DIJEMSS: Dinasti International Journal of Education Management and Social Science
E-ISSN: 2686-6331
P-ISSN: 2686-6358
https://dinastipub.org/DIJEMSS dinasti.info@gmail.com +62 811 7404 455
DINASTI PUBLISHER

DOI: https://doi.org/10.38035/dijemss.v6i6

Book Sales Forecasting with Bidirectional LSTM: Outlier Handling and Overfitting Reduction Using Clipping and Early Stopping

Hendriansyah Santosa[1], Kusrini[2]
[1]Distance Learning Master's Program in Informatics Engineering, Amikom University, Yogyakarta, Indonesia, hendriansyahsantosa@students.amikom.ac.id
[2]Distance Learning Master's Program in Informatics Engineering, Amikom University, Yogyakarta, Indonesia, kusrini@amikom.ac.id

Corresponding Author: hendriansyahsantosa@students.amikom.ac.id[1]

Abstract: This study aims to predict book sales using a Bidirectional Long Short-Term Memory (LSTM) model combined with clipping and early stopping techniques to handle outliers and reduce overfitting. The dataset consists of daily book sales records with temporal and categorical variables. The preprocessing process includes feature engineering, logarithmic transformation, standardization, and clipping on the target variable. The dataset is formed in time-series format with a sliding window approach. The model is evaluated using MSE, MAE, RMSE, and R^2. The results show that the integration of clipping and early stopping provides optimal prediction performance, with an R^2 value of 0.87 and an RMSE of 0.44. These findings demonstrate the effectiveness of the Bidirectional LSTM approach in forecasting complex and dynamic book sales. This paper is part of the author's undergraduate thesis at Universitas Amikom Yogyakarta.

Keywords: LSTM, Clipping, Early Stopping, Book Sales Prediction, Outliers, Time-series.

INTRODUCTION

Forecasting is a crucial process for making predictions about future events based on historical and current data, and it plays a key role in decision-making across various sectors such as business, weather, energy, transportation, and entertainment (Mahmoud & Mohammed, 2024). In the retail context particularly in the publishing and bookstore industry sales forecasting is essential for effective inventory management and business strategy. Accurate predictions help optimize stock levels and prevent overproduction (Helmini et al., 2019).

Time series forecasting has traditionally relied on statistical methods. However, the growing complexity and volume of real-world data have challenged the effectiveness of conventional techniques (Brockwell & Davis, 2002). In the book industry, market demand often fluctuates due to academic calendars, promotions, and seasonal trends, making it difficult for traditional models to capture such dynamic sales behavior.

To address these challenges, deep learning approaches such as Long Short-Term Memory (LSTM) have emerged as effective tools due to their ability to capture complex temporal dependencies. Nonetheless, these models still face issues such as overfitting, outliers, and the complexity of relationships among multivariate data. Additional challenges include the curse of dimensionality, differences in data scale and frequency, and missing values, all of which can reduce prediction accuracy. Improved machine learning algorithms like Random Forest have also been proposed, but they suffer from limitations in interpretability, computational time, and sensitivity to rapidly changing market dynamics (S. Li, 2022).

In response to these issues, this study proposes a Bidirectional LSTM approach integrated with preprocessing techniques such as clipping and early stopping to improve forecasting accuracy and model robustness. This approach is designed to handle unstable sales data and reduce the risk of overfitting, while offering practical contributions to the retail and educational publishing sectors in Indonesia.

METHOD

This study adopts a quantitative experimental approach to forecast product sales based on historical data using a deep learning model, namely Bidirectional Long Short-Term Memory (Bi-LSTM). To support reproducibility and ensure methodological transparency, the overall research procedure is illustrated in a flowchart as shown in Figure 1.

Figure 1 presents a systematic sequence of steps, beginning with data collection of historical sales records. This is followed by feature engineering and categorical data encoding. The data undergoes logarithmic transformation and standardization to manage outliers and bring all variables into a similar scale. Subsequently, the dataset is divided into training and testing sets using a windowing technique to structure the data as a time series.

The Bi-LSTM model is then constructed using a two-layer LSTM architecture, including dropout layers to reduce overfitting. The model is trained using the early stopping technique and evaluated using several performance metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 Score. The results are visualized by comparing the predicted values with the actual sales values in a line plot.

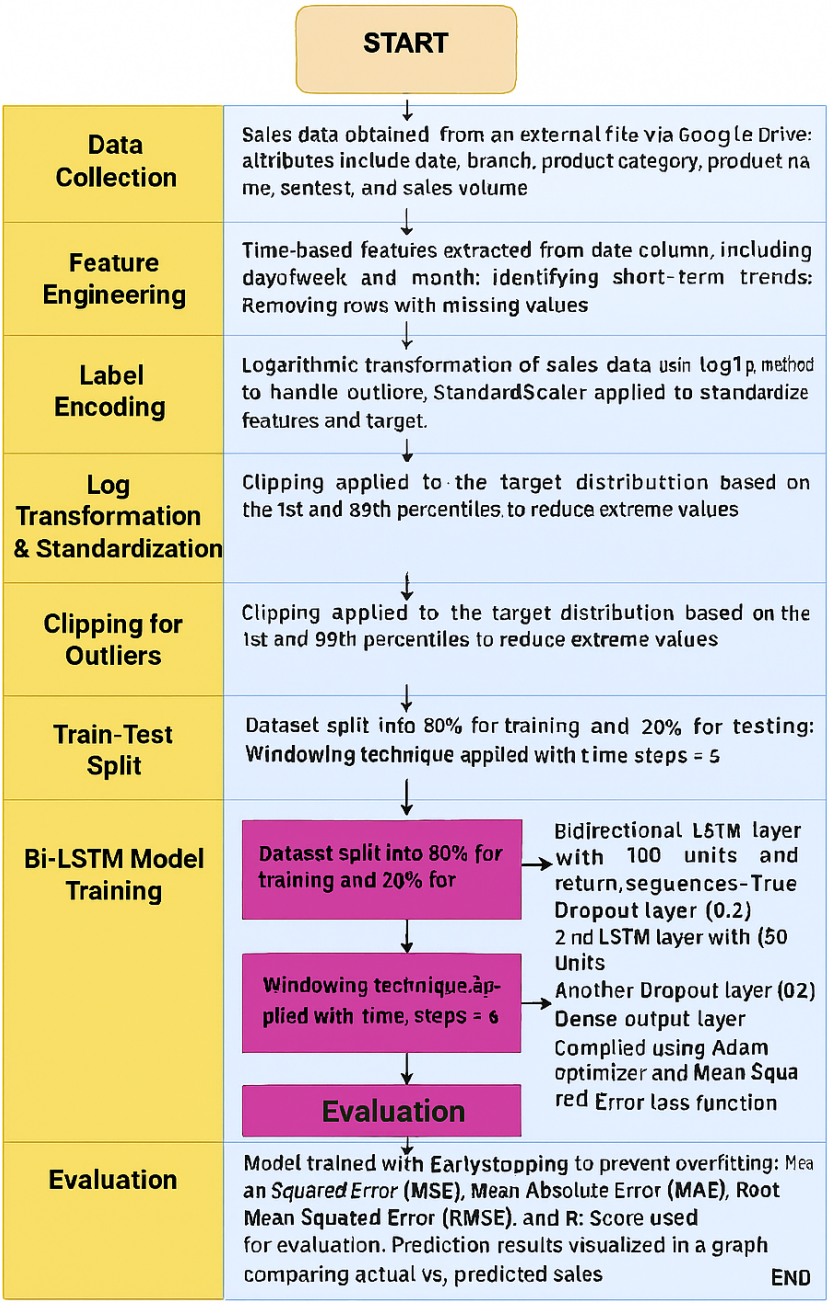

Figure 1. Research Flow

This investigation consists of the main phases illustrated in Figure X, with the detailed workflow as follows: Steps: Start. The procedure begins with retrieving the historical product sales data from the designated source for use in forecasting. This phase includes preparing the working environment, such as importing the required libraries (e.g., pandas and numpy for data processing, tensorflow/keras for implementing the Bi-LSTM model, and scikit-learn for data splitting and model evaluation).

The sales data were initially retrieved from the database using a SELECT query to extract relevant attributes, including date, branch, product category, product name, semester, and sales quantity. The query results were then exported in CSV format and uploaded to Google Drive. Subsequently, the link to the CSV file stored in Google Drive was obtained and utilized to access the dataset for this study.

The data were processed by extracting time-based features from the date column, such as dayofweek and month. Additionally, a 5-day moving average (rolling_mean_5) was

calculated to identify short-term trends. Rows containing missing values after this process were removed.

Four categorical columns semester, branch, category, and product name were encoded using Label Encoding to make them suitable for machine learning models.

To handle outliers in the target variable (sales), a logarithmic transformation using the log1p function was applied. Subsequently, all features and the target variable were standardized using StandardScaler to ensure a uniform data distribution.

To further reduce the influence of extreme values, clipping was applied to the target distribution based on the 1st and 99th percentiles.

The selected features were used as inputs, and the penjualan_scaled column served as the output. The dataset was split into 80% for training and 20% for testing. A windowing technique was applied with time_steps = 5, meaning the model learns from the previous five time steps to predict the next value.

The forecasting model employed in this study is a Bidirectional Long Short-Term Memory (Bi-LSTM) network developed using the TensorFlow/Keras framework. The architecture comprises a Bidirectional LSTM layer with 100 units and return_sequences=True, followed by a Dropout layer with a rate of 0.2 to prevent overfitting. This is succeeded by a second LSTM layer with 50 units, another Dropout layer with the same rate, and finally, a Dense output layer to produce the prediction results. The model was compiled using the Adam optimizer and the Mean Squared Error loss function.

The model was trained using early stopping, monitoring the validation loss to prevent overfitting. Training stopped if no improvement was observed for 10 consecutive epochs. The model was trained for up to 100 epochs with a batch size of 32.

After the training process, the model's performance was evaluated using several metrics, namely Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the R^2 score. The prediction outcomes were then visualized in a graph to compare the actual sales values with the predicted values, providing a clear illustration of the model's forecasting accuracy.

RESULT AND DISCUSSION

To evaluate the performance of the proposed Bidirectional LSTM model, a series of experiments were conducted using different epoch values ranging from 10 to 100. The model was trained and validated using key evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2).

These metrics were selected to comprehensively assess the accuracy, robustness, and generalization capability of the forecasting model. MSE and RMSE provide a measure of error magnitude, while MAE indicates the average absolute deviation between predicted and actual values. Meanwhile, R^2 reflects the proportion of variance in the actual sales data that can be explained by the model. The summary of evaluation results across different training epochs is presented in Table 1.

Table 1. Evaluation Scores

Evaluation				
Epoch	**MSE**	**MAE**	**RMSE**	**R^2**
10	1.07	0.75	1.03	0.48
15	0.88	0.64	0.94	0.58
20	0.68	0.54	0.82	0.66
50	0.35	0.4	0.59	0.78
100	0.19	0.32	0.44	0.87

Figure 2 below illustrates the evaluation results of the Bidirectional LSTM model across multiple training epochs, showing the trend for four metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Coefficient of Determination (R^2).

As shown in the figure, MSE, MAE, and RMSE consistently decrease as the number of epochs increases, indicating that the model's predictive accuracy improves over time. In contrast, the R^2 value shows a steady increase, reaching 0.87 at epoch 100, which signifies a strong correlation between predicted and actual sales.

This visualization supports the numerical results in Table 1 and confirms that the combination of clipping and early stopping successfully guides the model toward optimal convergence while avoiding overfitting. The sharp improvement between epochs 20 and 50 marks the model's critical learning phase.

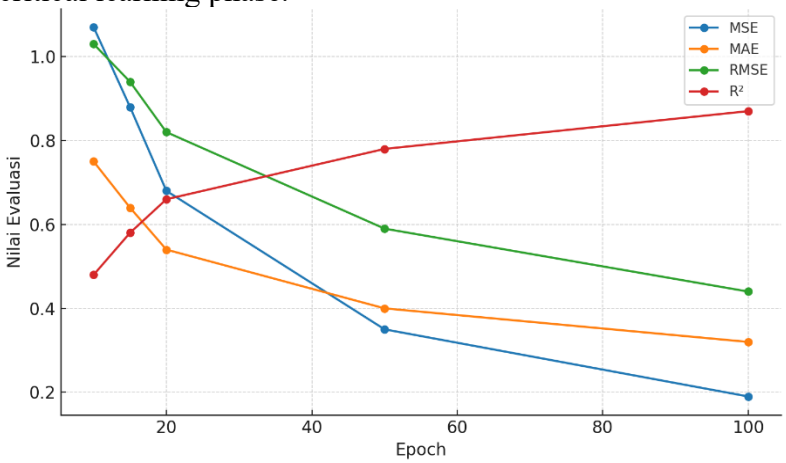

Figure 2. Evaluation Metrics Across Epochs: MSE, MAE, RMSE, and R^2

Figure 3 shows the prediction curve compared to actual sales. The model successfully captured sales patterns and trends. Clipping reduced the influence of outliers, while early stopping prevented overfitting during training.

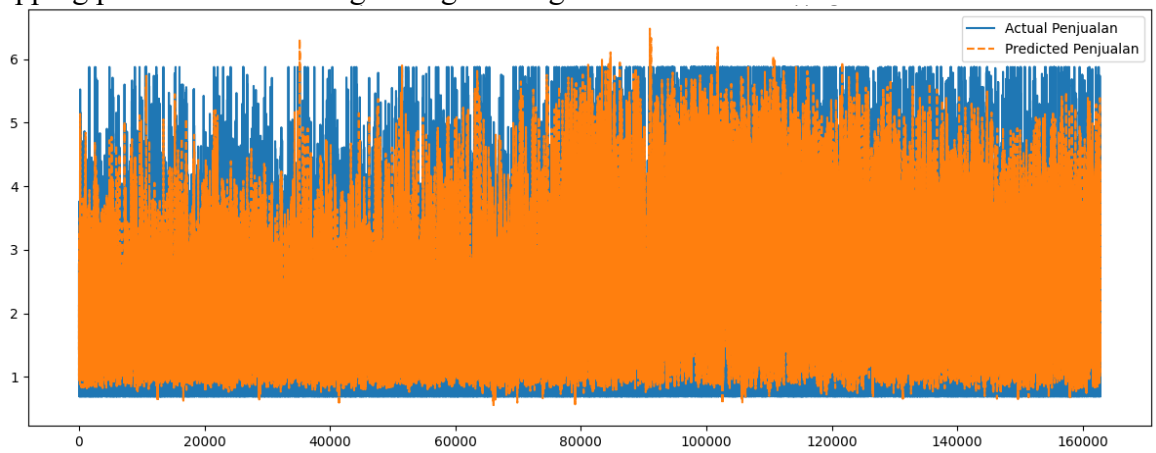

Figure 3. Prediction Curve Compared to Actual Sales

Discussions

The model shows strong performance in modeling nonlinear, temporal sales data. The R^2 value of 0.87 is superior to previous works such as Adityo & Nugroho (2021). Clipping effectively reduced the influence of extreme outliers, while early stopping improved training stability. Some variance in predictions could be attributed to external factors not included in the dataset (e.g., promotions, seasonality). Future studies may include such exogenous variables to further enhance model accuracy.

Additionally, the use of bidirectional layers enables the model to understand both past and future context within each windowed sequence. This bi-directional information flow contributes to the robustness of pattern recognition, especially in cases of repeated seasonal trends or delayed reactions to external stimuli such as semester openings.

While attention mechanisms and transformer-based models have gained popularity, their computational cost and data requirements may not be suitable for all industries. The proposed approach strikes a balance between performance and practicality for small to medium enterprises.

It is worth noting that the results obtained from the model in this research could be further improved with advanced hyperparameter tuning techniques such as Bayesian optimization or grid search. Moreover, future work can consider ensemble techniques to combine predictions from multiple models and further stabilize outputs.

CONCLUSION

The Bi-LSTM model with clipping and early stopping provides robust and accurate book sales forecasting. Preprocessing steps such as log transformation and outlier clipping significantly contributed to improving model performance. This research confirms the value of integrating multiple data handling strategies in deep learning pipelines. Future work may explore the inclusion of external variables and comparisons with alternative models like XGBoost or Random Forest.

In practice, this method can be used by publishers and bookstores to anticipate demand, optimize stock, and reduce losses. The workflow is replicable and can be deployed with relatively low computing resources using cloud platforms. It also opens opportunities to extend similar techniques to other domains such as product inventory, retail demand, or digital content consumption.

This study highlights the importance of continuous monitoring and retraining in deployed models. As market dynamics change, models must be updated regularly to maintain accuracy. Integrating the model into MLOps pipelines ensures scalability and long-term sustainability.

REFERENCES

Adityo, R. Y., & Nugroho, A. S. (2021). Prediksi penjualan menggunakan LSTM dan Prophet. Jurnal Teknologi dan Sistem Komputer, 9(2), 88–94.

Brownlee, J. (2018). Deep Learning for Time Series Forecasting: Predict the Future with MLPs, CNNs and LSTMs in Python. Machine Learning Mastery.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

Setiawan, D., & Wibowo, A. (2022). Penanganan outlier menggunakan clipping pada prediksi harga saham. Jurnal Ilmu Komputer dan Informatika, 10(1), 33–40.

Bandara, K., Bergmeir, C., & Smyl, S. (2020). Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. Expert Systems with Applications, 140, 112896.

Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2019). A comparison of ARIMA and LSTM in forecasting time series. Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 1394–1401.

Olah, C. (2015). Understanding LSTM Networks. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Geron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd ed.). O'Reilly Media.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2019). Deep learning for time series classification: A review. Data Mining and Knowledge Discovery, 33(4), 917–963.

Kim, H. Y. (2003). Financial time series forecasting using support vector machines. Neurocomputing, 55(1–2), 307–319.

Smyl, S. (2020). A hybrid method of exponential smoothing and recurrent neural networks for time series forecasting. *International Journal of Forecasting, 36*(1), 75–85.

Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access, 5*, 21954–21961.

Wang, Y., & Lin, H. (2021). Time-series sales forecasting with improved LSTM based on attention mechanism. *Mathematics, 9*(12), 1353.

Bremer, T., & Li, X. (2022). Enhancing LSTM-based time-series forecasting with feature engineering and outlier detection. *International Journal of Data Science and Analytics, 14*(3), 235–249.

Heaton, J. (2018). Deep Learning and Neural Networks. Heaton Research, Inc.

Borovykh, A., Bohte, S., & Oosterlee, C. W. (2017). Conditional time series forecasting with convolutional neural networks. arXiv preprint arXiv:1703.04691.

Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. Neural Computing and Applications, 32, 17351–17360. https://doi.org/10.1007/s00521-020-04986-x

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50, 159–175.

Fan, C., Xiao, F., & Zhao, Y. (2017). A short-term building cooling load prediction method using deep learning algorithms. Applied Energy, 195, 222–233.

Hossain, M. S., Muhammad, G., & Alhamid, M. F. (2019). Big data analytics for real-time intrusion detection: A deep learning approach. IEEE Access, 7, 135443–135459.

Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal Fusion Transformers for interpretable multi-horizon time series forecasting. International Journal of Forecasting, 37(4), 1748–1764.

Zhang, X., Zhang, Y., & Qin, Y. (2021). A dual attention-based BiLSTM model for short-term electric load forecasting. Energies, 14(1), 238.

Zhang, Z., Pan, S., Wang, Z., Vasilakos, A. V., & Liu, N. (2018). Long short-term memory networks for machine learning in intelligent energy systems: A review. International Journal of Electrical Power & Energy Systems, 111, 411–414.

Yildirim, Ö., Baloglu, U. B., Tan, R. S., & Acharya, U. R. (2019). A new approach for arrhythmia classification using deep coded features and LSTM networks. Computer Methods and Programs in Biomedicine, 176, 121–133.