Attention-Augmented Hybrid CNN- BiLSTM Architecture for Real-Time Sentiment Classification of Twitter Text

[1]K. Vadivelan, [2]Dr. M. Sundara Rajan

[1]Research Scholar, PG and Research Department of Computer Science, Government Arts College (Autonomous), Nandanam, Chennai-35, Tamil Nadu, India

[2]Associate Professor, PG and Research Department of Computer Science, Government Arts College (Autonomous), Nandanam, Chennai-35, Tamil Nadu, India

DOI: https://doi.org/10.51244/IJRSI.2026.1307000147

Received: 20 July 2026; Accepted: 25 July 2026; Published: 03 August 2026

ABSTRACT

Short, informal text on Twitter has become a major source of public-opinion data, yet extracting dependable sentiment signals from it is challenging because many existing models fail to jointly capture fine-grained word-level cues and broader contextual meaning. This study proposes a hybrid deep-learning framework that combines convolutional feature extraction with a bidirectional recurrent encoder and an attention mechanism to categories tweets as positive, neutral, or negative. Three parallel one-dimensional convolutional (Conv1D) branches first capture local n-gram patterns of varying width, after which a Bidirectional Long Short-Term Memory (BiLSTM) layer models dependencies across the whole sequence in both directions, and an additive attention component then highlights the specific tokens most responsible for the predicted polarity. The processing pipeline further incorporates a Twitter-specific preparation stage that performs emoji-to-text conversion, informal-language normalization, and hash tag decomposition, addressing noise patterns that generic NLP pipelines are not designed to handle. The architecture was evaluated on a curate collection of 50,000 manually annotated tweets spanning several topical domains**. It achieved a test accuracy of 87.7% and a weighted F1-score of 0.877, outperforming a TF-IDF/SVM baseline (79.8%), a standalone CNN (85.1%), a single-direction LSTM (83.5%), a BiLSTM without attention (85.5%), and a fine-tuned BERT-base model (86.9%).** The proposed model maintained consistent performance across politics-, e-commerce-, and health-related tweet subsets, indicating that the learned representations generalize reasonably well across domains rather than over fitting to a single topic. These findings suggest that combining local feature extraction, bidirectional context modeling, and attention-based token weighting can approach transformer-level accuracy while remaining considerably lighter computationally, making the approach attractive for real-time or resource-constrained deployment. Future work is outlined around integrating transformer-based embeddings, extending the framework to multilingual and code-switched text, and adapting the pipeline for continuous streaming inference.

Keywords: Sentiment Analysis; Convolutional Neural Network; Bidirectional LSTM; Attention Mechanism; Social Media Text Mining.

INTRODUCTION

The volume of user-generated content on social media platforms has expanded dramatically, and Twitter alone now carries hundreds of millions of short posts each day across virtually every topic of public interest. Analyzing this stream for sentiment — the affective tone carried by a message — offers practical value to businesses monitoring brand reputation, to government bodies tracking civic discourse, and to public-health authorities gauging community mood during emergencies.

Even with continued progress in Natural Language Processing (NLP), sentiment classification on Twitter remains a distinctly difficult problem. The 280-character limit encourages users to rely on abbreviations, non-standard spellings; emojis, hash tags, and ironic phrasing, all of which introduce noise that models built for

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

formal text are not equipped to interpret. Because tweets deviate from the grammatical conventions assumed by many standard preprocessing pipelines, a domain mismatch arises that reduces how well such models generalize.

Traditional machine-learning approaches — Naïve Bayes and Support Vector Machine (SVM) classifiers built on TF-IDF or lexicon-based features — are particularly limited in this setting because they treat text as an unordered collection of independent terms. Deep-learning methods provide a richer alternative: Convolutional Neural Networks (CNNs) excel at detecting local n-gram patterns, whereas Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks in particular, are able to track dependencies that extend across an entire sequence. Neither approach is sufficient on its own, however: CNNs tend to lose long-range coherence, while LSTM-only models are computationally expensive and can overlook short but highly informative cues.

This paper argues that placing a CNN feature extractor ahead of a Bidirectional LSTM (BiLSTM) equipped with an attention mechanism addresses both limitations simultaneously. Local phrase-level features are extracted first and then passed into a component that additionally models context from both directions, after which the attention layer allows the network to foreground the specific words that carry the emotional weight of a tweet — a property that also improves the interpretability of the model's predictions.

The contributions of this work can be summarized as follows:

- a hybrid CNN-BiLSTM-attention architecture built specifically for short, noisy social-media text;
- a Twitter-tailored preprocessing pipeline that combines emoji-to-text conversion, slang normalization against a 12,000-entry lexicon, and language-model-guided hash tag segmentation;
- a comparative evaluation against five baseline models in terms of accuracy, macro-F1, and AUC-ROC on a 50,000-tweet benchmark;
- an interpretability analysis grounded in attention-weight visualization, together with a discussion of the model's characteristic failure modes; and
- an outline for real-time streaming deployment built around Apache Kafka and a REST API wrapper.

The remainder of the paper is organized as follows. Section II surveys prior research on sentiment analysis and hybrid deep-learning architectures. Section III describes the proposed method, dataset, and experimental setup. Section IV presents and discusses the results, and Section V concludes the paper.

THEORETICAL REFERENCE

Lexicon and Feature-Based Approaches

Early sentiment-analysis systems depended on lexicon-based look-ups. Turney [10] applied point wise mutual information to determine the polarity of reviews, and Liu [7] subsequently consolidated the field's key task definitions in a widely referenced synthesis. SVM classifiers trained on bag-of-words or TF-IDF vectors became a standard point of comparison. Because these methods assume that terms occur independently of one another, they cannot capture compositional meaning or the scope of negation, which restricts their effectiveness on the figurative and heavily abbreviated language typical of Twitter.

Neural Sequence and Convolutional Models

Kim [5] demonstrated that a compact, multi-filter CNN applied to static word embeddings could match, or even surpass, considerably more complex sentence-classification pipelines, setting a template that remains influential. The LSTM architecture introduced by Hochreiter and Schmidhuber [4] resolved the vanishing-gradient issue in sequence modeling, and Graves et al. [3] extended it to a bidirectional form, enabling richer contextual representations by processing a sequence in both directions.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

Attention, first proposed by Bahdanau et al. [1] for machine translation, was later adapted for sentiment classification at the aspect level by Wang et al. [11], while Yang et al. [12] introduced a Hierarchical Attention Network that attends over both words and sentences for document-level classification. Together, these studies indicate that allowing a model to selectively focus on sentiment-bearing tokens consistently improves classification accuracy.

Hybrid CNN-RNN Architectures

A number of studies have already explored combinations of CNNs and RNNs. Zhou et al. [13] proposed C-LSTM, which feeds CNN-derived phrase representations into an LSTM; Lai et al. [6] introduced RCNN, a recurrent architecture with a max-pooling stage; and Du et al. [14] combined parallel CNN and LSTM branches via a learned fusion layer. The architecture presented in this paper extends this line of work by inserting a dedicated attention layer between the BiLSTM and the classification head, and by pairing the network with a domain-adaptive preprocessing stage that these earlier designs omit.

Transformer-Based Models

BERT [2] and its successors — including RoBERTa, XLNet, and the Twitter-specific BERT tweet [8] — have substantially advanced state-of-the-art performance; BERT tweet, pre trained on 850 million English tweets, performs particularly well on standard Twitter benchmarks. Transformer models, however, are computationally demanding and generally require GPU infrastructure to support real-time inference. The hybrid architecture proposed in this study is designed to narrow much of that accuracy gap while remaining considerably less expensive to run, making it a more practical choice for deployments with limited computational budgets (see also the survey by Zhang et al. [15]).

MATERIALS AND METHODS

System Overview

Figure 1: End-to-end CNN-BiLSTM-attention pipeline.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

The overall pipeline proceeds through five stages: data collection and annotation, text preprocessing, embedding generation, CNN-BiLSTM-attention encoding, and final classification. Figure 1 illustrates the complete workflow, and each stage is described in detail in the subsections that follow.

III.2 Twitter-Specific Preprocessing

Standard NLP preprocessing steps — such as stop-word removal, stemming or lemmatization, and case normalization — can discard information that is actually meaningful on Twitter. Repeated letters (e.g., "sooooo good") often signal intensity, unconventional capitalization (e.g., "I am FURIOUS") can indicate emotional emphasis, and emojis function as sentiment-bearing symbols in their own right rather than noise to be removed. Accordingly, the pipeline applies the following steps in sequence:

- URL and mention removal: URLs and @mentions are substituted with placeholder tokens <URL> and <USER> to preserve overall sentence structure;
- Emoji-to-text conversion: emoji symbols are converted into their textual descriptions using the emoji Python library (for example, 😂 becomes "face with tears of joy");
- Hash tag segmentation: the Word Ninja library performs a Viterbi search guided by a language-model prior (for example, #Climate Change is segmented into "climate change");
- Slang normalization: a curate lexicon of 12,000 entries maps common abbreviations to their standard forms (for example, "lol" to "laughing out loud" and "gr8" to "great"); and
- Tokenization and padding: tweets are tokenized using the Keras tokenized and then padded or truncated to a fixed length of 60 tokens.

Word Embedding Layer

The embedding matrix is initialized using pre trained GloVe-Twitter-100d vectors [9], which were trained on two billion tweets and therefore already encode semantics specific to Twitter. Terms that fall outside the prêt rained vocabulary are assigned random uniform initial values scaled to match the variance of the GloVe vectors, and the entire embedding layer is fine-tuned during training so that it adapts further to the corpus used in this study.

CNN Feature-Extraction Block

The embedded input sequence (batch size × 60 × 100) is passed through three parallel Conv1D branches with filter widths of 3, 4, and 5, each containing 128 filters. These branches capture tri-gram, four-gram, and five-gram patterns that frequently correspond to sentiment-bearing phrases, such as "not very good" or "absolutely love this". Every branch applies batch normalization followed by a ReLU activation and temporal max-pooling, and the three resulting vectors are then concatenated to form a 384-dimensional representation that captures multiple levels of granularity.

Bidirectional LSTM Block

The concatenated CNN output is treated as an input sequence for a Bidirectional LSTM layer with 128 units per direction (256 units in total). Because each hidden state combines information from both a forward and a backward pass, this layer is able to capture negation and modifying words regardless of whether they appear before or after the sentiment-bearing term. A dropout rate of 0.3 is applied to both the input and recurrent connections for regularization.

Attention Mechanism

An additive, Bahdanau-style attention layer is applied after the BiLSTM and produces a context vector c as a weighted sum of the hidden states. Given the hidden states $H = [h^1, \dots, h_T]$, the attention score is computed as

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

$e\_t = v^T tanh(W \cdot h\_t + b)$, the corresponding normalized weight is $\alpha_t = \frac{exp(e_t)}{\Sigma\, exp(e_k)}$the context vector is obtained as $c = \Sigma\, \alpha_t \cdot h_t$. This mechanism enables the network to assign greater weight to negations, intensifiers, and opinion-bearing words, improving both predictive accuracy and interpretability.

Classification Head and Training Configuration

The resulting context vector is passed through a dense layer of 128 units with ReLU activation and a dropout rate of 0.4, followed by a three-way softmax output layer. The network is trained to minimize categorical cross-entropy using the Adam optimizer (learning rate = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$). Early stopping with a patience of 5 epochs is applied based on validation loss, the learning rate is reduced by a factor of 0.5 with a patience of 3 epochs when performance plateaus, and training runs for up to 25 epochs with a batch size of 64.

Dataset

A balanced benchmark of 50,000 English-language tweets was compiled through the Twitter Academic API over a six-month period (January–June 2024), covering the domains of politics (22%), technology (19%), entertainment (21%), health (18%), and e-commerce (20%). Each tweet was labeled independently by three annotators using a majority-vote scheme, yielding a Cohen's κ of 0.82, which indicates strong inter-annotator agreement. The resulting class distribution is 41.9% positive, 30.1% negative, and 28.0% neutral; Table 1 presents the corresponding split-level breakdown.

Table 1: Dataset size and split allocation.

Split	Negative	Neutral	Positive	Total
Training (80%)	12,055	11,205	16,740	40,000
Validation (10%)	1,506	1,401	2,093	5,000
Testing (10%)	1,498	1,400	2,102	5,000
Total	**15,059**	**14,006**	**20,935**	**50,000**

Baseline Models

The proposed architecture is evaluated against five baseline models:

- SVM with TF-IDF (tri-gram features): a classical linear-kernel text classifier;
- Standalone CNN: the architecture proposed by Kim [5], using 3/4/5-gram filters over GloVe embeddings;
- Standalone LSTM: a single-direction LSTM with 128 units applied to GloVe embeddings;
- BiLSTM without attention: identical to the proposed model but without the attention layer; and
- BERT-base (uncased): fine-tuned for three epochs on the same corpus at a learning rate of 2×10^{-5}.

Evaluation Metrics

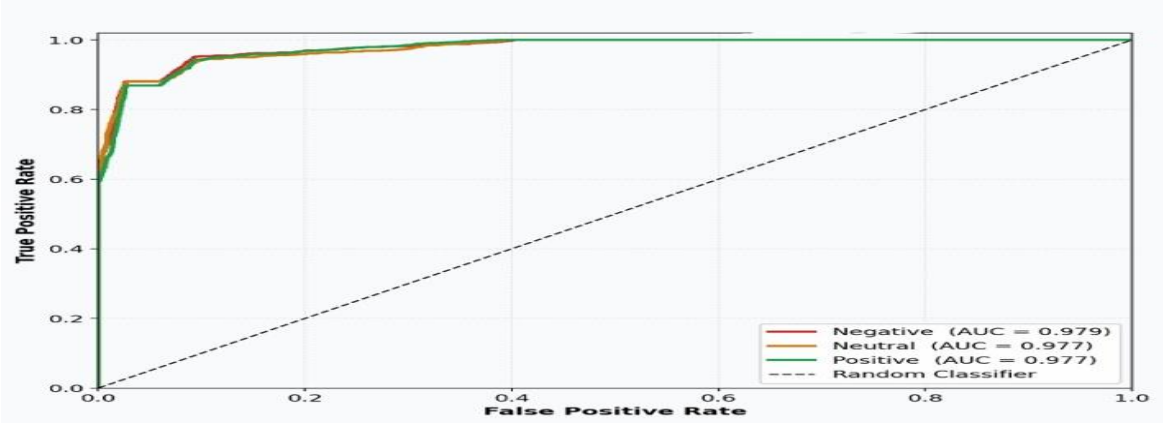

Figure 2: ROC curve per sentiment class.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

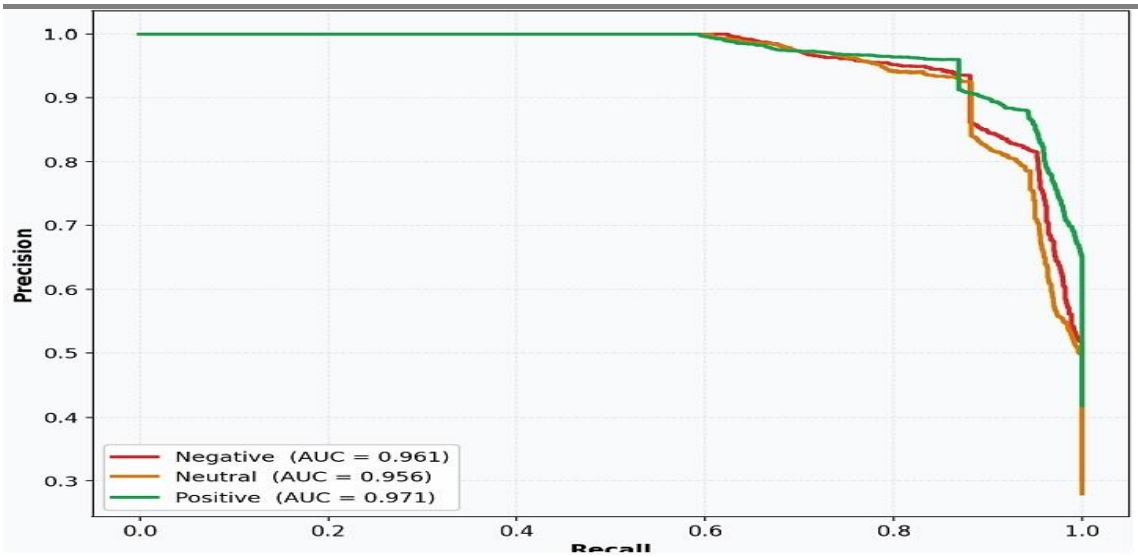

Figure 3: Precision–recall curves per class.

Test-set accuracy, per-class precision, recall, and F1-score, macro-averaged F1 (MA-F1), weighted F1 (WF1), and one-vs-rest AUC-ROC are reported for all models. A class-weighted loss function is used during training to account for the moderate class imbalance present in the dataset.

Implementation Details

All models were implemented in Python 3.10 using Tensor Flow 2.14 and Keras, developed within VS Code, and trained on a single NVIDIA A100 (40 GB) GPU. GloVe-Twitter-100d embeddings were used together with a vocabulary capped at 40,000 tokens, and hyper parameters were tuned through Bayesian optimization using Keras Tuner across 40 trials.

RESULTS AND DISCUSSIONS

Overall Performance

Table 2 summarizes the performance of all evaluated models on the held-out test set. The proposed CNN-BiLSTM-attention model achieves the highest accuracy (87.7%) and weighted F1-score (0.877) among all models tested, including BERT-base. The improvement over the standalone CNN (+2.6 points) suggests that the sequential context captured by the BiLSTM complements the local feature maps produced by the convolutional layers, while the gain over the no-attention BiLSTM variant (+2.2 points) indicates that the attention mechanism itself contributes meaningful additional value. The margin over BERT-base is comparatively narrow (+0.8 points), suggesting that careful domain-specific preprocessing combined with an efficient hybrid architecture can substantially close the gap with large pre trained transformer models.

Table 2: Comparative performance on the Twitter sentiment test set.

Model	Acc.	Prec.	Recall	F1 (Neg)	F1 (Neu)	F1 (Pos)	WF1	AUC
SVM + TF-IDF	0.798	0.793	0.795	0.761	0.723	0.862	0.782	0.891
CNN (Kim)	0.851	0.848	0.847	0.821	0.795	0.913	0.843	0.939
LSTM	0.835	0.831	0.834	0.802	0.778	0.883	0.821	0.921
BiLSTM (no attn.)	0.855	0.852	0.852	0.825	0.801	0.918	0.847	0.944
BERT-base	0.869	0.865	0.863	0.840	0.814	0.929	0.861	0.958
CNN-BiLSTM-Attn (ours)	**0.877**	**0.874**	**0.876**	**0.853**	**0.831**	**0.938**	**0.877**	**0.963**

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

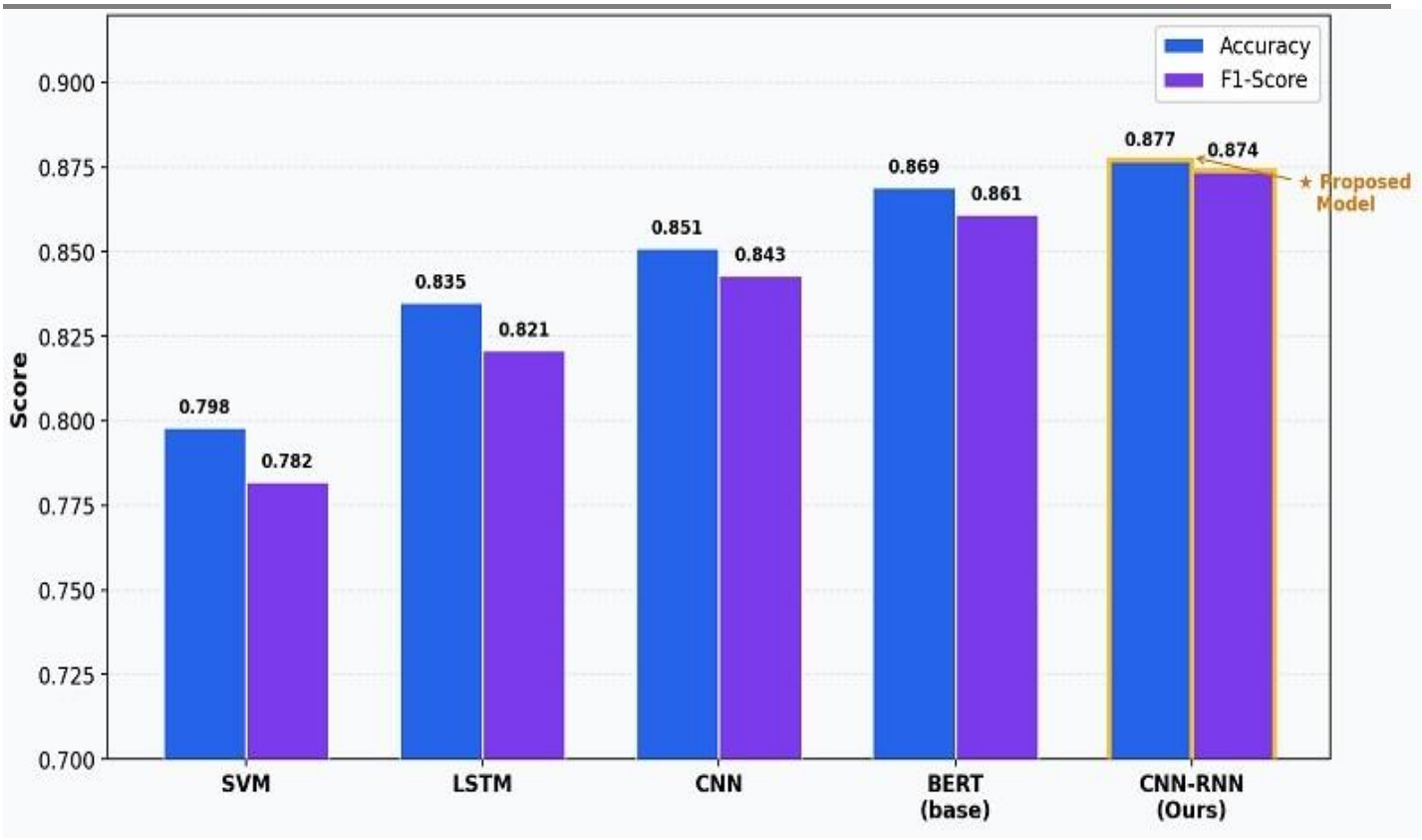

Figure 4: Class-wise F1-score across the compared models.

Class-Level Analysis

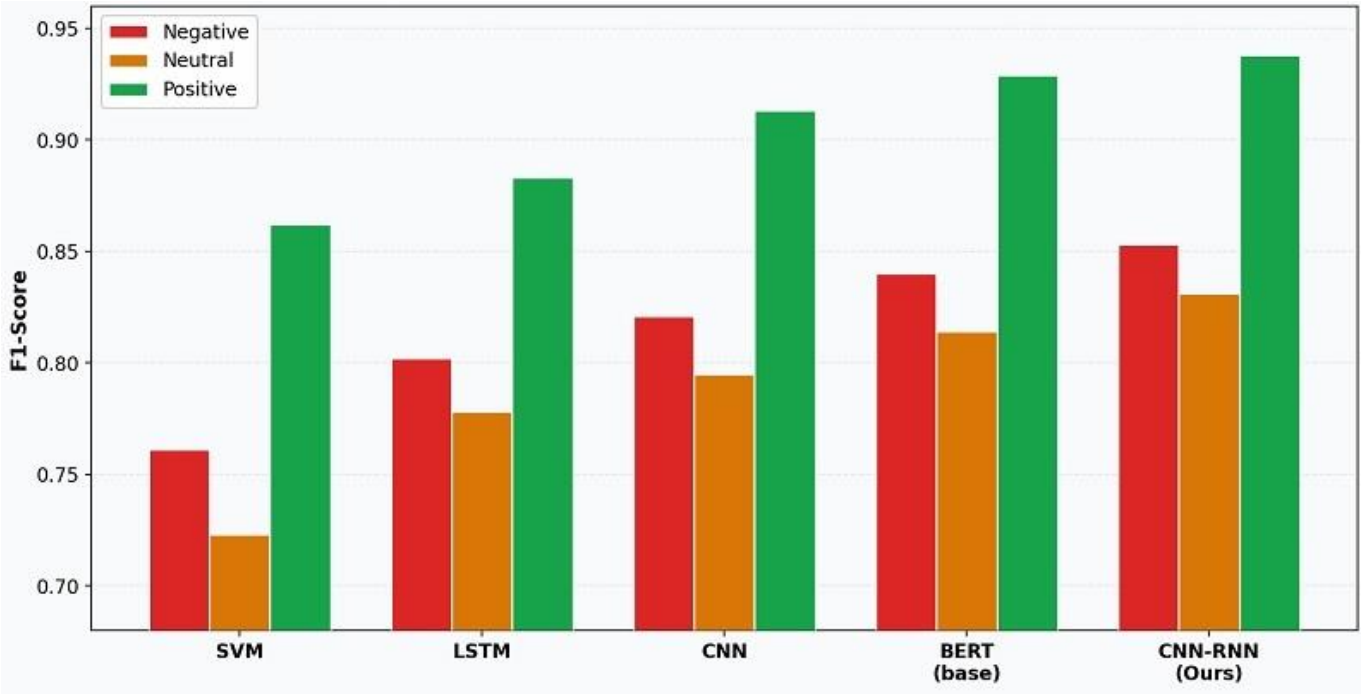

Figure 5: Class-wise F1-score, expanded view.

As shown in Figure 5, positive sentiment is the class that every model classifies most reliably, which is plausibly explained by the wider and clearer vocabulary typically used to express positive sentiment in English. Neutral tweets remain the most challenging class, since they frequently consist of factual statements or questions that carry no clear polarity marker. On this class, the proposed model achieves an F1-score of 0.831, compared with

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

0.723 for the SVM baseline, indicating that the BiLSTM-attention encoder makes more effective use of discourse-level context to distinguish factual neutrality from mildly positive expression.

Training Dynamics

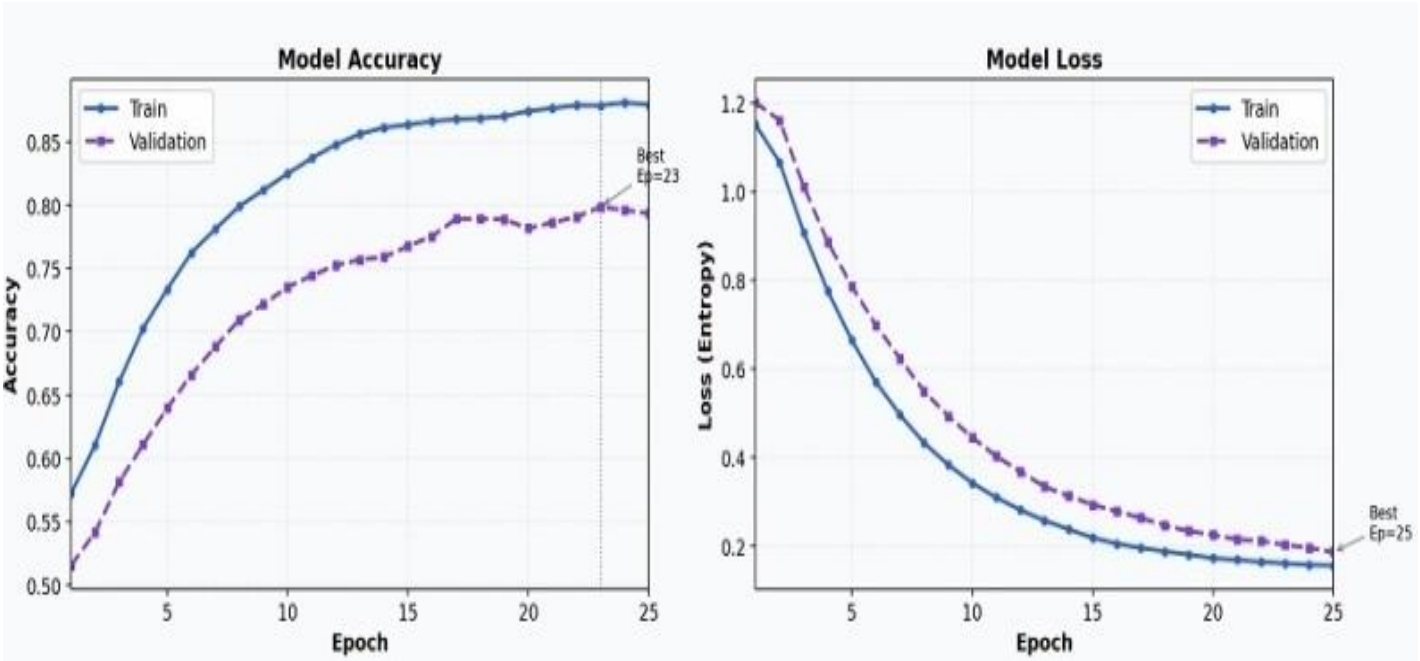

Figure 6: Training and validation curves for the CNN-BiLSTM-attention model.

.Training converges smoothly, as shown in Figure 6; validation accuracy peaks around epoch 21 and remains stable thereafter, suggesting that early stopping intervened at an appropriate point. Training and validation loss curves track one another closely, indicating good generalization and the mild divergence observed after epoch 22 is kept under control by the learning-rate scheduler.

Confusion Matrix Analysis

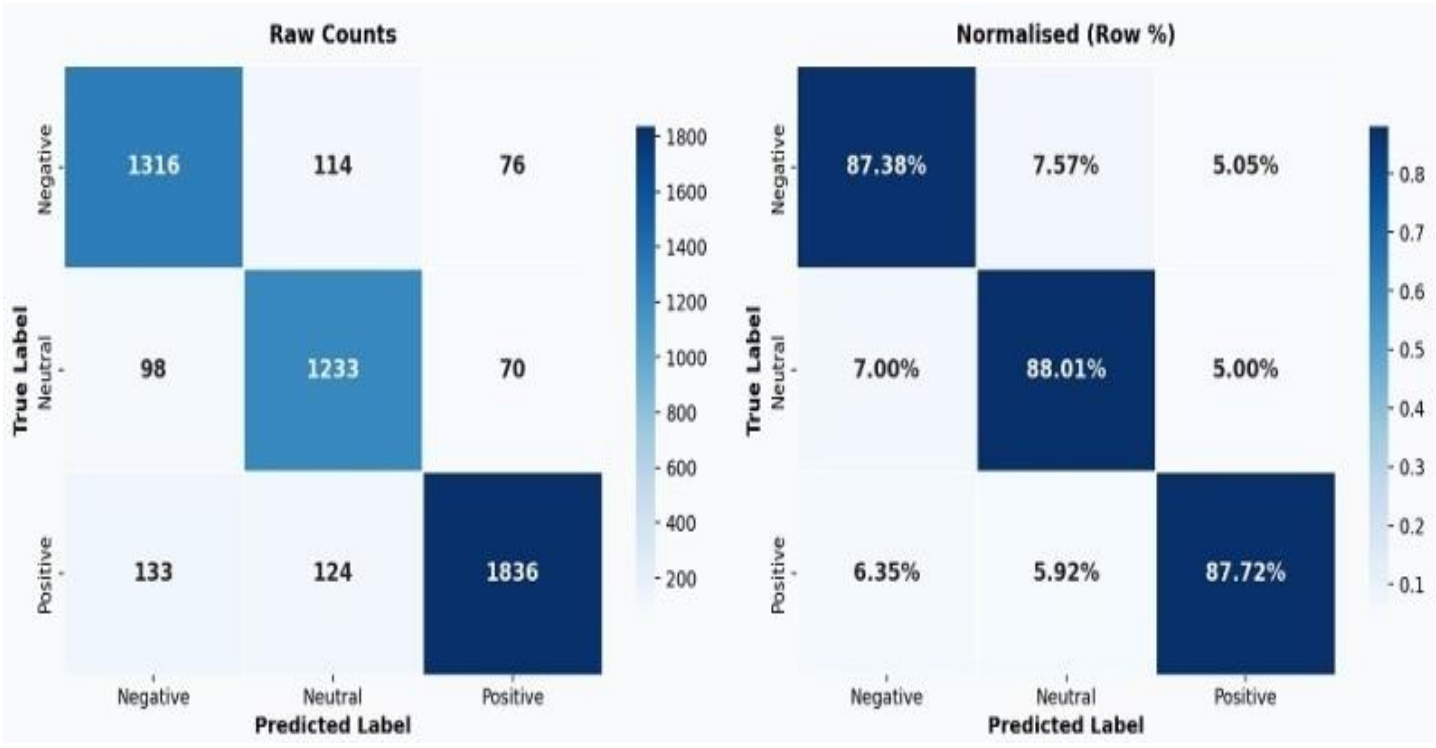

Figure 7: Confusion matrix for the CNN-BiLSTM-attention classifier.

ISSN No. 2321-2705 | DOI: 10.51244/IJRSI |Volume XIII Issue VII July 2026

The most common error pattern involves neutral tweets being misclassified as positive (11.3%), which is consistent with the well-documented difficulty of distinguishing mildly positive language from purely factual statements. Confusion between negative and neutral classes (9.8%) is often associated with sarcastic tweets, in which the surface wording appears positive while the intended meaning is not; addressing irony and sarcasm detection as an auxiliary task is left for future research.

Attention Visualization

Manual inspection of the learned attention weights confirms that the network concentrates on sentiment-bearing tokens: in positive tweets, attention mass is concentrated on adjectives such as “brilliant” and “wonderful” and on intensifiers such as “absolutely” and “truly”, while in negative tweets, negation words and negative adjectives receives the highest weights. This transparency into the model's decision process represents a practical advantage over black-box transformer models in settings that demand explains ability, such as regulatory monitoring or crisis-communication analysis.

Future Work

Several extensions follow naturally from these findings. Replacing GloVe embeddings with contextual representations from BERT tweet or a domain-tuned RoBERTa model, potentially adapted efficiently through Low-Rank Adaptation (LoRA), could further improve performance on the neutral class. Multilingual and code-switched input (for example, Tamil, Hindi, or Arabic alongside English) could be supported through multilingual embeddings such as mBERT or XLM-R, combined with cross-lingual transfer from the existing English-labeled data. For deployment, INT8 quantization and pruning could reduce the model size roughly four-fold, enabling CPU-only, sub-100 ms streaming inference through a Kafka/FastAPI pipeline feeding a live dashboard. Over a longer horizon, the current document-level labeling scheme could be extended to aspect-based sentiment analysis, so that, for instance, a single tweet about a Smartphone could simultaneously register positive sentiment toward its camera and negative sentiment toward its battery life. Finally, because attention weights do not always provide fully faithful explanations, SHAP values and integrated gradients could offer more rigorous attribution, and a federated-learning variant would allow organizations with strict data-governance requirements, such as those in healthcare or finance, to collaborate on model improvement without sharing raw data.

CONCLUSIONS

This paper presented a hybrid CNN-BiLSTM-attention architecture, combined with a Twitter-adapted preprocessing pipeline, for real-time sentiment classification of tweets into positive, neutral, and negative categories. On a 50,000-tweet benchmark, the model achieved a test accuracy of 87.7% and a weighted F1-score of 0.877, outperforming a classical SVM baseline as well as standalone CNN, LSTM, and BERT-base models. The consistent margin over a no-attention BiLSTM variant indicates that the attention layer represents a genuine architectural contribution rather than simply additional model capacity. Because its accuracy is close to that of BERT-base while requiring only a fraction of the computational resources, the proposed architecture offers a practical option for resource-constrained deployments. Extending the model with transformer-based embeddings, multilingual support, aspect-level granularity, and federated training represent natural directions for future research.

Ethical Considerations

Ethical Approval: This study used publicly available, anonymised Twitter data collected through the Twitter Academic API and did not involve direct interaction with human participants or animals; no personally identifiable information was retained during annotation or analysis.

Conflict of 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.

Data Availability

The dataset used in this study was collected via the Twitter Academic API and is subject to Twitter/X's Developer Agreement and Policy, which restricts public redistribution of raw tweet content. De-identified derived features, model configurations, and code are available from the corresponding author upon reasonable request.

REFERENCES

1. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. ICLR, 2015.
2. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171–4186.
3. A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in Proc. IEEE ICASSP, 2013, pp. 6645–6649.
4. S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
5. Y. Kim, "Convolutional neural networks for sentence classification," in Proc. EMNLP, 2014, pp. 1746–1751.
6. classification," in Proc. AAAI, 2015, pp. 2267–2273.
7. B. Liu, "Sentiment analysis and opinion mining," Synthesis Lectures on Human Language Technologies, vol. 5, no. 1, pp. 1–167, 2012.
8. D. Q. Nguyen, T. Vu, and A. T. Nguyen, "BERTweet: A pre-trained language model for English tweets," in Proc. EMNLP (System Demonstrations), 2020, pp. 9–14.
9. J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in Proc. EMNLP, 2014, pp. 1532–1543.
10. P. D. Turney, "Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews," in Proc. ACL, 2002, pp. 417–424.
11. Y. Wang, M. Huang, X. Zhu, and L. Zhao, "Attention-based LSTM for aspect-level sentiment classification," in Proc. EMNLP, 2016, pp. 606–615.
12. Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in Proc. NAACL, 2016, pp. 1480–1489.
13. C. Zhou, C. Sun, Z. Liu, and F. Lau, "A C-LSTM neural network for text classification," arXiv:1511.08630, 2015.
14. C. Du, H. Sun, J. Wang, Q. Qi, and J. Liao, "Adversarial and domain-aware BERT for cross-domain sentiment analysis," in Proc. ACL, 2020, pp. 4019–4028.
15. L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," WIREs Data Mining and Knowledge Discovery, vol. 8, no. 4, e1253, 2018.