of the input sequence [4]. These approaches represent the current state-of-the-art in sentiment

Submitted 5 June 2026
Accepted 20 August 2026
Published 17 November 2026

Corresponding author
Aris Budiasono,
25MA41D010@student.amikompurwokerto.ac.id

Additional Information and Declarations can be found on page 280

DOI: 10.63913/jds.v2i4.45

Copyright
2026 Budiasono and Khafid

Distributed under
Creative Commons CC-BY 4.0

How to cite this article: A. Budiasono and M. Khafid, "Hybrid IndoBERT–BiLSTM with Attention Mechanism for Sentiment Analysis in Imbalanced Digital Society Data," *J. Digit. Soc.*, vol. 2, no. 4, pp. 271-282, 2026.

analysis, especially when applied individually or in simple combinations.

Despite these advancements, several research gaps remain. First, many existing studies rely solely on transformer-based models without incorporating sequential modeling, potentially limiting their ability to capture fine-grained temporal dependencies in text. Second, although hybrid architectures have been explored, limited work has specifically combined IndoBERT, BiLSTM, and attention mechanisms in a unified framework for Indonesian sentiment analysis. Third, the issue of class imbalance—commonly found in real-world social media datasets—has not been adequately addressed, often resulting in biased models that perform poorly on minority classes [5]. This limitation is critical, as accurate detection of minority sentiments (e.g., negative opinions) is essential for reliable decision-making in digital society applications.

To address these gaps, this study proposes a hybrid deep learning model that integrates IndoBERT for contextual representation, BiLSTM for sequential modeling, and an attention mechanism for feature weighting. Additionally, a weighted loss function is employed to mitigate the impact of class imbalance. The main contributions of this study are as follows: (1) the development of a hybrid Transformer–BiLSTM–Attention architecture tailored for Indonesian sentiment analysis, (2) the incorporation of class imbalance handling using weighted loss, and (3) a comprehensive evaluation using multiple performance metrics and visualization techniques. The proposed approach aims to provide a more robust and effective solution for sentiment analysis in digital society contexts.

Literature Review and Related Works

Sentiment analysis has become a critical research area in natural language processing, particularly in analyzing user-generated content within digital society. Traditional machine learning approaches such as Naïve Bayes and Support Vector Machines have been widely used for sentiment classification tasks; however, these methods often rely heavily on manual feature engineering and fail to capture deep semantic relationships in text [6], [7].

The emergence of deep learning has significantly improved sentiment analysis performance. Recurrent neural networks (RNN), particularly Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), have demonstrated strong capabilities in modeling sequential dependencies in textual data [8], [9]. These models are able to capture contextual information across sequences, making them suitable for sentiment classification tasks involving complex sentence structures. However, RNN-based models often struggle with long-range dependencies and may fail to focus on the most relevant parts of the input [10].

To address these limitations, attention mechanisms have been introduced to enhance model performance by assigning different weights to input tokens based on their importance [11], [12]. Attention-based models allow the network to selectively focus on informative features, leading to improved classification accuracy. Several studies have shown that integrating attention mechanisms with BiLSTM can significantly enhance sentiment classification performance [13].

More recently, transformer-based architectures such as BERT have achieved state-of-the-art results in various NLP tasks by capturing bidirectional contextual representations [14], [15]. These models leverage self-attention mechanisms to

understand semantic relationships within text more effectively than traditional sequential models. In the Indonesian context, IndoBERT has been developed as a pretrained language model specifically designed for Bahasa Indonesia, demonstrating strong performance in sentiment analysis and other NLP tasks [16].

Despite their effectiveness, transformer-based models alone may not fully capture sequential patterns in text. Therefore, hybrid models combining transformers with sequential architectures such as LSTM or BiLSTM have been proposed [17], [18]. These hybrid approaches aim to leverage the strengths of both contextual embedding and sequential modeling, resulting in improved performance compared to standalone models.

Another important challenge in sentiment analysis is class imbalance, which commonly occurs in real-world datasets, particularly in social media data. Imbalanced data can lead to biased models that favor majority classes while neglecting minority classes [19]. Various techniques have been proposed to address this issue, including class weighting, resampling methods, and specialized loss functions [20].

Based on the existing literature, it can be observed that while significant progress has been made in sentiment analysis using deep learning and transformer-based models, limited studies have explored the integration of IndoBERT, BiLSTM, and attention mechanisms in a unified framework while simultaneously addressing class imbalance. Therefore, this study aims to fill this gap by proposing a hybrid architecture that combines contextual understanding, sequential modeling, and attention mechanisms to improve sentiment classification performance in imbalanced digital society datasets.

Methodology

Research Framework

This study follows a structured research framework consisting of several stages, including data preprocessing, text representation, model development, training, and evaluation. The overall research workflow is illustrated in figure 1.

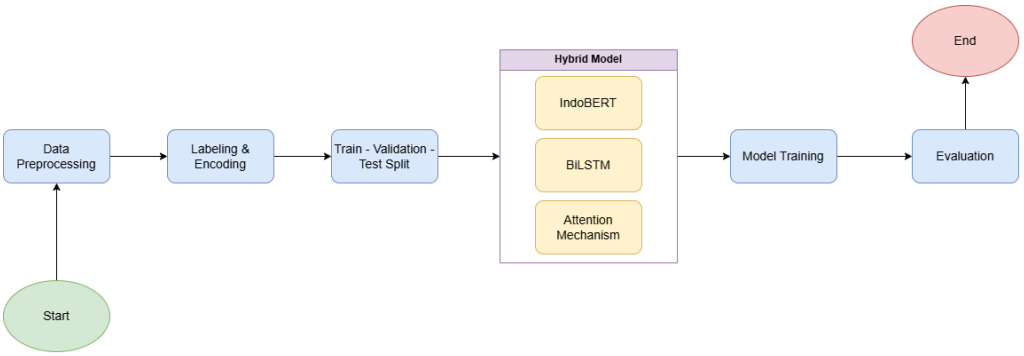


Figure 1 Research Framework of the Proposed Method

As shown in figure 1, the process begins with data cleaning and preprocessing, followed by tokenization using IndoBERT. The processed data is then fed into the proposed hybrid model consisting of IndoBERT, BiLSTM, and an attention mechanism. Finally, the model is evaluated using various performance metrics and visualization techniques.

Dataset

The dataset used in this study consists of textual data collected from various social media platforms, where user's express opinions, emotions, and reactions in an unstructured manner. Each text instance is manually or automatically labeled into three sentiment categories: positive, negative, and neutral, enabling supervised learning for sentiment classification tasks. The dataset captures the characteristics of real-world digital communication, including informal writing styles, slang, abbreviations, misspellings, emojis, and inconsistent grammar. Additionally, it contains noise such as redundant characters, hyperlinks, hashtags, and user mentions, which require preprocessing steps before analysis. This diversity and complexity make the dataset highly representative of actual online discourse, providing a realistic benchmark for evaluating the performance and robustness of sentiment analysis models.

Data Preprocessing

Text preprocessing was conducted to normalize and clean the raw textual data prior to analysis. This process involved several steps, including converting all text to lowercase to ensure consistency, removing URLs and hyperlinks that do not contribute to sentiment meaning, eliminating non-alphabetic characters such as numbers and symbols, and normalizing whitespace to reduce irregular spacing. These steps help reduce noise and improve the quality of the input data for modeling. Furthermore, the sentiment labels were standardized into three predefined classes, namely positive, negative, and neutral, and subsequently transformed into numerical representations using label encoding to facilitate their use in machine learning algorithms.

Data Splitting

The dataset was divided into three subsets, namely training, validation, and testing sets, using a stratified sampling approach to ensure that the distribution of sentiment classes remained consistent across all subsets. Specifically, 72% of the data was allocated to the training set, which was used to train the model, while 8% was assigned to the validation set for hyperparameter tuning and model selection. The remaining 20% constituted the testing set, which was used to evaluate the final performance of the model on unseen data. This splitting strategy helps maintain representativeness of each class and ensures a reliable assessment of the model's generalization capability.

Text Representation

Text data was tokenized using the IndoBERT tokenizer, which is specifically designed to handle Indonesian language characteristics and subword representations. Each input sequence was then processed to ensure a uniform length by applying padding and truncation, with a maximum sequence length set to 64 tokens. Sequences shorter than this limit were padded, while longer sequences were truncated to prevent exceeding the model's input capacity. The tokenizer outputs two main components, namely input IDs that represent the tokenized text and attention masks that indicate which tokens are actual data versus padding. These components are essential for enabling the model to correctly process and focus on relevant parts of the input during training and inference.

Proposed Model Architecture

The proposed model integrates four main components to effectively capture contextual, sequential, and important semantic features from the input text. First, a pretrained IndoBERT layer is utilized to generate rich contextual embeddings, where most of the layers are frozen to retain pretrained knowledge, while the last three layers are fine-tuned to adapt to the specific task. Second, a Bidirectional Long Short-Term Memory (BiLSTM) layer with a hidden size of 128 is employed to capture sequential dependencies from both forward and backward directions, allowing the model to better understand the context of each token within the sequence. Third, an attention mechanism is applied on top of the BiLSTM outputs to assign different weights to each token, enabling the model to focus more on important and sentiment-relevant words. Finally, the resulting representation is passed through a fully connected layer, where normalization and dropout with a rate of 0.3 are applied to reduce overfitting, followed by a dense layer that performs the final classification into sentiment categories.

Handling Class Imbalance

To address class imbalance, this study applies a weighted cross-entropy loss function, defined as follows:

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

$w_i$is the class weight,

$y_i$is the true label, and

$\hat{y}_i$is the predicted probability for class i.

This formulation ensures that minority classes receive higher importance during training.

Model Training

The model is trained using a carefully designed configuration to ensure stable convergence and optimal performance. The AdamW optimizer is employed due to its effectiveness in handling weight decay and improving generalization. Different learning rates are applied to different components of the model, where the IndoBERT layer uses a smaller learning rate of $2 \times 10^{-5}$to preserve pretrained knowledge, while the BiLSTM and classification layers use a higher learning rate of $1 \times 10^{-3}$to allow faster adaptation. The training process uses a batch size of 8 and runs for 6 epochs. Additionally, a linear learning rate scheduler with a warmup phase is implemented to gradually increase the learning rate at the beginning of training, which helps stabilize updates. To further enhance training stability, gradient clipping with a maximum norm of 1.0 is applied to prevent exploding gradients, and early stopping with a patience of 2 epochs is used to halt training when no improvement is observed, thereby reducing the risk of overfitting.

Evaluation Metrics

The model performance is evaluated using several standard classification metrics to provide a comprehensive assessment of its effectiveness. Accuracy is used to measure the overall proportion of correctly predicted instances across

all classes. Precision evaluates the correctness of positive predictions by measuring how many predicted instances for a class are actually relevant, while recall measures the model's ability to correctly identify all relevant instances within each class. The F1-score, which is the harmonic mean of precision and recall, is used to provide a balanced evaluation, especially in the presence of class imbalance. In addition to these metrics, a confusion matrix is utilized to give a detailed breakdown of correct and incorrect predictions across all sentiment classes, allowing for error analysis. Furthermore, the Receiver Operating Characteristic (ROC) curve is employed to assess the model's discriminative ability by illustrating the trade-off between true positive rate and false positive rate across different threshold settings.

Algorithm 1: Hybrid IndoBERT–BiLSTM with Attention

Input: Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$

Output: Predicted sentiment label $\hat{y}_i$

Process:

Start

Preprocess input text

$$x_i' = clean(x_i)$$

Tokenize text using IndoBERT tokenizer

$$(input_ids, mask) = Tokenizer(x_i')$$

Generate contextual embeddings

$$H = IndoBERT(input_ids, mask)$$

Apply BiLSTM for sequence modelling

$$Z = BiLSTM(H)$$

Compute attention weights

$$\alpha = softmax(W_a \cdot Z)$$

Compute context vector

$$c = \sum_{t=1}^{T} \alpha_t \cdot Z_t$$

Apply normalization and dropout

$$\tilde{c} = Dropout(LayerNorm(c))$$

Classification layer

$$o = W_c \cdot \tilde{c} + b_c$$
$$\hat{y}_i = softmax(o)$$

Compute weighted loss

$$L = -\sum_{j=1}^{C} w_j \cdot y_{i,j} \cdot \log(\hat{y}_{i,j})$$

Update model parameters

$$\theta = \theta - \eta \nabla_\theta L$$

Predict final label

$$\hat{y}_i = argmax(\hat{y}_i)$$

End

Results

Overall Performance

The proposed hybrid IndoBERT–BiLSTM with attention model was evaluated using multiple performance metrics, including accuracy, precision, recall, and F1-score, to ensure a comprehensive assessment of its classification capability.

The model achieved an overall accuracy of 91.16%, indicating a high proportion of correctly classified instances across all sentiment classes. In addition, it obtained a weighted precision of 0.8899, reflecting the reliability of its positive predictions, and a weighted recall of 0.9116, demonstrating its strong ability to identify relevant instances. The weighted F1-score of 0.9005 further confirms a balanced performance between precision and recall. Overall, these results suggest that the proposed model is effective and robust for general sentiment classification tasks, with strong predictive performance across different evaluation metrics.

Class-wise Performance

A detailed evaluation of classification performance for each class is presented in table 1. The results indicate that the model achieves its best performance on the neutral class, with the highest precision and recall, suggesting that neutral sentiments are more consistently and accurately identified. The positive class shows moderate performance, indicating that while the model can reasonably detect positive sentiments, there is still some level of misclassification. In contrast, the negative class is not well predicted, highlighting a significant limitation of the model in capturing minority class patterns. This issue is likely influenced by class imbalance, where fewer negative samples make it more difficult for the model to learn representative features, ultimately affecting its ability to generalize for that class.

Table 1 Classification Performance per Class

Class	Precision	Recall	F1-score	Support
Negative	0.00	0.00	0.00	3
Neutral	0.94	0.97	0.95	131
Positive	0.58	0.54	0.56	13

Confusion Matrix Analysis

The confusion matrix of the proposed model is shown in figure 2. As illustrated, all negative samples are misclassified into other classes, indicating that the model fails to correctly identify instances from the minority class. In contrast, the majority of neutral samples are accurately predicted, confirming that the model performs strongly on the dominant class. Additionally, some positive samples are incorrectly classified as neutral, suggesting overlap in feature representation between these two classes. Overall, these findings indicate that the model is biased toward the dominant class, which is likely caused by class imbalance and results in reduced performance for underrepresented classes.

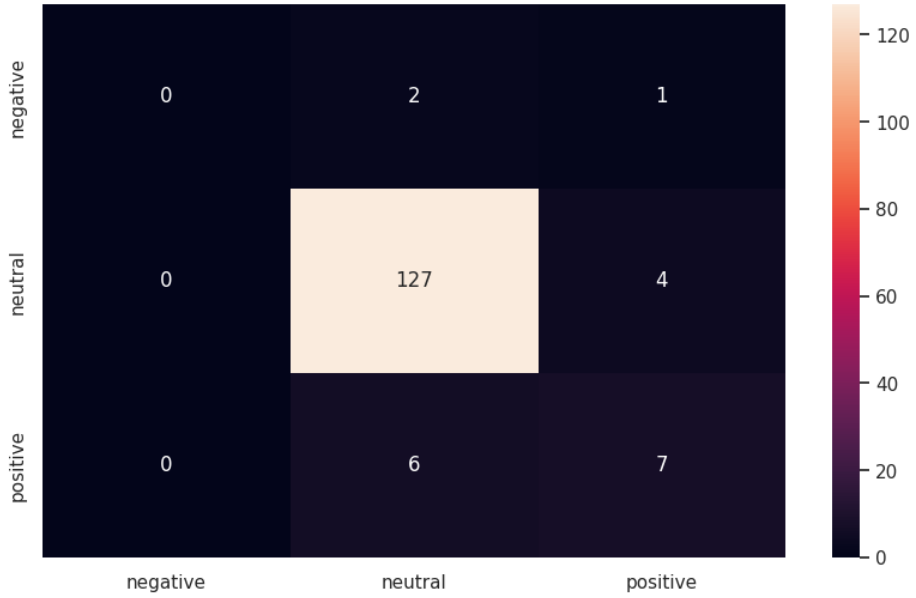


Figure 2 **Confusion Matrix of Proposed Model**

Training Performance

The training and validation loss curves are presented in figure 3. As shown in the figure, the training loss decreases steadily across epochs, indicating that the model is effectively learning patterns from the training data. Meanwhile, the validation loss remains relatively stable with only minor fluctuations, suggesting that the model maintains a consistent performance on unseen data. However, the gap between training and validation loss, along with the slight instability in validation loss, may indicate the onset of overfitting, particularly toward the majority class. This implies that while the model continues to improve on the training set, its ability to generalize, especially for underrepresented classes, may be limited.

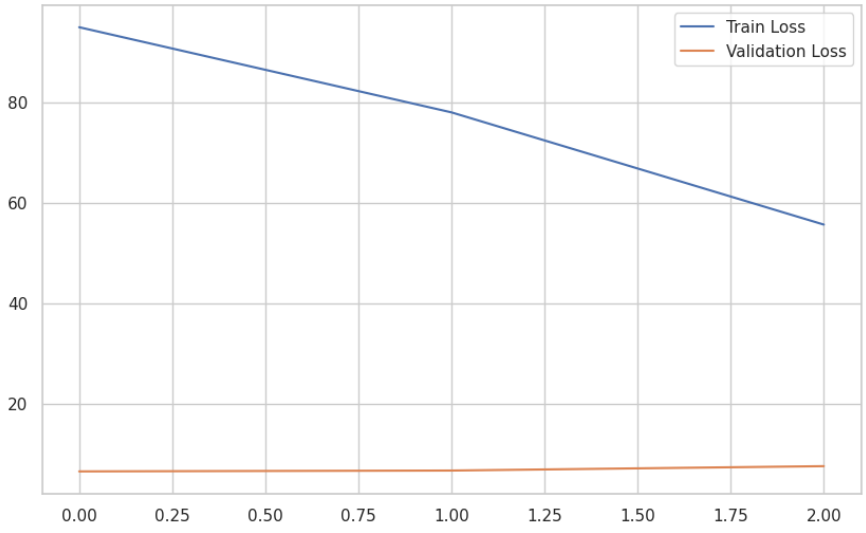


Figure 3 **Training Curve of the Model**

Data Distribution Analysis

The distribution of sentiment labels in the dataset is illustrated in figure 4. As shown, the dataset is highly imbalanced, with the neutral class comprising the majority of the samples, followed by the positive class, while the negative class represents only a small fraction of the data. This uneven distribution significantly impacts the model's learning process, as it tends to prioritize patterns from the dominant class while failing to adequately capture features of the minority class. Consequently, the model's ability to correctly classify negative sentiments is reduced, leading to biased predictions and lower performance on underrepresented classes.

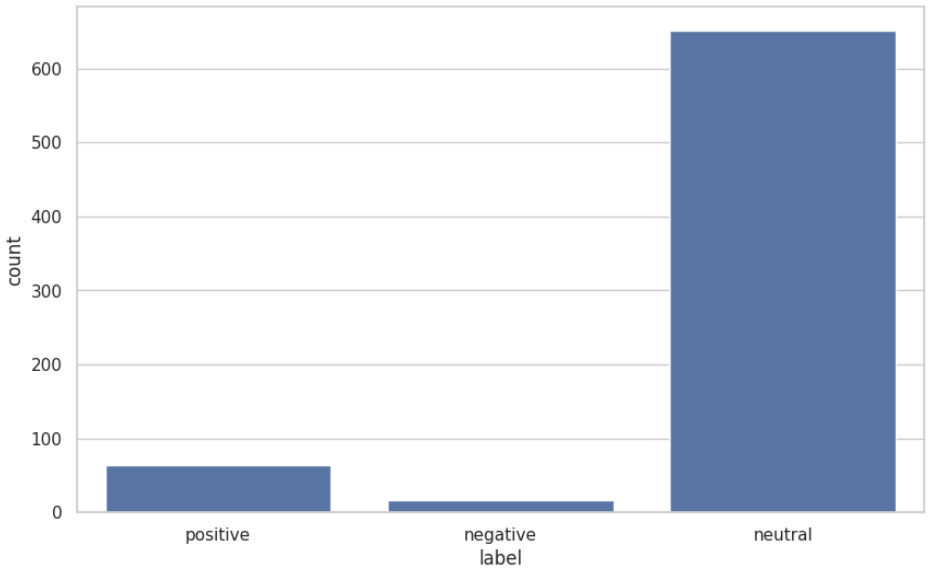


Figure 4 Label Distribution

Summary of Results

Overall, the proposed model achieves high accuracy and demonstrates strong performance on the majority class, particularly in correctly identifying neutral sentiments. However, a more detailed class-wise evaluation reveals a notable limitation in handling minority classes, especially the negative class. This weakness is primarily attributed to the severe class imbalance in the dataset, which causes the model to be biased toward the dominant class and limits its ability to learn meaningful representations for underrepresented samples. As a result, despite strong overall metrics, the model's performance is not equally robust across all sentiment categories.

Discussion

The experimental results demonstrate that the proposed hybrid IndoBERT–BiLSTM with attention model is effective in capturing both contextual and sequential features of text data, as reflected by the high overall accuracy and weighted F1-score. The integration of IndoBERT enables the model to understand semantic context in Indonesian language, while the BiLSTM layer enhances the modeling of word dependencies within sequences. Furthermore, the attention mechanism contributes by assigning higher importance to relevant tokens, allowing the model to focus on critical information during classification. This combination results in strong performance, particularly for the dominant neutral class, where the model achieves high precision and recall. The training

curve also indicates stable learning behavior, suggesting that the architecture is capable of converging efficiently during training.

However, a significant limitation is observed in the model's inability to correctly classify the negative class, as indicated by an F1-score of 0.00. This issue is primarily attributed to the severe class imbalance present in the dataset, where negative samples are extremely underrepresented compared to neutral and positive classes. Despite the use of weighted cross-entropy loss, the model remains biased toward the majority class, indicating that class weighting alone is insufficient in handling extreme imbalance scenarios. This finding highlights an important challenge in sentiment analysis within digital society data, where real-world datasets are often skewed. Therefore, future improvements should consider more advanced imbalance handling techniques, such as data augmentation, oversampling, or focal loss, to enhance minority class detection. These enhancements are expected to improve the robustness and generalization capability of the proposed model across all sentiment categories.

Conclusion

This study proposes a hybrid IndoBERT–BiLSTM model enhanced with an attention mechanism for sentiment analysis in digital society contexts. The experimental results demonstrate that the proposed architecture is effective in capturing both contextual semantics and sequential dependencies, achieving high overall performance with an accuracy of 91.16% and a weighted F1-score of 0.9005. The integration of transformer-based representation, sequential modeling, and attention enables the model to perform particularly well on the dominant class. However, the findings also reveal a critical limitation in handling severely imbalanced data, as the model fails to accurately classify the minority negative class despite the use of weighted loss. This highlights that high overall accuracy can be misleading in imbalanced datasets and emphasizes the need for more robust imbalance handling techniques. Therefore, future work should focus on incorporating advanced strategies such as data augmentation, resampling methods, and improved loss functions to enhance minority class detection and ensure more balanced and reliable sentiment classification performance.

Declarations

Author Contributions

Conceptualization: A.B., M.K.; Methodology: M.K.; Software: A.B.; Validation: A.B., M.K.; Formal Analysis: A.B., M.K.; Investigation: A.B.; Resources: M.K.; Data Curation: M.K.; Writing Original Draft Preparation: A.B., M.K.; Writing Review and Editing: M.K., A.B.; Visualization: A.B.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

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

References

[1] B. Liu, *Sentiment Analysis and Opinion Mining*, Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, May 2012, doi: 10.2200/S00416ED1V01Y201204HLT016.

[2] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, “New avenues in opinion mining and sentiment analysis,” *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, Mar.–Apr. 2013, doi: 10.1109/MIS.2013.30.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 2019, no. October, pp. 1-16, 2019, doi: 10.48550/arXiv.1810.04805.

[4] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, vol. 2015, no. September, pp. 1-15, 2015, doi: 10.48550/arXiv.1409.0473.

[5] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009, doi: 10.1109/TKDE.2008.239.

[6] P. D. Turney, “Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, PA, USA, 2002, vol. 2022, no. December, pp. 417–424, doi: 10.48550/arXiv.cs/0212032.

[7] T. Joachims, “Text categorization with support vector machines,” 1998, doi: 10.17877/DE290R-5097.

[8] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[9] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997, doi: 10.1109/78.650093.

[10] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar. 1994, doi: 10.1109/72.279181.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 2017, no. June, pp. 1-8, 2017, doi: 10.48550/arXiv.1706.03762.

[12] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical attention networks for document classification,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 2016, no. June, pp. 1480–1489, 2016, doi: 10.18653/v1/N16-1174.

[13] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, “Attention-based bidirectional long short-term memory networks for relation classification,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), vol. 2: Short Papers*, no. August, pp. 207–212, 2016, doi: 10.18653/v1/P16-2034.

[14] T. Wolf, L. Debut, V. Sanh, and J. Chaumond, “Transformers: State-of-the-art natural language processing,” in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP Demos)*, vol. 2020, no. October, pp. 38–45, 2020, doi: 10.18653/v1/2020.emnlp-demos.6.

[15] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, “What does BERT look at? An analysis of BERT’s attention,” in *Proceedings of the BlackboxNLP Workshop*, vol. 2019, no. June, pp. 1-11, 2019, doi: 10.48550/arXiv.1906.04341.

[16] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, “IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP,” in *Proceedings of the International Conference on Computational Linguistics (COLING)*, vol. 2020, no. November, pp. 1-14, 2020, doi: 10.48550/arXiv.2011.00677.

[17] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune BERT for text classification?” *arXiv preprint* arXiv:1905.05583, vol. 2019, no. May, pp. 1-10, 2019, doi: 10.48550/arXiv.1905.05583.

[18] K. L. Tan, C.-P. Lee, K. S. M. Anbananthen, and K. M. Lim, “RoBERTa-LSTM: A hybrid model for sentiment analysis with transformers and recurrent neural network,” *IEEE Access*, vol. 10, no. February, pp. 1–1, 2022, doi: 10.1109/ACCESS.2022.3152828.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. June, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[20] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.