Comparative Study of Deep Learning Architectures for Financial Sentiment Analysis in Digital Market Environments

Lakshmi Dandapani[1,*], Manimegalai Subramanian[2]

[1,2]Department of Electrical and Electronics Engineering, Faculty of EEE, AMET University, Chennai, India
[2]Department of Electrical and Electronics Engineering, Faculty of EEE, P.T. Lee Chengalvaraya Naicker College of Engineering and Technology, Kanchipuram, India

ABSTRACT

This study investigates the effectiveness of deep learning architectures for financial sentiment classification using textual data from financial reports and market statements. Four models, namely Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN), were trained and evaluated on a dataset of 5,842 labeled financial sentences categorized as positive, negative, or neutral. The evaluation employed accuracy and weighted F1-score metrics to assess predictive performance and class balance. The results revealed that the CNN model achieved the highest performance with an accuracy of 70.32 percent and an F1-score of 70.22 percent, followed by the BiLSTM model with 67.49 percent accuracy and 66.86 percent F1-score. The CNN's superior performance demonstrates its ability to capture local sentiment-bearing features efficiently, while BiLSTM excels in understanding bidirectional contextual dependencies. In contrast, the LSTM and GRU models performed less effectively due to their reliance on sequential memory mechanisms, which are less applicable to short and factual financial sentences. These findings highlight that CNN provides the best balance between accuracy, efficiency, and adaptability for financial sentiment classification and can serve as a foundation for intelligent digital market analysis systems and AI-driven financial decision support tools.

Keywords Financial Sentiment Analysis, Deep Learning, CNN, BiLSTM, Digital Market

INTRODUCTION

The rapid growth of digital financial ecosystems has increased the availability of financial text from news portals, trading platforms, corporate reports, and social media. Such text contains information about investor sentiment that can support financial analysis and market-related decision making. Financial sentiment analysis therefore plays an important role in identifying positive, negative, and neutral opinions from financial text [1], [2]. However, financial language is often domain-specific, concise, and context-dependent, making sentiment difficult to identify using simple textual features.

Traditional machine learning methods, including Support Vector Machines (SVM), Naïve Bayes, and Random Forest, have been widely applied to financial sentiment classification [3], [4]. These methods generally rely on handcrafted features such as n-grams and term frequency-inverse document frequency (TF-IDF), which may not adequately capture semantic relationships and contextual dependencies.

Submitted: 10 February 2026
Accepted: 20 May 2026
Published: 16 September 2026

Corresponding author
Lakshmi Dandapani,
lakshmi.d@ametuniv.ac.in

Additional Information and Declarations can be found on page 249

DOI: 10.47738/jdmdc.v3i3.83

How to cite this article: L. Dandapani and M. Subramanian, "Comparative Study of Deep Learning Architectures for Financial Sentiment Analysis in Digital Market Environments," *J. Digit. Mark. Digit. Curr.*, vol. 3, no. 3, pp. 229-243, 2026.

As a result, deep learning methods have increasingly been adopted because they can automatically learn relevant representations from text [5], [6].

Several deep learning architectures have shown promising results for financial sentiment analysis. Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) can capture sequential dependencies, while Bidirectional LSTM (BiLSTM) considers contextual information from both directions. Convolutional Neural Network (CNN), in contrast, is effective in extracting local patterns and important phrases. Previous studies have reported competitive performance from CNN, LSTM, and GRU in financial sentiment classification [7], [8], [9].

Despite these developments, direct comparisons of LSTM, BiLSTM, GRU, and CNN under the same dataset and experimental conditions remain limited. Differences in datasets, preprocessing methods, and evaluation settings make it difficult to determine how architectural characteristics affect classification performance. Recent studies also indicate that financial sentiment classification can benefit from improved text representations and contextual models [10], [11], [12]. Therefore, a controlled comparison of major deep learning architectures is needed to provide clearer empirical evidence for financial sentiment classification.

This study compares LSTM, BiLSTM, GRU, and CNN for classifying financial sentiment using 5,842 sentences labeled as positive, negative, and neutral. Accuracy and weighted F1-score are used to evaluate model performance across the three classes. The study aims to examine the effectiveness of each architecture and provide empirical evidence on the relationship between model architecture and financial sentiment classification performance. The findings are expected to support the development of automated financial text analysis systems and provide a foundation for future research using attention-based and transformer-based approaches.

Literature Review and Related Works

Financial sentiment analysis has become an important component of digital market research as investors and financial institutions increasingly use textual information to assess market conditions. Early approaches relied on lexicon and rule-based methods, but their ability to handle ambiguity and financial language complexity was limited [6], [7]. Traditional machine learning methods such as Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression were subsequently adopted using bag-of-words and term frequency-inverse document frequency (TF-IDF) features [8], [9]. However, their dependence on manually engineered features limited their ability to capture semantic and contextual relationships in financial text [10].

The development of deep learning improved financial sentiment analysis by enabling models to learn representations directly from text. Recurrent neural networks (RNNs) were initially used to model sequential dependencies but were limited by vanishing gradient problems [11], [12]. Long short-term memory (LSTM) addressed this limitation through gated mechanisms for retaining long-term information [13], while gated recurrent unit (GRU) provided a simpler and more computationally efficient alternative [14]. Bidirectional LSTM (BiLSTM) further improved contextual representation by processing text in both directions [15].

Convolutional neural networks (CNNs) have also been widely applied to financial text classification because of their ability to extract local features and sentiment-bearing phrases [16]. CNNs can effectively identify important expressions in concise financial statements and have demonstrated competitive performance compared with recurrent models [17], [18]. Hybrid approaches combining CNN with recurrent architectures have also been explored to capture both local and sequential textual features [19].

Despite these advances, systematic comparisons of LSTM, BiLSTM, GRU, and CNN under identical experimental conditions using the same financial dataset remain limited. Differences in datasets, preprocessing, and evaluation settings make it difficult to determine the relative effectiveness of these architectures, particularly for neutral and weakly polarized financial texts. Therefore, this study empirically compares LSTM, BiLSTM, GRU, and CNN using consistent preprocessing, identical experimental settings, and standardized evaluation metrics to assess their performance in financial sentiment classification and support the development of AI-based financial text analysis systems.

Methodology

This study uses a quantitative experimental approach to compare four deep learning architectures, Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN), for financial sentiment classification. As shown in figure 1, the research consists of five stages: dataset preparation, text preprocessing, word embedding, model training, and performance evaluation. All models were trained under identical configurations to ensure a fair and reproducible comparison. The experiments were implemented in Python 3.10 using TensorFlow 2.11 and Keras, with NumPy, Pandas, and Scikit-learn for data processing and evaluation, and Matplotlib and Seaborn for visualization. Experiments were conducted on a workstation with an Intel Core i7 processor, 16 GB RAM, and NVIDIA RTX 3060 GPU.

The dataset contains 5,842 financial sentences collected from publicly accessible financial news, market reports, and corporate announcements. Each sentence was manually labeled as positive,

negative, or neutral, representing favorable outcomes, adverse conditions, or factual statements, respectively. After removing duplicates and inconsistent entries, the data were divided into training and testing sets using an 80:20 ratio, with a random seed of 42 for reproducibility.

Text preprocessing included lowercasing, removal of punctuation and symbols, tokenization using the Keras Tokenizer, and vocabulary limitation to 10,000 words. Out-of-vocabulary terms were mapped to an <OOV> token, while sequences were padded or truncated to 50 tokens. Sentiment labels were encoded using Scikit-learn's LabelEncoder and converted into one-hot vectors for multiclass classification.

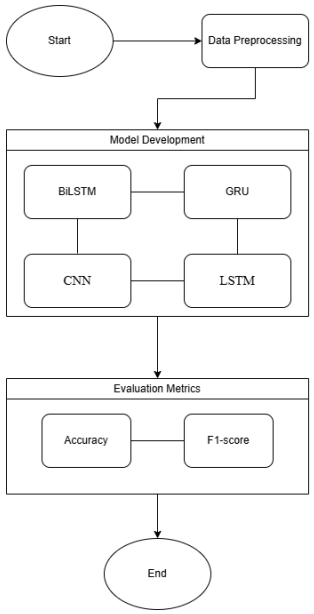

Figure 1 Research Steps

The data representation relied on a trainable word embedding layer that projected words into dense vector spaces. The embedding process transforms each input token into a 128-dimensional vector, allowing the model to learn semantic and contextual relationships among words. Mathematically, this process is defined as:

$$E = f(W \cdot X)X \in R^{n \times t} \quad (1)$$

represents the input matrix containing sequences of word indices for n samples of sequence length t, $W \in R^{|V| \times d}$ denotes the embedding weight matrix for vocabulary size $|V|$ and embedding dimension $d = 128$, and $f(\cdot)$ is the mapping function that produces dense embeddings. This representation enables models to identify contextual similarity between semantically related financial terms such as "growth," "increase," and "profit," thereby enhancing their ability to detect sentiment.

Four deep learning models were developed based on this embedding representation. The LSTM model employs gated memory cells to retain important sequential information over time. The core computations of an LSTM unit can be expressed as:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tilde{C}_t \\ = tanh(W_C[h_{t-1}, x_t] + b_C) C_t = f_t * C_{t-1} + i_t * \tilde{C}_t h_t \\ = tanh(C_t) f_t \quad (2)$$

, i_t, and $\tilde{C}_t$represent the forget, input, and candidate memory gates respectively; $C_t$is the updated cell state; and $h_t$is the hidden output at time t. The BiLSTM extends this mechanism by processing sequences in both directions, capturing dependencies from preceding and succeeding words within a sentence. The GRU simplifies the gating process by combining the forget and input gates into a single update gate, computed as:

$$z_t = \sigma(W_z[h_{t-1}, x_t]) r_t = \sigma(W_r[h_{t-1}, x_t]) \tilde{h}_t = tanh(W_h[r_t * h_{t-1}, x_t]) h_t \\ = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (3)$$

This configuration allows GRU to achieve faster convergence with fewer parameters compared to LSTM while maintaining competitive accuracy.

In contrast, the CNN model extracts local features through convolutional filters that detect sentiment-relevant word groups. The convolutional operation is mathematically expressed as:

$$h_i = ReLU(W_c * x_{i:i+k-1} + b) W_c \quad (4)$$

is the convolutional kernel of size k, $x_{i:i+k-1}$denotes a segment of the input sequence, and $b$is the bias term. CNN's strength lies in its ability to detect localized sentiment expressions such as "market surge," "loss in profit," or "steady growth." After convolution, a global max pooling layer was applied to extract the most salient features, followed by a dense ReLU layer with 64 neurons, a dropout rate of 0.3, and a softmax classifier to output probabilities for the three sentiment classes.

All models were trained using the Adam optimizer with a learning rate of 0.001, which adaptively updates weights based on first- and second-order moment estimates of gradients. The loss function minimized during training was categorical cross-entropy, defined as:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} log(\hat{y}_{i,c}) N \quad (5)$$

denotes the total number of samples, $C$is the number of sentiment classes, $y_{i,c}$is the true one-hot label, and $\hat{y}_{i,c}$is the predicted probability. Each model was trained for 20 epochs with a batch size of 64, and 20 percent of the training data was used as validation to monitor generalization performance. Dropout regularization was applied to

reduce overfitting, and model convergence was tracked through accuracy and loss curves for both training and validation datasets.

To evaluate model performance, two main metrics were used: accuracy and weighted F1-score. Accuracy measures the proportion of correctly classified instances:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

while the weighted F1-score provides a balanced evaluation across all classes, especially under potential class imbalance:

$$F1_{weighted} = \sum_{i=1}^{C} w_i \cdot \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} P_i \tag{7}$$

and $R_i$denote precision and recall for class i, and $w_i$is the class weight. The confusion matrix was also computed to visualize the distribution of true versus predicted classes, offering detailed insight into model misclassifications between neutral and weakly polarized sentiments.

Finally, visualizations were generated using Matplotlib and Seaborn to support interpretability. Training history plots demonstrated model learning behavior across epochs, bar charts compared overall accuracy and F1-score among the four architectures, and confusion matrix heatmaps provided visual evidence of classification performance for the best-performing model. This combination of quantitative evaluation and graphical analysis allowed for a holistic assessment of each model's strengths and limitations.

In summary, this methodological design ensures fairness, transparency, and reproducibility in evaluating the comparative performance of LSTM, BiLSTM, GRU, and CNN models for financial sentiment classification. The inclusion of mathematical formulations, unified hyperparameter settings, and consistent evaluation procedures establishes a rigorous foundation for empirical benchmarking. The results of this methodology serve as a reliable guide for developing artificial intelligence-driven tools in digital finance, supporting market prediction, and enhancing sentiment-aware decision-making systems.

Algorithm 1 Financial Sentiment Classification

Input: Dataset $D = \{(s_i, y_i)\}_{i=1}^{N}$, containing financial sentences $s_i$and sentiment labels $y_i \in \{0,1,2\}$
Output: Predicted sentiment class $\hat{y}_i$, performance metrics $Accuracy$and $F1_{weighted}$
Process:
Start
Data Preprocessing
Clean text by converting to lowercase and removing special characters.
Tokenize sentences: $s_i \rightarrow X_i = [x_1, x_2, \ldots, x_T]$.
Pad or truncate each sequence to a uniform length $T = 50$.
Transform sentiment labels into one-hot encoded vectors:

$$Y = OneHot(y_i)$$

Split the dataset into training and testing subsets:

$$(X_{train}, Y_{train}), (X_{test}, Y_{test}) = Split(X_p, Y, 0.8)$$

Embedding Representation

Initialize the embedding matrix $W_E \in R^{|V| \times d}$.

Transform token sequences into dense embeddings:

$$E_i = W_E[X_i]$$

This maps each word to a 128-dimensional vector representation.

Model Construction

If model = LSTM:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)\ i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)\ \tilde{C}_t = tanh(W_C[h_{t-1}, x_t] + b_C)\ C_t = f_t * C_{t-1} + i_t * \tilde{C}_t\ h_t = tanh(C_t)$$

If model = BiLSTM:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$$

If model = GRU:

$$z_t = \sigma(W_z[h_{t-1}, x_t])\ r_t = \sigma(W_r[h_{t-1}, x_t])\ \tilde{h}_t = tanh(W_h[r_t * h_{t-1}, x_t])\ h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

If model = CNN:

$$h_i = ReLU(W_c * x_{i:i+k-1} + b)$$

Apply Global Max Pooling to extract the most salient features.

Output Layer and Optimization

Pass the output to a dense layer with ReLU activation:

$$z = ReLU(W_d h_t + b_d)$$

Apply a softmax layer to compute sentiment probabilities:

$$\hat{y}_i = \frac{e^{z_i}}{\sum_{j=1}^{C} e^{z_j}}$$

Define the loss function as categorical cross-entropy:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} log(\hat{y}_{i,c})$$

Update weights using the Adam optimizer:

$$\theta_{t+1} = \theta_t - \eta \frac{m_t}{\sqrt{v_t} + \epsilon}$$

Model Evaluation

Predict sentiment classes:

$$\hat{Y} = arg\,max(\hat{y})$$

Compute model accuracy:

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

Compute weighted F1-score:

$$F1_{weighted} = \sum_{i=1}^{C} w_i \cdot \frac{2 \cdot P_i \cdot R_i}{P_i + R_i}$$

Construct confusion matrix:

$$M_{ij} = |\{x \in X : y = i, \hat{y} = j\}|$$

Visualize accuracy, F1-score, and confusion matrix to assess model performance. Select the best-performing architecture:

$$M^* = arg\,max_M(Accuracy_M)$$

Return M^* and corresponding evaluation results.

End

Results

The results of this study show distinct differences in the performance of the four deep learning models, namely LSTM, BiLSTM, GRU, and CNN, when applied to the task of financial sentiment classification. Each model was trained and tested on the same dataset consisting of 5,842 financial sentences that were manually labeled into three sentiment categories: positive, negative, and neutral. The preprocessing stage involved tokenization, sequence padding, and label encoding to ensure consistent

input across models. The models were trained under identical conditions, using the Adam optimizer with a learning rate of 0.001, categorical cross-entropy as the loss function, and 20 epochs with a batch size of 64. To objectively evaluate their performance, accuracy and weighted F1-score were selected as the primary metrics, as they provide a comprehensive assessment of both prediction correctness and balance across multiple sentiment classes. This approach ensures that models are not only rewarded for correctly classifying dominant classes but are also evaluated for their ability to identify minority sentiment categories effectively.

The results indicate that there are clear variations in how well each model captures sentiment from financial text data. The CNN model achieved the highest overall performance, demonstrating superior ability to extract sentiment-bearing features from localized text patterns. This can be attributed to the convolutional layer's capability to identify important n-gram structures, such as expressions of financial growth or decline, without relying heavily on long-term dependencies. The BiLSTM model also performed competitively, benefiting from its bidirectional processing that allows it to interpret contextual information from both preceding and succeeding words in a sentence. However, both LSTM and GRU models showed limited success in this task, achieving relatively low accuracy and F1-scores, suggesting that their reliance on sequential memory mechanisms provides less advantage when dealing with short, information-dense financial statements. Overall, the comparative performance across models underscores the importance of model architecture selection, where CNN and BiLSTM demonstrate stronger generalization and better adaptability to the linguistic and contextual complexity of financial language.

Table 1 Performance Comparison of Deep Learning Models

Model	Accuracy	F1-score
LSTM	0.5321	0.3696
BiLSTM	0.6749	0.6686
GRU	0.5321	0.3696
CNN	0.7032	0.7022

As presented in table 1, the CNN model produced the highest overall performance among all evaluated architectures, with an accuracy of 70.32 percent and a weighted F1-score of 70.22 percent. This indicates that the CNN was able to effectively capture the discriminative features that define sentiment in financial language. The model's convolutional filters played a crucial role in identifying local patterns, such as key phrases and short expressions that carry strong sentiment cues. These features often appear as concise textual indicators of financial status, for example, "rise in revenue," "drop in earnings," or "strong market performance." The CNN's ability to process these short n-gram structures

allowed it to generalize efficiently without requiring long-term contextual information. In contrast, the BiLSTM model achieved slightly lower performance, with an accuracy of 67.49 percent and an F1-score of 66.86 percent. The BiLSTM's bidirectional mechanism enabled it to analyze each sentence from both forward and backward directions, capturing dependencies between preceding and succeeding words that contributed to a richer contextual understanding. However, this contextual strength came at a higher computational cost and slightly reduced efficiency compared to CNN, which appears to be better suited for short and information-dense financial statements.

The LSTM and GRU models demonstrated relatively poor performance compared to CNN and BiLSTM, each recording 53.21 percent accuracy and 36.96 percent F1-score. Their limited ability to capture sentiment nuances in financial text suggests that sequential memory mechanisms alone are insufficient for modeling sentiment when linguistic structures are compact and context shifts occur rapidly. These models are designed to retain long-term dependencies, but in financial discourse, where sentences often convey condensed and factual information, such mechanisms contribute less to improving classification accuracy. The comparative visualization in figure 2 further reinforces these findings by clearly illustrating the gap between CNN and the other architectures. The bar chart shows that CNN consistently outperformed the remaining models across both accuracy and F1-score metrics, followed closely by BiLSTM. The figure visually emphasizes that convolution-based architectures can better extract essential features from financial data, while traditional recurrent structures like LSTM and GRU tend to lose efficiency when dealing with short, domain-specific text inputs.

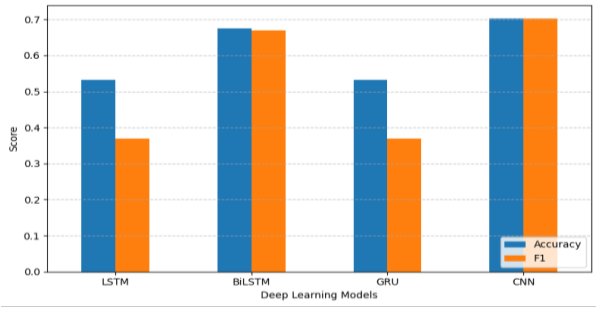

Figure 2 Model Performance Comparison Based on Accuracy and F1-score

Figure 2 illustrates that the CNN model demonstrates remarkable stability during training and strong generalization when tested on unseen financial text data. The model consistently maintains high performance across multiple epochs, indicating that it successfully learned relevant sentiment patterns without overfitting. The convolutional layers in CNN enable the extraction of local semantic features, allowing the model to focus on specific sequences of words that convey emotional or evaluative meaning in financial statements. This capacity makes it highly effective at

recognizing sentiment-bearing phrases such as “increase in revenue,” “market decline,” “loss in profit,” or “strong quarterly growth,” which often appear as compact expressions with high predictive value. These localized features are particularly important in financial language, where sentiment is frequently expressed through short and precise terminology rather than long descriptive narratives. In comparison, the BiLSTM model also performed well, benefiting from its ability to analyze textual sequences in both forward and backward directions. This bidirectional context enables BiLSTM to capture dependencies between words that occur earlier and later in a sentence, providing a deeper understanding of linguistic nuances. However, its slightly lower accuracy compared with CNN suggests that in financial sentiment analysis, identifying local key phrases is often more critical than modeling broader contextual relationships, especially when dealing with concise sentences found in reports and market summaries.

To gain a more detailed understanding of the CNN model’s classification behavior, a confusion matrix was generated and is presented in figure 3. The matrix offers a clear view of how well the CNN categorized sentences into positive, negative, and neutral classes. The results show that the model correctly classified the majority of neutral statements, reflecting its effectiveness in identifying objective or factual expressions commonly used in financial communication.

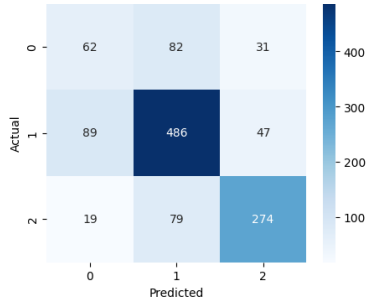

Figure 3 Confusion Matrix of the CNN Model for Financial Sentiment Classification

However, several instances of confusion occurred between positive and neutral categories, where sentences expressing mild optimism, such as “steady improvement” or “marginal growth,” were often labeled as neutral. A smaller number of errors were also observed between negative and neutral statements, where phrases indicating minor declines were misinterpreted as neutral due to their subtle tone. Despite these misclassifications, the confusion matrix reveals that CNN maintained a balanced distribution of correct predictions across all sentiment classes. This outcome confirms that the CNN model not only achieves high quantitative performance but also demonstrates robustness in understanding the linguistic complexity and ambiguity typical of financial text.

As presented in figure 3, the CNN model successfully identified most of the neutral samples, indicating its strong ability to detect factual and objective statements that frequently occur in financial reports and market summaries. This result highlights the model's capability to differentiate between purely informative content and sentiment-bearing expressions, a crucial aspect of financial text analysis where neutrality often dominates. Nevertheless, a number of misclassifications were observed between the positive and neutral categories. Several positive sentences, which contained expressions of mild optimism such as "steady improvement," "slight profit growth," or "stable performance," were predicted as neutral. This confusion is understandable, given that financial language often uses cautious or understated phrasing that can blur the line between neutrality and positivity. Similarly, a few neutral statements were misinterpreted as positive due to the presence of words that carry slight positive connotations, even when used in a factual context. The model also displayed minor difficulties in distinguishing negative sentences that described limited or moderate declines from neutral statements, particularly in cases where the phrasing lacked strong emotional or evaluative markers. These misclassifications demonstrate the inherent linguistic complexity of financial communication, where tone and sentiment are often implicit and context-dependent rather than explicitly stated.

Despite these minor inaccuracies, the CNN model maintained consistent and balanced classification performance across all sentiment categories. Its strong generalization ability suggests that it effectively learned to adapt to the subtle variations and domain-specific vocabulary characteristic of financial discourse. The quantitative outcomes, combined with the visual evidence from the confusion matrix, confirm that CNN achieved the highest predictive accuracy and F1-score among all evaluated models. This reflects the network's capability to extract key sentiment-bearing features from short and information-rich statements without overfitting to specific terms or sentence structures. The BiLSTM model also produced competitive results, particularly in handling context-dependent expressions, making it a suitable alternative when more nuanced understanding of text is required. However, the LSTM and GRU models exhibited relatively weak performance, indicating their limited capacity to process concise and fact-heavy financial language effectively. Overall, these findings demonstrate that CNN provides the most reliable and computationally efficient framework for financial sentiment analysis, while BiLSTM offers added value in scenarios that demand a higher level of contextual interpretation.

Overall, these results establish CNN as the most effective and reliable deep learning architecture for financial sentiment classification within digital market contexts. Its ability to extract key sentiment-bearing features with high precision and computational efficiency makes it a

promising tool for practical implementations in digital finance, automated trading sentiment systems, and market intelligence applications.

Discussion

The results demonstrate that the four deep learning architectures behave differently in financial sentiment classification. CNN achieved the best overall performance, indicating its effectiveness in extracting local patterns and short semantic sequences from concise financial text. Expressions such as “profit increase,” “revenue decline,” and “market loss” can strongly determine sentiment polarity, making CNN’s local feature extraction particularly suitable for this domain [20], [21]. BiLSTM also achieved strong performance by capturing contextual information from both preceding and succeeding words, which is useful for comparative and conditional financial expressions. However, its higher training time and computational cost make it less efficient than CNN for large-scale or real-time applications [22], [23].

LSTM and GRU produced lower performance, suggesting that sequential memory mechanisms may provide less advantage for short and information-dense financial texts. Financial statements typically emphasize concise facts, metrics, and expressions rather than long narrative structures [24]. The results also show that distinguishing neutral from mildly polarized sentiment remains challenging, as even CNN occasionally classified slightly positive or negative statements as neutral. This difficulty reflects the restrained and ambiguous nature of financial language [25]. Future studies could address this limitation by incorporating domain-specific embeddings, attention mechanisms, or transformer-based models such as FinBERT, which can provide richer contextual representations. Overall, the findings indicate that CNN offers an effective balance between classification performance and computational efficiency, while BiLSTM remains useful when deeper contextual understanding is required.

Conclusion

This study compared four deep learning architectures, LSTM, BiLSTM, GRU, and CNN, for financial sentiment classification using 5,842 labeled financial sentences. Using accuracy and weighted F1-score, CNN achieved the highest performance, demonstrating its effectiveness in extracting local sentiment features from short and information-dense financial texts. BiLSTM achieved the second strongest performance by capturing bidirectional contextual dependencies, although it required greater computational resources than CNN. Meanwhile, LSTM and GRU performed less effectively, suggesting that long-term sequential modeling provides less advantage for concise financial texts.

The findings indicate that model selection should consider the characteristics of financial data and analytical objectives. CNN provides

an efficient approach for concise sentiment-rich texts, while BiLSTM may be more suitable for longer and context-dependent financial documents. Distinguishing neutral from weakly positive or negative sentiments remains challenging due to the subtle and cautious nature of financial language. Future research should therefore explore transformer-based models such as FinBERT and RoBERTa to capture deeper financial context and improve sentiment classification. These findings contribute to the development of AI-based financial sentiment analysis for digital finance, investment analysis, and real-time market decision support.

Declarations

Author Contributions

Conceptualization: L.D., M.S.; Methodology: L.D.; Software: M.S.; Validation: L.D.; Formal Analysis: L.D.; Investigation: M.S.; Resources: L.D., M.S.; Data Curation: M.S.; Writing – Original Draft Preparation: L.D.; Writing – Review and Editing: L.D., M.S.; Visualization: M.S.; 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] S. Kumar and P. Bhatia, “LSTM Based Sentiment Analysis of Financial News,” *SN Computer Science*, vol. 4, no. 5, pp. 584–594, 2023, doi: 10.1007/s42979-023-02018-2.

[2] M. García, L. Serrano, and F. Herrera, “Financial Sentiment Analysis: Classic Methods vs. Deep Learning Models,” *Information*, vol. 14, no. 5, pp. 451–463, 2023, doi: 10.3233/IDT-230478.

[3] F. Tao, W. Wang, and R. Lu, “A deep learning-based sentiment flow analysis model for predicting financial risk of listed companies,” *Engineering Applications*

of Artificial Intelligence, vol. 150, no. June, Art. no. 110522, 2025, doi: 10.1016/j.engappai.2025.110522.

[4] R. Patel and M. Banerjee, "Deep Learning for Financial Forecasting: A Review of Recent Trends," *International Review of Economics & Finance*, vol. 104, no. November, pp. 104719–104734, 2025, doi: 10.1016/j.iref.2025.104719.

[5] H. Zhu, X. Yang, and S. Liu, "Fintech Sentiment Analysis Using Deep Learning Models," in *Proc. International Congress on ICT (ICICT)*, *Lecture Notes in Networks and Systems (LNNS)*, vol. 1413, no. October, pp. 325–333, 2025, doi: 10.1007/978-981-96-6435-1_25.

[6] D. K. Nasiopoulos, K. I. Roumeliotis, D. P. Sakas, K. Toudas, and P. Reklitis, "Financial sentiment analysis and classification: A comparative study of fine-tuned deep learning models," *International Journal of Financial Studies*, vol. 13, no. 2, Art. no. 75, 2025, doi: 10.3390/ijfs13020075.

[7] J. Delgadillo, J. Kinyua, and C. Mutigwe, "FinSoSent: Advancing financial market sentiment analysis through pretrained large language models," *Big Data and Cognitive Computing*, vol. 8, no. 8, Art. no. 87, 2024, doi: 10.3390/bdcc8080087.

[8] T. Nguyen, H. Vu, and Q. Tran, "Integrative Analysis of Financial Market Sentiment Using CNN and GRU for Risk Prediction and Alert Systems," *arXiv preprint arXiv:2412.10199*, vol. 2024, no. December, pp. 1-6, Dec. 2024. Available: https://arxiv.org/abs/2412.10199.

[9] P. Wang and X. Li, "A Deep Learning Framework Integrating CNN and BiLSTM for Financial Systemic Risk Analysis," *arXiv preprint arXiv:2502.06847*, vol. 2025, no. February, pp. 1-5, Feb. 2025. Available: https://arxiv.org/abs/2502.06847.

[10] Y. Yang, E. Liang, and X. Zhang, "FinBERT: Financial Sentiment Analysis with Pre-trained Language Models," *arXiv preprint arXiv:1908.10063*, vol. 2019, no. August, pp. 1-11, Aug. 2019. Available: https://arxiv.org/abs/1908.10063.

[11] S. Singh, P. Verma, and R. Gupta, "SEntFiN 1.0: Entity-Aware Sentiment Analysis for Financial News," *arXiv preprint arXiv:2305.12257*, vol. 2023, no. May, pp. 1-32, May 2023. Available: https://arxiv.org/abs/2305.12257.

[12] J. Brown, L. Shen, and A. Rahman, "FinEAS: Financial Embedding Analysis of Sentiment," *arXiv preprint arXiv:2111.00526*, vol. 2021, no. November, pp. 1-14, Nov. 2021. Available: https://arxiv.org/abs/2111.00526.

[13] A. Moore and J. P. Qiu, "Lancaster A at SemEval-2017 Task 5: Evaluation Metrics Matter—Predicting Sentiment from Financial News Headlines," *arXiv preprint arXiv:1705.00571*, vol. 2017, no. May, pp. 1-5, May 2017. Available: https://arxiv.org/abs/1705.00571.

[14] M. Chen, Y. Li, and H. Zhao, "Comparative Study for Sentiment Analysis of Financial Tweets with Deep Learning Methods," *Applied Sciences*, vol. 14, no. 2, pp. 1–16, 2024, doi: 10.3390/app14020588.

[15] Das, N., Sadhukhan, B., Ghosh, R. *et al.* "Developing Hybrid Deep Learning Models for Stock Price Prediction Using Enhanced Twitter Sentiment Score and Technical Indicators", *Comput Econ* vol. **64**, no. March, pp. 3407–3446, (2024). https://doi.org/10.1007/s10614-024-10566-9

[16] J. R. Jim, M. A. R. Talukder, P. Malakar, M. M. Kabir, K. Nur, and M. F. Mridha, "Recent advancements and challenges of NLP-based sentiment analysis: A state-

of-the-art review," *Natural Language Processing Journal*, vol. 6, no. March, Art. no. 100059, pp. 1-30, 2024, doi: 10.1016/j.nlp.2024.100059.

[17] Gil, O., Martinez, E., Mansilla-Lopez, J., Loza, R. (2027). Stock Price Forecasting Integrating Deep Learning Models and Sentiment Analysis. In: Rocha, A., Ferrás, C., Loo, L. (eds) Information Technology & Systems. ICITS 2026. Lecture Notes in Networks and Systems, vol 1965, no. August, pp. 420-434, Springer, Cham. https://doi.org/10.1007/978-3-032-25652-2_35

[18] R. Khalil, "AI driven sentiment analysis in financial markets: Using transformer base models and social media signals for stock market predictions," *Journal of Modelling in Management*, vol. 2026, no. January, pp. 1-25, 2026, doi: 10.1108/JM2-08-2025-0415.

[19] S. De Silva, M. Jayasena, and N. Perera, "Quantitative Financial Modeling for Sri Lankan Markets: Sentiment and Time-Series Forecasting," arXiv preprint arXiv:2512.20216, vol. 2025, no. December, pp. 1-8, Dec. 2025. Available: https://arxiv.org/abs/2512.20216.

[20] G. L. Duan, S. Yan, and M. Zhang, "A hybrid neural network model for sentiment analysis of financial texts using topic extraction, pre-trained model, and enhanced attention mechanism methods," *IEEE Access*, vol. 12, no. July, pp. 98207-98224, 2024, doi: 10.1109/ACCESS.2024.3429150.

[21] T. Mawere, S. Rajalakshmi, V. Kuthadi, and O. Dinekanyane, "Adaptive sentiment analysis for stock markets using deep learning," *International Journal of Advances in Applied Sciences*, vol. 15, no. 1, pp. 416-426, 2026. DOI: 10.11591/ijaas.v15.i1.pp416-426

[22] G. W. Chen and I. C. Hsu, "Integrating Taiwan financial BERT sentiment analysis with CNN-BiLSTM-SA model for stock prediction," *Discover Computing*, vol. 28, no. November, pp. 1-21, 2025, doi: 10.1007/s10791-025-09515-3.

[23] B. H. C. and I. Jacob, "An efficient hybrid model for stock market price prediction using CNN-BiLSTM with attention mechanism and sentimental analysis," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 25565-25571, 2025, doi: 10.48084/etasr.11886.

[24] K. Du, F. Xing, R. Mao, and E. Cambria, "Financial sentiment analysis: Techniques and applications," *ACM Computing Surveys*, vol. 56, no. 9, pp. 1-42, 2024, doi: 10.1145/3649451.

[25] K. Mishev, A. Gjorgjevikj, I. Vodenska, L. T. Chitkushev, and D. Trajanov, "Evaluation of sentiment analysis in finance: From lexicons to transformers," *IEEE Access*, vol. 8, no. July, pp. 131662-131682, 2020, doi: 10.1109/ACCESS.2020.3009626.