Deep Learning-Based Sentiment Analysis for Digital Market Intelligence through Comparative Evaluation of BiLSTM and CNN Models

Hendro Budiyanto[1,*], Muhammad Shahid Khan[2], Winda Widyanty[3]

[1]Faculty of Business, Universitas Multimedia Nusantara, Indonesia
[2]Faculty of Business, Suan Sunandha Rajabhat University, Thailand
[3]Faculty of Business and Management, Universitas Mercu Buana, Indonesia

ABSTRACT

The rapid growth of digital communication platforms has generated vast amounts of textual data containing valuable insights into public sentiment toward products, services, and digital assets. Accurately analyzing this data is crucial for understanding consumer behavior and market trends in digital marketing and cryptocurrency ecosystems. This study presents a comparative analysis of two deep learning architectures, Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN), for sentiment classification using textual data. Both models were trained for twenty epochs without early stopping to evaluate their full learning potential. The experimental results revealed that the BiLSTM model achieved superior performance with an accuracy of 73.82% and an F1-score of 72.40%, while the CNN model obtained 65.10% accuracy and an F1-score of 64.28%. The BiLSTM demonstrated a higher capability to capture sequential dependencies and contextual semantics through its bidirectional processing, whereas the CNN primarily relied on local feature extraction, limiting its contextual understanding. Class-wise analysis showed that BiLSTM performed strongly in identifying neutral and positive sentiments but struggled with negative sentiment detection due to data imbalance and linguistic ambiguity. These findings highlight BiLSTM's robustness and suitability for sentiment analysis in digital market applications. The study emphasizes the importance of context-aware deep learning models in supporting data-driven marketing strategies, customer sentiment monitoring, and digital asset analysis. Future research is recommended to integrate transformer-based architectures such as BERT or RoBERTa to further enhance contextual comprehension and improve overall classification performance.

Keywords Sentiment Analysis, BiLSTM, CNN, Digital Marketing, Artificial Intelligence

Submitted 15 February 2026
Accepted 20 April 2026
Published 27 August 2026

Corresponding author
Hendro Budiyanto,
hendro.budiyanto@umn.ac.id

Additional Information and Declarations can be found on page 192

DOI: 10.47738/jdmdc.v3i3.72

INTRODUCTION

The rapid advancement of digital technologies and online communication platforms has fundamentally changed the way individuals express opinions, interact with brands, and make purchasing or investment decisions. In the digital economy, vast amounts of textual data are continuously generated through social media platforms, e-commerce reviews, online forums, and cryptocurrency communities. These digital spaces serve as rich sources of information that reflect consumer attitudes, preferences, and emotions toward various products and services. For businesses and organizations operating in digital markets,

How to cite this article: H. Budiyanto, M. S. Khan, W. Widyanty, "Deep Learning-Based Sentiment Analysis for Digital Market Intelligence through Comparative Evaluation of BiLSTM and CNN Models," *J. Digit. Mark. Digit. Curr.*, vol. 3, no. 3, pp. 175-194, 2026.

understanding public sentiment has become a crucial aspect of decision-making, marketing strategy formulation, and brand management. Accurate sentiment analysis enables companies to identify emerging consumer trends, gauge customer satisfaction, and predict market movements with greater precision [1]. However, the unstructured and linguistically diverse nature of textual data presents significant challenges for automated analysis, particularly when sentiments are expressed indirectly or ambiguously.

In recent years, Artificial Intelligence (AI) and Deep Learning (DL) have become central to solving complex problems in text analytics and natural language processing [2]. Traditional machine learning techniques, such as Support Vector Machines (SVM), Naïve Bayes, and Logistic Regression, have been widely used in early sentiment analysis applications. Although these models can achieve reasonable performance when combined with handcrafted features like term frequency and n-gram statistics, they often fail to capture the contextual and sequential relationships between words in a sentence. Deep learning methods, in contrast, automatically learn feature representations from data without manual intervention, enabling more accurate understanding of linguistic patterns [3]. Among these, Convolutional Neural Networks (CNN) have gained attention for their ability to identify local patterns and key phrases, while Recurrent Neural Networks (RNN) and their variants, including Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), have proven highly effective in modeling long-term dependencies and contextual semantics in textual data. This evolution in neural network design represents the current state of the art in sentiment analysis, where deep learning architectures outperform traditional models in both efficiency and predictive capability [4].

Despite these technological advancements, several gaps remain in the current body of research related to sentiment analysis within digital markets and digital currency ecosystems. Many existing studies have focused on general-purpose sentiment datasets, such as movie reviews or social media posts, rather than domain-specific data that reflects the linguistic patterns and emotional tone of digital market communication. As a result, there is limited understanding of how well deep learning models perform when applied to data drawn from financial discussions, cryptocurrency forums, or online reviews in the context of digital commerce. Moreover, most previous comparative studies between CNN and BiLSTM have utilized early stopping or short training durations, potentially limiting each model's opportunity to reach full learning capacity [5]. Consequently, there is insufficient empirical evidence illustrating how the two architectures perform over extended training periods and whether the difference in their sequential learning structures meaningfully influences model generalization and sentiment classification accuracy. These issues highlight the need for a deeper and

more comprehensive comparison between CNN and BiLSTM under consistent experimental conditions.

Another area that remains underexplored is the interpretability and adaptability of deep learning models for sentiment analysis across varying sentiment categories. Many models demonstrate high performance on dominant or neutral sentiment classes but struggle to accurately detect negative sentiments, which are often linguistically subtle and context-dependent. Negative opinions can be expressed through indirect phrases, sarcasm, or mixed emotions that are easily misinterpreted by models focused solely on local lexical patterns. Additionally, data imbalance, where positive or neutral classes significantly outnumber negative examples, can further reduce classification performance for underrepresented categories. Addressing these challenges requires models capable of not only identifying explicit sentiment cues but also capturing implicit emotional undertones embedded within context. Therefore, evaluating BiLSTM and CNN in this setting provides valuable insights into their relative strengths and weaknesses in managing linguistic ambiguity, class imbalance, and contextual variability within sentiment data.

The primary objective of this research is to conduct a detailed comparative analysis between the BiLSTM and CNN architectures for sentiment classification in textual data relevant to digital market applications. Both models were trained for twenty full epochs without the use of early stopping, allowing observation of their complete learning behavior and convergence patterns. Model performance was evaluated using multiple metrics, including accuracy, F1-score, precision, and recall, with further analysis conducted using class-wise performance and confusion matrices. By systematically comparing the two architectures, this study aims to identify which model is better suited for capturing sentiment patterns in digital communication and consumer-generated text. The findings contribute to a broader understanding of how deep learning can enhance sentiment-driven analytics in digital marketing and digital currency domains. Ultimately, the results of this research are expected to assist organizations, marketers, and financial analysts in implementing AI-based systems that can more accurately interpret market sentiment and support data-driven strategic decisions.

Literature Review

Sentiment analysis has become an essential research area in natural language processing (NLP) due to the exponential growth of online textual content across social media, e-commerce, and financial platforms. The objective of sentiment analysis is to extract opinions, attitudes, and emotions expressed by users in written text, which is valuable for understanding consumer behavior and market dynamics [6]. Early approaches relied heavily on traditional machine learning

techniques such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression [7]. These models required manual feature engineering using representations such as bag-of-words or term frequency–inverse document frequency (TF-IDF) [8]. While effective to some extent, these approaches often failed to capture complex linguistic structures, semantic meaning, and contextual relationships between words, leading to limitations in handling longer and more nuanced text [9].

The introduction of deep learning brought significant advancements in sentiment analysis by allowing models to automatically learn high-level features from raw text data [10]. Convolutional Neural Networks (CNN) were among the earliest architectures adopted for text classification tasks, excelling at identifying localized features and phrase-level sentiment indicators [11]. CNN-based sentiment models utilize convolutional and pooling layers to detect patterns within fixed-length word embeddings, enabling efficient training and strong generalization on short text sequences [12]. However, CNN models tend to struggle with capturing long-term dependencies and contextual meaning since their convolutional operations are inherently limited to fixed receptive fields [13]. Consequently, CNNs are less effective in processing complex sentences where sentiment depends on relationships between distant words or clauses.

To address these limitations, sequential neural network models such as the Recurrent Neural Network (RNN) and its enhanced variant, the Long Short-Term Memory (LSTM), were developed to capture sequential dependencies within text [14]. The Bidirectional LSTM (BiLSTM) further extended this capability by processing text in both forward and backward directions, enabling the model to learn context from surrounding words [15]. This bidirectional architecture has been shown to significantly improve sentiment classification accuracy by understanding the full context of a sentence rather than relying solely on preceding information. Comparative studies have demonstrated that BiLSTM consistently outperforms CNN models on various sentiment analysis datasets, particularly when dealing with longer or more context-dependent expressions [16]. Moreover, BiLSTM networks are better equipped to handle linguistic constructs such as negations, intensifiers, and compound statements, which are critical for distinguishing subtle sentiment variations in digital communication.

Recent research has also explored hybrid and transformer-based approaches that integrate multiple neural architectures to improve overall model performance. Hybrid CNN–BiLSTM frameworks combine the local feature extraction capability of CNN with the contextual learning ability of BiLSTM, resulting in enhanced sentiment prediction accuracy [17]. Attention mechanisms and transformer-based models such as BERT and RoBERTa have further advanced the field by leveraging pre-trained

contextual embeddings that capture deeper semantic representations [18]. These approaches have achieved state-of-the-art results across multiple benchmark datasets, demonstrating the importance of contextual understanding in modern sentiment analysis. Nevertheless, despite their success, transformer-based models require substantial computational resources and large amounts of labeled data, which limit their applicability in certain domains, particularly for small and domain-specific datasets.

Despite the progress in deep learning for sentiment analysis, significant research gaps remain, especially in domain-specific applications such as digital marketing and digital currency analytics. Many previous studies focused on general-purpose text corpora and did not examine sentiment expressed in financial or digital market contexts, where language patterns often differ significantly from everyday communication [19]. Moreover, most comparative analyses between CNN and BiLSTM employed early stopping or short training periods, which may not accurately reflect each model's full learning potential. There is also limited research that directly examines model stability, generalization performance, and class-wise prediction behavior in the context of multi-class sentiment classification. Addressing these gaps is essential to improve the reliability and interpretability of sentiment models for decision-making applications in digital business environments [20]. Therefore, this study aims to contribute to this gap by performing a comprehensive evaluation of CNN and BiLSTM models under identical configurations, using extended training to assess their long-term learning behavior and effectiveness in classifying sentiments related to digital market data.

Methodology

This study employed a quantitative experimental approach to evaluate and compare the performance of two deep learning architectures Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN) for the task of sentiment classification. The methodological framework was designed to ensure a controlled and objective comparison, where both models were trained and evaluated under identical preprocessing, embedding, and optimization settings. The overall methodological process, illustrated in figure 1, consists of several sequential stages: data collection, preprocessing, embedding initialization, model construction, model training, and performance evaluation. Each stage was structured to systematically transform raw text data into a form suitable for deep learning, followed by the learning and assessment of the models' capability to capture linguistic and contextual sentiment features. Figure 1 provides an overview of these stages, depicting the flow of the research from input data through preprocessing and embedding, continuing to the model development and

training phases, and culminating in the evaluation and comparative analysis of the two architectures.

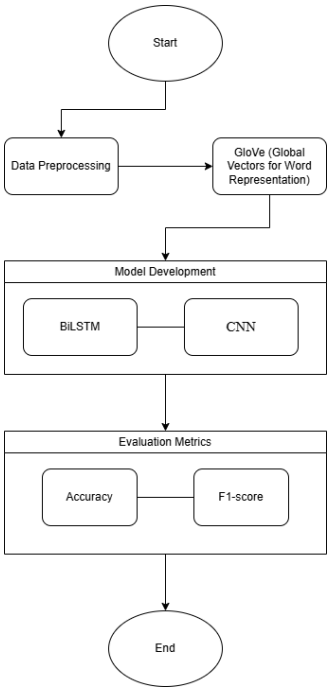

Figure 1 Research Steps outlining the methodological framework of the study.

The dataset used in this study was organized in a CSV file named data.csv, consisting of two key attributes: Sentence and Sentiment. The Sentence attribute contained textual data derived from user-generated opinions, while the Sentiment attribute represented the sentiment label categorized into three classes—positive, neutral, and negative. These labels were designed to capture the emotional polarity of the text, allowing for multi-class sentiment classification. The dataset was divided into training and testing subsets using an 80:20 ratio, ensuring that the training set provided sufficient diversity for model learning while the testing set contained unseen data for validation. A stratified sampling approach was used to preserve the class distribution across both subsets, reducing potential bias caused by uneven sentiment representation. This dataset preparation step was critical for maintaining consistency in model training and reliable performance comparison.

Prior to the training process, extensive preprocessing was applied to clean and transform the raw text data into a numerical format compatible with deep learning architectures. All sentences were converted to lowercase to maintain consistency, and a Keras Tokenizer was employed to generate word indices based on frequency, with a maximum vocabulary size of 15,000 words. Words not included in the vocabulary were replaced with a special token (“<OOV>”) to manage unseen words. Each sentence was tokenized and converted into a sequence of integers

representing word indices. Since deep learning models require uniform input dimensions, all sequences were standardized using post-padding to a fixed maximum length of 80 tokens. The sentiment labels were then encoded numerically using the LabelEncoder and further transformed into one-hot encoded vectors via the to_categorical() function from Keras. These preprocessing steps ensured that the dataset was clean, consistent, and numerically structured for input into the BiLSTM and CNN models.

To represent semantic meaning more effectively, the study employed the GloVe (Global Vectors for Word Representation) embedding with 100-dimensional pre-trained word vectors. GloVe embeddings capture global word co-occurrence statistics from large-scale text corpora, enabling the model to understand semantic relationships between words. The embedding process constructs a word embedding matrix that maps each word index to its corresponding vector representation. This is achieved by minimizing the following objective function:

$$J = \sum_{i,j=1}^{V} f(P_{ij})\left(w_i^{\top}\tilde{w}_j + b_i + \tilde{b}_j - log\, P_{ij}\right)^2 \quad (1)$$

P_{ij}represents the probability of word $j$appearing in the context of word i, $w_i$and $\tilde{w}_j$are the word and context vectors, and $f(P_{ij})$is a weighting function that controls the impact of frequent co-occurrences. The resulting embeddings were stored in an embedding matrix and used as the initial weights of the models' embedding layers. The embedding layer was set as non-trainable to retain the semantic information encoded in the GloVe vectors, allowing both architectures to benefit from rich linguistic knowledge during training.

The BiLSTM model architecture was designed to capture long-term dependencies and bidirectional contextual information. The architecture consisted of an embedding layer initialized with the GloVe matrix, followed by two stacked bidirectional LSTM layers with 128 and 64 hidden units, respectively. Dropout and recurrent dropout values of 0.3 were applied to prevent overfitting, and batch normalization was incorporated to stabilize the training process. The output was passed through a fully connected dense layer with 64 neurons using the ReLU activation function, followed by a dropout layer (rate 0.4) to enhance generalization. Finally, a softmax output layer with three neurons was used to classify the text into the three sentiment categories. The BiLSTM processes sequences from both forward and backward directions, enabling it to understand contextual dependencies across sentences, making it particularly effective for sentiment analysis.

Meanwhile, the CNN model was constructed to identify localized linguistic patterns and n-gram features from the input text. The architecture included an embedding layer with GloVe embeddings,

followed by two one-dimensional convolutional layers with 128 and 64 filters and kernel sizes of 5 and 3, respectively. Batch normalization was used after each convolutional layer to enhance training stability. The resulting feature maps were passed through a global max pooling layer, which selected the most salient features. The pooled output was fed into a dense layer with 64 neurons and a dropout layer (rate 0.4). The final output layer also used softmax activation for multi-class sentiment classification. While CNN excels at extracting localized sentiment patterns efficiently, it lacks the ability to model long-term dependencies, distinguishing it conceptually from the BiLSTM model.

Both models were compiled using the Adam optimizer with a learning rate of 0.001 and trained using the categorical cross-entropy loss function:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} log(\hat{y}_{i,c}) \quad (2)$$

$N$is the total number of samples, $C$is the number of sentiment classes, $y_{i,c}$represents the true label, and $\hat{y}_{i,c}$denotes the predicted probability for each class. Training was performed for 20 epochs with a batch size of 64 and a validation split of 20%. The ReduceLROnPlateau callback was implemented to adaptively reduce the learning rate by a factor of 0.5 whenever the validation loss did not improve for three consecutive epochs. Both models were trained without early stopping to allow full observation of their learning and convergence behaviors.

The evaluation of model performance employed standard classification metrics, including accuracy, precision, recall, and F1-score. Accuracy (Acc) measured the overall proportion of correctly classified instances and was defined as:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (3)$$

TP, TN, FP, and FNrepresent true positives, true negatives, false positives, and false negatives, respectively. The F1-score ($F1$) provided a harmonic mean between precision and recall, offering a balanced measure of accuracy and robustness, expressed as:

$$F1 = 2 \times \frac{P \times R}{P + R} \quad (4)$$

A confusion matrix was also analyzed to visualize the classification results, indicating how well each model performed for each sentiment class. These evaluation metrics enabled a comprehensive comparison between the BiLSTM and CNN architectures, identifying their respective strengths and weaknesses in modeling textual sentiment.

In summary, this methodological framework integrated structured data preprocessing, semantically informed embedding, and rigorous training

and evaluation protocols. By maintaining consistency across preprocessing, embedding, and optimization configurations, this study ensured that the observed performance differences between BiLSTM and CNN were solely due to architectural distinctions. The overall experimental design thus provided a replicable and objective foundation for understanding how sequential context modeling and localized feature extraction influence the accuracy and generalization of sentiment analysis in digital market applications.

Algorithm 1 Sentiment Classification using BiLSTM and CNN

Input: Raw dataset $D = \{(s_i, y_i)\}$, pretrained embeddings E_{GloVe}
Output: Trained models M_{BiLSTM}, M_{CNN}, performance metrics (Accuracy, F1-score)
Process:
Start
Load dataset $D = \{(s_i, y_i)\}_{i=1}^{N}$ and perform preprocessing:
convert to lowercase, tokenize $T(s_i')$, pad to fixed length $L = 80$,
and one-hot encode sentiment labels y_i'.
Initialize GloVe embeddings:

$$W_e[i] = \{E_{GloVe}(w_i), if\ w_i \in E_{GloVe}\ Uniform(-a, a), otherwise$$

Construct embedding layer $E(x_i) = W_e \cdot x_i$.
BiLSTM Architecture:

$$\overrightarrow{h_t} = LSTM_f(E(x_t), \overrightarrow{h_{t-1}})\ h_t^{\leftarrow} = LSTM_b(E(x_t), h_{t+1}^{\leftarrow})\ h_t = [\overrightarrow{h_t}; h_t^{\leftarrow}]\ \hat{y} = Softmax(W_2\ ReLU(W_1 h_t + b_1) + b_2)$$

CNN Architecture:

$$c_i = \sigma(W_c * E(x_i) + b_c)\ p = max(c_i)\ \hat{y} = Softmax(W_o\ ReLU(W_d p + b_d) + b_o)$$

Train both models using categorical cross-entropy loss:

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

Update parameters via Adam optimizer:

$$\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$$

Set learning rate $\eta = 0.001$, batch size = 64, epochs = 20,
and validation split = 0.2 with dynamic learning rate reduction.
Evaluate model performance:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}\ F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Select the best-performing model:

$$M^* = arg\ max_{M \in \{M_{BiLSTM}, M_{CNN}\}}(Acc_M + F1_M)$$

End

Results

This section presents the experimental results obtained from the comparative analysis between two deep learning architectures: Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN). Both models were trained using identical configurations for twenty full epochs without early stopping to ensure fair comparison and to observe their complete learning dynamics. The

experiments aimed to evaluate how well each model performs in classifying sentiment into three categories: positive, neutral, and negative.

The overall performance comparison between the two deep learning architectures is summarized in table 1. The BiLSTM model demonstrated significantly higher performance with an accuracy of 73.82% and an F1-score of 72.40%, outperforming the CNN model, which achieved 65.10% accuracy and an F1-score of 64.28%. This clear performance gap indicates that BiLSTM has a stronger capacity to capture the intricate dependencies and linguistic structures present in textual data. Because BiLSTM processes input sequences in both forward and backward directions, it is able to retain information from earlier and later parts of a sentence simultaneously. This bidirectional approach enhances the model's ability to understand the context surrounding each word, which is essential in sentiment classification tasks where meaning often depends on the order and relationship between words. For example, phrases such as "not good" or "hardly positive" require contextual interpretation that extends beyond adjacent word patterns, which BiLSTM handles effectively.

In contrast, the CNN model relies primarily on convolutional filters to extract local patterns or short n-gram features within text sequences. While this makes CNN computationally efficient and effective in detecting specific keyword patterns, it limits its ability to grasp long-range dependencies and sentence-level semantics. As a result, CNN often struggles to interpret complex linguistic constructs, negations, and subtle emotional cues that are common in human expressions of sentiment. Moreover, CNN tends to generalize poorly when dealing with variable sentence lengths or when important contextual information lies outside the receptive field of its convolutional filters. This explains why the CNN model's overall performance plateaued at lower accuracy and F1-score levels compared to BiLSTM. The superior performance of BiLSTM therefore highlights the importance of incorporating temporal and contextual awareness in sentiment analysis, particularly for applications that involve nuanced or context-dependent text such as product reviews, customer opinions, or digital market discussions.

Table 1 Model performance comparison based on Accuracy and F1-score.

Model	Accuracy	F1-score
BiLSTM	0.7382	0.7240
CNN	0.6510	0.6428

The same performance comparison is visually presented in figure 2, which highlights the clear superiority of the BiLSTM model over the CNN across both accuracy and F1-score metrics. The bar chart shows that BiLSTM achieved consistently higher scores in every evaluation aspect, with a performance margin of approximately eight to nine percentage points above CNN. This margin reflects a meaningful improvement in the

model's ability to learn and generalize from sequential data. The visual representation emphasizes that BiLSTM not only achieves better quantitative performance but also maintains stability across training and validation phases. This indicates that its architecture can effectively retain important contextual information over longer textual sequences, which is crucial for distinguishing subtle differences in sentiment. The ability to process input in both directions allows BiLSTM to consider how earlier and later words influence meaning, leading to a more complete semantic understanding of each sentence.

In contrast, the CNN model's lower bars in figure 2 suggest that it struggles to extract sufficient contextual depth from text data, particularly when sentiment is influenced by word order or syntactic relationships. CNN tends to focus on detecting local features through convolutional filters, which is beneficial for identifying frequent word patterns but less effective for interpreting longer dependencies that span multiple phrases or sentences. The smaller gap between CNN's accuracy and F1-score further implies that while the model can recognize some surface-level sentiment cues, it often fails to generalize effectively to unseen examples. This limitation becomes more evident when dealing with complex or ambiguous sentences where emotional tone is implied rather than explicitly stated. Therefore, the visual findings in figure 2 reinforce the conclusion that BiLSTM's bidirectional structure provides a more powerful mechanism for contextual learning, allowing it to outperform CNN consistently in sentiment classification tasks.

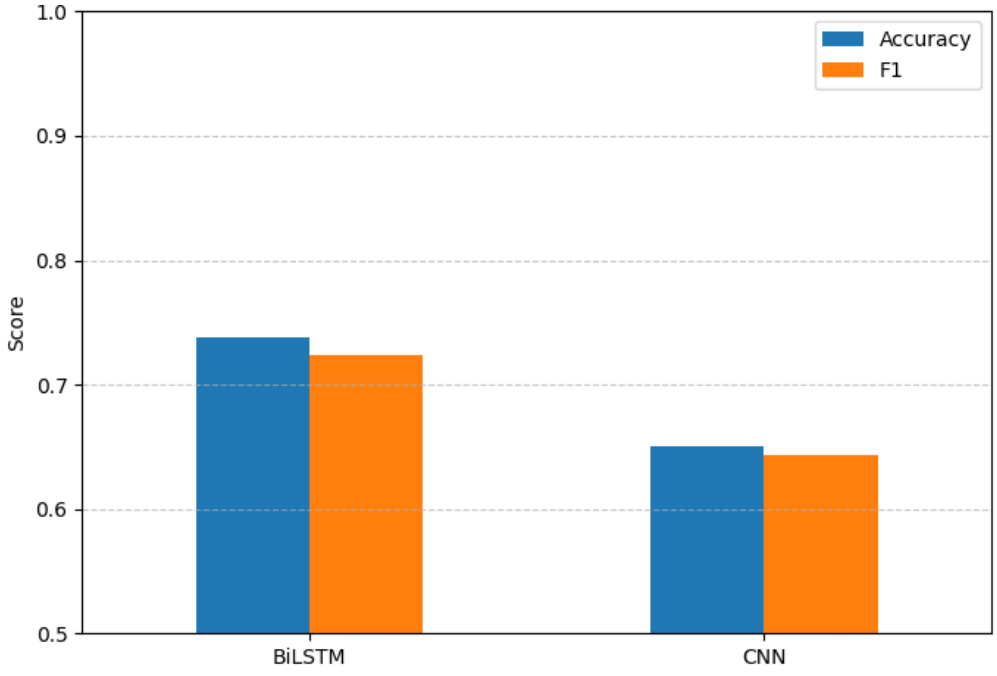

Figure 2 Full Epoch Model Comparison between BiLSTM and CNN.

During the training phase, both models were closely monitored for their validation accuracy across twenty epochs to analyze their convergence behavior and learning progression. As illustrated in figure 3, both BiLSTM and CNN demonstrated a rapid increase in validation accuracy during the initial epochs, suggesting that both models quickly captured fundamental sentiment patterns from the dataset. However, starting around the seventh epoch, the BiLSTM began to exhibit a clear advantage, showing a steadier and more continuous improvement in performance compared to CNN. By the later stages of training, the BiLSTM reached a validation accuracy of approximately 0.71, indicating that it continued to learn and refine its internal representations even after multiple epochs. This sustained improvement suggests that BiLSTM possesses a higher learning capacity, enabling it to extract more meaningful contextual and sequential features from text data. The gradual rise in validation performance also demonstrates that the BiLSTM model was able to maintain stability throughout the training process, avoiding fluctuations that typically indicate unstable or erratic learning behavior.

In contrast, the CNN model's validation accuracy plateaued around 0.65 after a few epochs, showing minimal improvement as training progressed. This stagnation suggests that CNN reached its optimal feature extraction limit relatively early, which may be due to its focus on local text patterns rather than long-range dependencies. CNN models are often effective at identifying frequent keyword combinations but struggle when sentiment meaning depends on broader sentence structure or context. The difference in learning trends between the two models indicates that BiLSTM was better suited to generalize across unseen data, as it could incorporate temporal and contextual information during training. Additionally, the BiLSTM's smooth and consistent validation curve implies that the use of dropout and batch normalization effectively prevented overfitting, allowing the model to retain generalization capabilities across epochs. Meanwhile, the CNN's early convergence may reflect underfitting, as it failed to capture deeper semantic relationships required for accurate sentiment prediction in complex textual data.

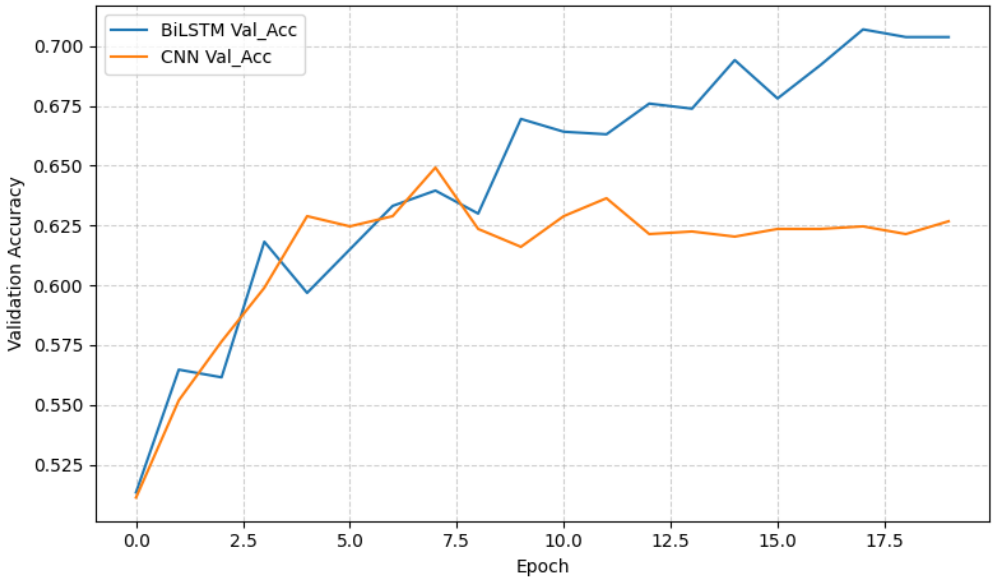

Figure 3 Validation Accuracy Across.

A more detailed evaluation of the BiLSTM model's classification performance based on precision, recall, and F1-score provides deeper insight into how effectively the model handled each sentiment category. The model performed exceptionally well in identifying neutral sentiments, achieving a precision of 0.76, recall of 0.89, and an F1-score of 0.82, which indicates that BiLSTM was able to recognize and generalize neutral expressions with high reliability. This strong performance suggests that the model effectively captured the structural and lexical patterns commonly associated with neutral language, such as objective statements or emotionally balanced expressions. The positive sentiment class also showed commendable results, with an F1-score of 0.71, reflecting the model's ability to detect affirmative tone and positive contextual indicators within the text. In contrast, the negative sentiment class recorded an F1-score of 0.42, which reveals that BiLSTM struggled to differentiate subtle negative expressions from neutral ones. This limitation can be attributed to the relatively lower frequency of negative samples in the dataset and the inherent ambiguity in how negativity is expressed linguistically. Negative sentiments are often conveyed through indirect cues, sarcasm, or mild disapproval, which may cause the model to confuse them with neutral phrases. The imbalance in class distribution likely amplified this issue, leading to reduced recall for negative instances. Despite this, the overall class-wise results indicate that BiLSTM maintained strong generalization for dominant sentiment categories and achieved reliable consistency across the dataset, confirming its robustness in multi-class sentiment classification tasks.

To gain a more comprehensive understanding of the BiLSTM model's predictive behavior, a confusion matrix was constructed, as presented in figure 4. The matrix provides a clear visualization of how accurately the

model categorized each sentiment class and where misclassifications occurred. The results reveal that the model performed strongly for the neutral and positive sentiment categories, with the majority of instances correctly identified. Specifically, the BiLSTM correctly predicted 551 out of 622 neutral samples, demonstrating its ability to accurately capture the linguistic and contextual patterns commonly associated with neutral statements. For the positive category, the model achieved 255 correct predictions out of 372 samples, which indicates its capability to detect affirmative or optimistic expressions effectively. This strong performance in the neutral and positive classes suggests that the BiLSTM architecture was able to capture the dominant sentiment features in the dataset, particularly those associated with moderate and positive tones. These outcomes align with the model's high F1-scores for both classes and support the notion that sequential dependency modeling is effective for understanding sentiment polarity when the language used is explicit and contextually consistent.

However, the confusion matrix also highlights notable weaknesses in the model's classification of negative sentiments. Out of 175 negative samples, only 57 were correctly identified, while a significant portion, totaling 80 instances, was misclassified as neutral. This pattern suggests that the model struggled to recognize subtle linguistic cues that convey negative sentiment, particularly when they are expressed indirectly or embedded within complex sentence structures. Many negative expressions in textual data tend to use mild criticism, sarcasm, or mixed sentiment, which can resemble neutral language and lead to misclassification. This misinterpretation could also stem from class imbalance, where the model was exposed to fewer negative examples during training, limiting its ability to learn distinctive negative patterns. The overall results emphasize that while the BiLSTM achieved high reliability for neutral and positive sentiments, future research could improve its performance on negative sentiment detection through techniques such as data augmentation, class reweighting, or fine-tuning with sentiment-specific embeddings. These enhancements would help the model better capture nuanced emotional expressions and increase its robustness across all sentiment categories.

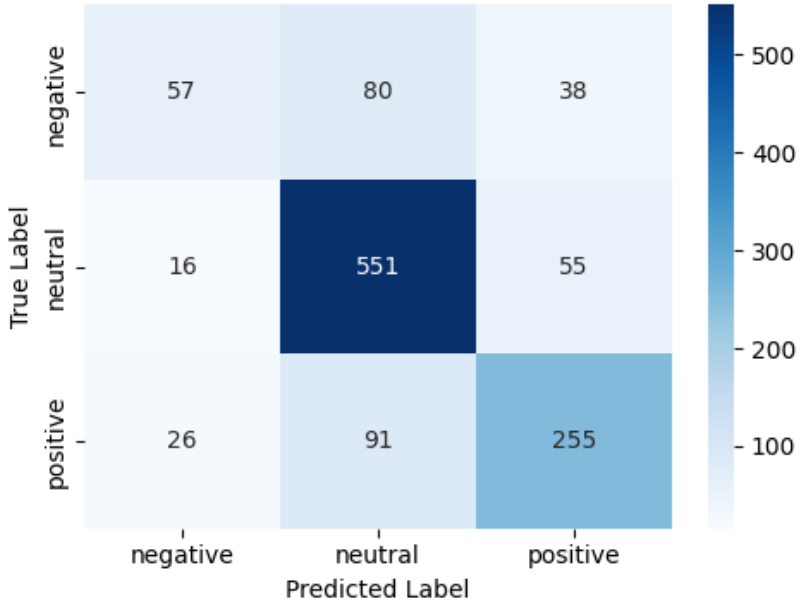

Figure 4 Confusion Matrix of the BiLSTM Model.

In comparison, the CNN model exhibited weaker performance across all sentiment categories, showing particular difficulty in identifying negative sentiments, where it achieved an F1-score of only 0.26. The confusion matrix for CNN, although not displayed here, indicated a higher degree of misclassification and more dispersed prediction patterns across all classes. This dispersion suggests that the CNN model struggled to distinguish between sentiment categories that shared overlapping linguistic features, especially in sentences containing subtle contextual or emotional cues. The architecture of CNN, which primarily relies on convolutional filters to extract local n-gram features, allows it to detect surface-level patterns such as frequently co-occurring words or short expressions of emotion. However, this localized feature extraction limits its ability to capture the broader contextual dependencies that often determine the true sentiment of a statement. For example, expressions containing negations or multi-word constructs that invert sentiment polarity are difficult for CNN to interpret accurately because it processes text in smaller, independent segments. As a result, the CNN tends to overgeneralize, leading to confusion between neutral and negative or positive and neutral categories. Although CNN models are known for their computational efficiency and effectiveness in text classification tasks with strong local features, these results demonstrate that sentiment analysis requires a deeper contextual understanding that CNN alone cannot provide. This limitation reinforces the need for sequential models such as BiLSTM, which can account for the dynamic relationships between words and phrases across an entire sentence, resulting in more accurate and context-aware sentiment classification.

Overall, the experimental results demonstrate that the BiLSTM model outperformed CNN in both overall accuracy and class-specific performance. The BiLSTM's sequential bidirectional processing enabled more effective extraction of contextual features, leading to better generalization across sentiment classes. The consistent improvement in validation accuracy and strong classification performance further indicate that BiLSTM is highly suitable for sentiment analysis in digital market-related applications, where understanding nuanced expressions and contextual sentiment is essential for accurate interpretation of consumer opinions and market trends.

Discussion

The findings of this study provide a deeper understanding of how architectural differences between BiLSTM and CNN affect sentiment analysis performance on textual data. The BiLSTM model demonstrated superior accuracy and F1-score, confirming its strength in learning sequential dependencies and contextual nuances that are essential for language understanding [21]. This advantage stems from the bidirectional structure of BiLSTM, which allows the model to analyze each word in a sentence with awareness of both its preceding and succeeding context [22]. As a result, BiLSTM can interpret linguistic phenomena such as negations, intensifiers, and dependencies that span across multiple words or clauses. This capability is particularly important for sentiment analysis, where meaning is often influenced by sentence structure and subtle variations in phrasing [23]. For instance, sentences containing phrases like "not bad" or "could be better" require the model to understand that the presence of negation changes the overall sentiment polarity. CNN, in contrast, relies heavily on localized convolutional filters that capture short-term patterns and n-gram relationships, which makes it less effective in interpreting long-range dependencies [24]. Although CNN models can identify strong sentiment signals based on specific keywords, they often fail to accurately classify sentences where context alters the meaning of those words. This explains why CNN tends to perform well on explicitly polarized text but poorly on more complex or context-dependent statements [25].

Another key insight from this research is the stability and generalization strength of the BiLSTM model during training. The validation accuracy of BiLSTM continued to improve gradually across all twenty epochs without signs of overfitting, suggesting that the regularization strategies applied, such as dropout and batch normalization, were effective in maintaining model robustness [26]. This steady learning curve indicates that BiLSTM was able to progressively refine its internal representation of sentiment features, leading to more consistent performance across different data samples. In contrast, the CNN model converged more rapidly but reached a plateau at a lower accuracy level, implying limited capacity to extract deeper semantic features after early training stages. This behavior aligns

with the structural characteristics of CNN, which is generally better suited for analyzing local patterns, whereas sequential models are more capable of representing contextual relationships in textual data [21], [24]. The findings highlight that sentiment analysis, particularly in the realm of digital communication and consumer feedback, requires architectures that can capture linguistic dependencies over longer contexts rather than relying solely on local textual cues [22], [23]. From a practical standpoint, the superior contextual understanding of BiLSTM makes it highly applicable in digital market analytics, where organizations must interpret nuanced opinions from social media, product reviews, and cryptocurrency discussions. The ability to accurately model sentiment polarity from these sources can significantly enhance market intelligence, consumer insight generation, and strategic decision-making in data-driven marketing and digital asset management [25].

Conclusion

The results of this study clearly demonstrate that the Bidirectional Long Short-Term Memory (BiLSTM) model outperformed the Convolutional Neural Network (CNN) in sentiment analysis tasks involving textual data. Through a comprehensive evaluation over twenty epochs, the BiLSTM achieved higher accuracy and F1-scores, confirming its superior ability to model sequential dependencies and capture contextual semantics. The bidirectional processing of BiLSTM allows it to understand both past and future word relationships, making it more effective in handling complex sentence structures and nuanced expressions compared to the CNN, which primarily focuses on local patterns. This enhanced contextual understanding enabled the BiLSTM to achieve strong classification performance, particularly in identifying neutral and positive sentiments. However, the model exhibited some difficulty in recognizing negative sentiments, likely due to dataset imbalance and linguistic ambiguity. Overall, the findings reaffirm that BiLSTM is better suited for natural language tasks that require comprehensive contextual comprehension and semantic interpretation.

From a practical perspective, the outcomes of this research hold significant implications for digital marketing, e-commerce analytics, and cryptocurrency sentiment monitoring. Accurate sentiment classification using deep learning can provide organizations with valuable insights into consumer behavior, brand perception, and market sentiment, which can inform strategic decision-making and enhance customer engagement. The success of BiLSTM in this study demonstrates that deep contextual models are highly effective tools for processing large-scale unstructured data in digital markets. Future work can build upon these findings by exploring transformer-based architectures such as BERT or RoBERTa, which are designed to capture even more sophisticated contextual relationships. Additionally, addressing class imbalance and expanding the dataset to include more diverse sentiment expressions would further

improve the model's robustness and generalization capability. In conclusion, this study contributes to the growing application of deep learning in sentiment analysis for digital market intelligence, underscoring the potential of AI-driven models in supporting data-informed marketing and financial strategies.

Declarations

Author Contributions

Conceptualization: H.B., M.S.K.; Methodology: H.B., W.W.; Software: M.S.K., W.W.; Validation: H.B., M.S.K.; Formal Analysis: H.B.; Investigation: M.S.K., W.W.; Resources: H.B., W.W.; Data Curation: M.S.K.; Writing – Original Draft Preparation: H.B.; Writing – Review and Editing: M.S.K., W.W.; Visualization: W.W.; 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] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 36, no. 4, pp. 1–12, Apr. 2024, doi: 10.1016/j.jksuci.2024.102048.

[2] A. Kumar Durairaj and A. Chinnalagu, "Transformer based contextual model for sentiment analysis of customer reviews: A fine-tuned BERT," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 11, pp. 353–361, Nov. 2021, doi: 10.14569/IJACSA.2021.0121153.

[3] M. Yekrangi and N. S. Nikolov, "Domain-specific sentiment analysis: An optimized deep learning approach for the financial markets," *IEEE Access*, vol. 11, no. 2023, pp. 70248–70262, Jul. 2023, doi: 10.1109/ACCESS.2023.3293733.

[4] H. You, J. Yang, and Z. Xu, “Sentiment analysis based on CNN and BiLSTM model,” *Commun. Signal Process. Syst.*, vol. 2023, no. Jun., pp. 175–182, Jun. 2023, doi: 10.1007/978-981-99-1260-5_22.

[5] H. Bashiri and H. Naderi, “Comprehensive review and comparative analysis of transformer models in sentiment analysis,” *Knowl. Inf. Syst.*, vol. 66, no. 2024, pp. 7305–7361, 2024, doi: 10.1007/s10115-024-02214-3.

[6] A. Mu and X. Li, “An enhanced CNN-BiLSTM hybrid model for natural disaster tweets sentiment analysis,” *Biomimetics*, vol. 9, no. 9, pp. 1–12, Sep. 2024, doi: 10.3390/biomimetics9090533.

[7] Z. Ahmed and J. Wang, “A fine-grained deep learning model using embedded CNN with BiLSTM for exploiting product sentiments,” *Alexandria Eng. J.*, vol. 61, no. 2022, pp. 731–744, 2022, doi: 10.1016/j.aej.2022.10.037.

[8] A. Kaur, A. Kaur, and A. Deotare, “Sentiment analysis of consumer reviews using enhanced attention-CNN-BiLSTM networks,” *Data Anal. Decision Making Toward Business Excellence*, vol. 2025, no. Jun., pp. 157–178, Jun. 2025, doi: 10.1007/978-981-95-2429-7_6.

[9] S. Yang, J. Xing, Z. Liu, and Y. Sun, “Short-text sentiment classification model based on BERT and gated attention transformer,” *Electronics*, vol. 14, no. 19, pp. 1–12, Oct. 2025, doi: 10.3390/electronics14193904.

[10] M. M. Rahman, A. I. Shiplu, Y. Watanobe, and M. A. Alam, “RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis,” *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 9, no. 6, pp. 3788–3805, Dec. 2025, doi: 10.1109/TETCI.2025.3572150.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL*, vol. 2019, no. Jun., pp. 4171–4186, Jun. 2019, doi: 10.18653/v1/N19-1423.

[12] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” *ICLR Workshop*, vol. 2020, no. Apr., pp. 1–10, Apr. 2020, doi: 10.48550/arXiv.1908.10063.

[13] Y. Kim, “Convolutional neural networks for sentence classification,” *EMNLP*, vol. 2014, no. Oct., pp. 1746–1751, Oct. 2014, doi: 10.3115/v1/D14-1181.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *NeurIPS*, vol. 30, no. Dec., pp. 5998–6008, Dec. 2017, doi: 10.48550/arXiv.1706.03762.

[15] R. Socher *et al.*, “Recursive deep models for semantic compositionality over a sentiment treebank,” *EMNLP*, vol. 2013, no. Oct., pp. 1631–1642, Oct. 2013, doi: 10.5555/2075116.2075281.

[16] I. Hagenau, M. Liebmann, and D. Neumann, “Automated news reading: Stock price prediction based on financial news using context-capturing features,” *Decis. Support Syst.*, vol. 50, no. 2, pp. 191–201, Jan. 2011, doi: 10.1016/j.dss.2010.11.017.

[17] D. Araci, “FinBERT: Financial sentiment analysis with pre-trained language models,” *IEEE Big Data*, vol. 2019, no. Dec., pp. 166–172, Dec. 2019, doi: 10.1109/BigData.2019.8902558.

[18] S. Yadav, C. K. Jha, A. Sharan, and V. Vaish, “Sentiment analysis of financial news using unsupervised approaches,” *Procedia Comput. Sci.*, vol. 167, no. 2020, pp. 2205–2214, 2020, doi: 10.1016/j.procs.2020.03.325.

[19] P. C. Tetlock, “Giving content to investor sentiment: The role of media in the stock market,” *J. Finance*, vol. 62, no. 3, pp. 1139–1168, Jun. 2007, doi: 10.1111/j.1540-6261.2007.01232.x.

[20] M. Yekrangi and N. Abdolvand, “Financial markets sentiment analysis: Developing a specialized lexicon,” *J. Intell. Inf. Syst.*, vol. 57, no. 2021, pp. 127–146, 2021, doi: 10.1007/s10844-020-00630-9.

[21] M. R. Mahim, F. B. Z. Islam, M. Mahmud, M. Hasan, and S. R. Ahona, “Context-aware sentiment analysis of e-commerce reviews using a BERT-CNN-BiLSTM hybrid model,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 17, no. 6, pp. 1–10, Jun. 2026, doi: 10.14569/ijacsa.2026.0170699.

[22] N. Zheng and C. Du, “BiLSTM-attention for sentiment analysis and reader review topic mining in online literature works,” *Big Data Inform. Education*, vol. 2026, no. Jun., pp. 1–10, Jun. 2026, doi: 10.1145/3806980.3807084.

[23] M. I. U. Haq, K. Mahmood, Q. Li, A. K. Das, S. Shetty, and M. Hussain, “Efficiently learning an encoder that classifies token replacements and masked permuted network-based BIGRU attention classifier for enhancing sentiment classification of scientific text,” *IEEE Access*, vol. 12, no. 2024, pp. 190240–190254, Dec. 2024, doi: 10.1109/ACCESS.2024.3516946.

[24] Y. Dong, X. Li, M. He, and J. Li, “DC-BiLSTM-CNN algorithm for sentiment analysis of Chinese product reviews,” *Appl. Artif. Intell.*, vol. 39, no. 2025, pp. 1–15, 2025, doi: 10.1080/08839514.2025.2461809.

[25] N. Abdelhady, T. Soliman, and M. F. Farghally, “Stacked-CNN-BiLSTM-COVID: An effective stacked ensemble deep learning framework for sentiment analysis of Arabic COVID-19 tweets,” *J. Cloud Comput.*, vol. 13, no. 2024, pp. 1–15, 2024, doi: 10.1186/s13677-024-00644-6.

[26] A. Abulfaraj, “BiLSTM-based parallel CNN models with attention and ensemble mechanism for Twitter sentiment analysis,” *IEEE Access*, vol. 13, no. 2025, pp. 141139–141155, Aug. 2025, doi: 10.1109/ACCESS.2025.3597413.