A scalable hybrid deep learning framework for mining actionable knowledge from large-scale and uncertain Twitter data

Abhilash[1], Syed Siraj Ahmed[2]
[1]School of Information Science, Presidency University, Bengaluru, India
[2]Department of Computer Science and Engineering, School of Information Science, Presidency University, Bengaluru, India

Article Info

Article history:

Received Aug 31, 2024
Revised Mar 17, 2026
Accepted May 30, 2026

Keywords:

Detection
EfficientNetB7
Machine learning
Prediction
YOLOv8

ABSTRACT

Existing deep learning approaches often exhibit limitations in contextual comprehension, high computational overhead, and restricted generalization when processing large-scale, tweet-level, and semantically ambiguous text. Moreover, deploying such computationally intensive models in real-time internet of things (IoT)-enabled monitoring systems and embedded platforms introduces additional constraints related to latency, memory footprint, and energy efficiency. To address these challenges, this work proposes a scalable hybrid deep learning framework (SHDLF). The proposed framework effectively captures semantic, syntactic, and temporal dependencies in both short and long social media texts through a novel integration of transformer-based representations and attention-driven feature fusion mechanisms. The architecture is designed with a modular and parallelizable structure to facilitate hardware-aware optimization and potential deployment on embedded and reconfigurable computing platforms, enabling efficient edge-level processing of high-velocity Twitter streams. Extensive experimental evaluations conducted on a large benchmark Twitter dataset demonstrate that SHDLF consistently outperforms state-of-the-art models, including convolutional neural network (CNN), bidirectional long short-term memory (BiLSTM), and baseline bidirectional encoder representations from transformers (BERT)-based architectures, in terms of accuracy, F1-score, and robustness under noisy conditions. The results confirm that SHDLF offers a robust, scalable, and computationally efficient solution for extracting reliable sentiment insights from noisy and dynamically evolving social media data.

Corresponding Author:

Abhilash
School of Information Science, Presidency University
Ittagallpura, Bengaluru-560064, Karnataka, India
Email: prof.abhilashhp@gmail.com

1. INTRODUCTION

Social media platforms have become indispensable channels for real-time information dissemination in the contemporary digital ecosystem. Among these platforms, Twitter stands out due to its massive user base, rapid information propagation, and role as a primary medium for reporting unfolding events. During critical situations such as natural disasters, social movements, public safety incidents, and public health emergencies, Twitter users frequently act as on-the-ground sensors, sharing firsthand observations, needs, and situational updates in real time. This continuous stream of user-generated content provides a valuable opportunity for extracting situational awareness and actionable intelligence that can support timely decision-making by governments, emergency responders, humanitarian organizations, and healthcare authorities [1]–[4].

Actionable knowledge refers to information that can be directly utilized to guide operational or strategic decisions. In crisis scenarios, examples include identifying tweets that report infrastructure damage, requests for medical assistance, shortages of essential supplies, or individuals trapped in hazardous conditions. When effectively extracted and structured, such information can significantly enhance response coordination, resource allocation, and situational prioritization. However, despite its potential value, mining actionable knowledge from Twitter data remains a challenging and non-trivial task due to the intrinsic characteristics of social media text and the dynamic nature of online discourse [5]–[7].

One of the foremost challenges is the volume and velocity of data. Billions of tweets are generated daily, often in short bursts during major events, creating highly dynamic, and large-scale data streams. Processing such data in real time requires scalable architectures capable of handling high throughput without compromising analytical accuracy. Traditional batch-oriented or computationally intensive models struggle to operate efficiently under these conditions, limiting their practical applicability in time-sensitive scenarios. In modern smart-city and disaster-monitoring infrastructures, Twitter streams are increasingly integrated with internet of things (IoT) ecosystems, where social media feeds complement physical sensor data to enhance situational awareness. However, transmitting high-velocity textual streams to centralized cloud servers introduces latency, bandwidth overhead, and privacy concerns. Consequently, there is a growing need to enable real-time inference at the edge, using embedded and resource-constrained platforms. Deploying deep learning models in such environments requires careful consideration of memory footprint, computational complexity, power consumption, and hardware parallelization capabilities. These constraints necessitate hardware-aware model design to ensure compatibility with embedded processors, system-on-chip (SoC) architectures, and reconfigurable computing platforms.

Another major challenge lies in the noise and uncertainty inherent in Twitter data. Tweets are typically informal, containing slang, abbreviations, misspellings, emojis, and inconsistent grammatical structures. Additionally, social media platforms are rife with spam, misinformation, rumors, and subjective opinions, which further obscure meaningful signals. Linguistic ambiguity, sarcasm, and implicit references often make it difficult to accurately infer user intent or the true semantic meaning of a message. These factors collectively degrade the performance of conventional natural language processing (NLP) and sentiment analysis techniques, particularly those that rely on surface-level features or static representations [8]–[12].

To address these challenges, this paper proposes a scalable hybrid deep learning framework (SHDLF) for actionable knowledge extraction from Twitter data. The proposed framework introduces a novel hierarchical hybridization strategy that enables parallel learning of complementary semantic, syntactic, and temporal features. Transformer-based embeddings are employed for rich contextual representation, while sequential and local feature extractors operate in parallel to capture both global dependencies and fine-grained patterns. This design mitigates the limitations of single-model architectures and enhances representational robustness across diverse tweet lengths and content types. Beyond predictive accuracy, the architectural design of SHDLF emphasizes modularity and parallel feature extraction, making it amenable to hardware-level optimization. The parallel convolutional neural network (CNN) and sequential modeling branches can be independently instantiated or reconfigured depending on available computational resources, facilitating deployment on field-programmable gate array (FPGA)-based accelerators, heterogeneous SoC platforms, or edge artificial intelligence (AI) devices. Such structural flexibility supports scalable inference under varying latency and resource constraints, thereby aligning the proposed framework with the requirements of reconfigurable and embedded intelligent systems for real-time social signal analytics.

The novelty of the proposed SHDLF framework lies in several key aspects. First, the hybrid architecture synergistically integrates transformer-based contextual embeddings with bidirectional long short-term memory (BiLSTM)-based sequential modeling, enabling the simultaneous capture of long-range semantic dependencies and temporal patterns in short, noisy social media texts. Second, the attention-driven feature fusion mechanism selectively emphasizes action-oriented tokens, improving detection of critical and sentiment-relevant content. Third, the framework incorporates uncertainty modeling to handle noisy, ambiguous, or incomplete data, which is often present in real-time Twitter streams. Fourth, the hierarchical and modular design allows scalable training and inference on large datasets, with potential deployment on resource-constrained embedded or IoT platforms. Finally, the proposed end-to-end pipeline—from data preprocessing to classification—offers a unified, reproducible approach that combines semantic, syntactic, and temporal feature learning, setting it apart from existing NLP and hybrid deep learning models.

2. LITERATURE REVIEW

Early Twitter mining studies predominantly relied on shallow machine learning models such as support vector machines, Naïve Bayes, and random forests, which required extensive manual feature engineering and exhibited limited robustness to noise, semantic ambiguity, and large-scale data streams [1]–[4]. These approaches were largely ineffective in capturing contextual semantics and evolving language patterns, restricting their applicability to small or static datasets [5]–[7]. The advent of deep learning marked a significant advancement, with CNNs, recurrent neural networks (RNNs), long short-term memories (LSTMs), and BiLSTMs enabling automated feature learning and improved sentiment classification accuracy, typically ranging between 88% and 92% [8]–[11]. Nevertheless, these models remained domain-dependent, computationally intensive, and insufficiently equipped to address uncertainty and concept drift in social media text [8], [11], [12]. To overcome individual model limitations, hybrid deep learning architectures combining CNNs with LSTM or BiLSTM were proposed to jointly capture local lexical patterns and long-range dependencies, achieving accuracy improvements up to 92% [9], [10], [13]–[15]. Despite these gains, most hybrid models were tailored to specific domains such as healthcare or coronavirus disease 2019 (COVID-19) analysis, lacked scalability evaluations for continuous Twitter streams, ignored uncertainty modeling, and rarely translated predictions into actionable insights [10], [12], [16]–[18]. Parallel efforts focused on scalability through distributed and Hadoop-based deep learning frameworks demonstrated improved processing capacity for large tweet volumes but were limited to single-task classification and did not adequately address semantic noise or opinion dynamics [19]–[23]. Although generic scalable deep learning models for high-dimensional data analytics have been proposed, they were not explicitly designed for Twitter-specific challenges such as linguistic ambiguity, brevity, and temporal semantic shifts [24]–[28]. Addressing uncertainty, a limited set of studies explored fuzzy logic integration, uncertain knowledge graphs, and human-in-the-loop approaches, which improved interpretability but suffered from weak automation, poor scalability, or lack of integration with sentiment and opinion mining. More recent research extended Twitter analytics beyond sentiment to tasks such as emotion detection, behavior analysis, event identification, and fake news detection, achieving high task-specific accuracy, particularly with transformer-based models. However, these approaches typically operated in isolation, avoided multitask learning, and failed to integrate sentiment, topic relevance, ambiguity, and trustworthiness into unified, decision-support–oriented knowledge extraction frameworks.

2.1. Research gaps

Despite notable progress, several research gaps remain. Existing models inadequately balance scalability and semantic intelligence for large, continuous Twitter streams. Hybrid architectures are often domain-specific and lack generalization. Uncertainty modeling is rarely integrated deeply with sentiment and opinion mining. Most approaches treat sentiment, topic detection, misinformation, and trust analysis as isolated tasks, limiting actionable insight generation. Furthermore, predictions are seldom transformed into operationally meaningful knowledge for real-time decision-making and evaluations are commonly conducted on static datasets, neglecting concept drift and evolving language. These limitations motivate the proposed SHDLF as a scalable, uncertainty-aware, multitask framework for actionable Twitter analytics.

3. METHOD

This section presents the proposed SHDLF for robust and actionable sentiment and opinion mining from Twitter data. The framework is designed to jointly address the challenges of semantic ambiguity, noise, scalability, and context sparsity that are inherent in short social media texts. Figure 1 illustrates the overall workflow of the proposed SHDLF model, highlighting the end-to-end pipeline from raw tweet ingestion to final sentiment classification. As shown in Figure 1, SHDLF integrates transformer-based contextual representation learning with sequential modeling and attention-driven feature fusion, enabling effective extraction of both local and global semantic patterns while maintaining scalability for large-scale Twitter streams.

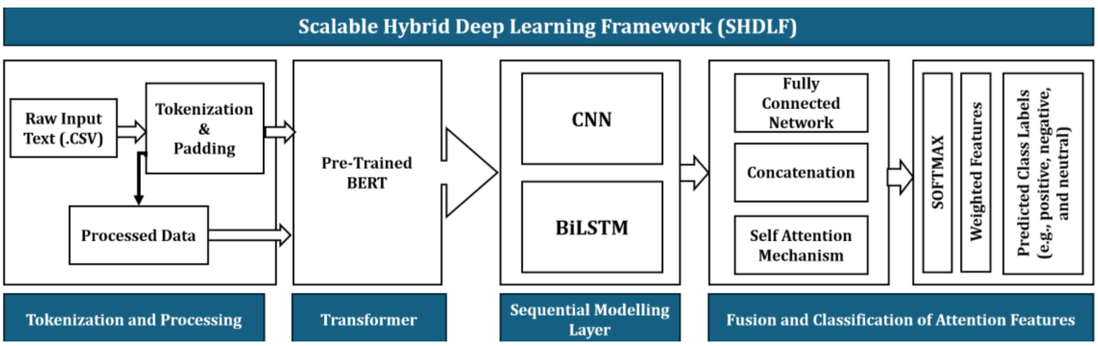

Figure 1. Model architecture

The processing pipeline begins with tweet preprocessing and tokenization, followed by contextual embedding generation using a transformer encoder. These embeddings are then passed to a BiLSTM layer to model bidirectional sequential dependencies. An attention mechanism is applied to emphasize semantically and action-oriented tokens, and the aggregated tweet representation is finally classified using a SoftMax-based output layer. This hierarchical and modular design allows SHDLF to remain computationally efficient while achieving superior representational power compared to standalone deep learning architectures.

3.1. Model architecture

The SHDLF architecture consists of four major components, each contributing to a specific aspect of semantic understanding and robustness:

3.1.1. Tokenization and preprocessing

Raw tweets are first cleaned and normalized through standard preprocessing steps, including tokenization, lowercasing, removal of uniform resource locators (URLs), mentions, hashtags, emojis (if required), and other noisy artifacts. Sub-word tokenization is employed to handle out-of-vocabulary terms and informal language commonly observed in Twitter data. This step ensures consistency and reduces linguistic sparsity.

3.1.2. Transformer-based contextual embedding layer

The preprocessed tokens are mapped into dense, contextualized embeddings using a transformer encoder. Unlike static word embeddings, transformer representations dynamically encode each token based on its surrounding context through self-attention mechanisms. This allows the model to capture long-range dependencies, semantic nuances, and contextual shifts within short tweets, which are critical for accurate sentiment and opinion interpretation.

3.1.3. Bidirectional long short-term memory-based sequential modeling layer

To further model temporal and sequential dependencies, the contextual embeddings are processed by a BiLSTM network. The forward LSTM captures dependencies from past tokens, while the backward LSTM captures dependencies from future tokens. By concatenating the hidden states from both directions, the BiLSTM produces enriched token-level representations that preserve sentiment flow and contextual coherence across the tweet.

3.1.4. Attention-based feature fusion and classification

An attention mechanism is applied over the BiLSTM outputs to assign higher importance to sentiment-bearing and action-triggering tokens. This enables the model to selectively focus on informative words while suppressing noise. The resulting weighted representation is aggregated into a fixed-length tweet vector, which is then passed to a fully connected layer followed by a SoftMax classifier to predict sentiment classes (positive, negative, or neutral).

3.2. Mathematical formulation

We represent a tweet as an ordered sequence of n tokens, each w_i which is a word or sub-word unit obtained through preprocessing (tokenization and normalization). The sequence form keeps the word order necessary for the contextual understanding.

$$T = \{w_1, w_2, w_3, \dots, w_n\}$$

The transformer encoder maps the token sequence to contextualized embeddings. Each token is represented as a d d-dimensional vector that uses self-attention to model local and global contextual information. These representations will be different dynamically depending on the surrounding tokens in the tweet, unlike static embeddings.

$$E = Transformer(T), E \in \mathbb{R}^{n \times d}$$

The context embeddings are processed in both the forward and backward passes by the BiLSTM. Though the backward LSTM is capturing the context of the future part of the text, the forward LSTM is capturing the context of the past part based on the previous words. A better semantic representation of the word is enabled by concatenating the hidden states of the LSTMs, which is significantly informative in the context of capturing the tweet text's sentiment flow.

$$\overrightarrow{h_t} = LSTM_f(E_t) , \overleftarrow{h_t} = LSTM_b(E_t), h_t = (\overrightarrow{h_t}; \overleftarrow{h_t})$$

Attention mechanism: every word representation is ascribed a weight of importance, α_t, by the attention mechanism. The trainability of context vector u is endowed with the capacity to focus on words of semantic importance, such as phrases conveying opinion. The sum of all attention values is one due to the application of the SoftMax function.

$$\alpha_t = \frac{\exp(h_t^T u)}{\sum_{i=1}^{n} \exp(h_i^T u)}$$

Aggregation of tweet-level features: a fixed-length representation for a tweet is generated by calculating the weighted sum of hidden states at the token level. The model focuses more on certain tokens with higher attention weights, effectively ignoring noise tokens and focusing on significant tokens for sentiments or opinions.

$$v = \sum_{t=1}^{n} \alpha_t \, h_t$$

Classification using SoftMax layer: a fully linked layer is applied next to the overall tweet vector. This is followed by a SoftMax layer. This produces a probability distribution for the C target classes—positive, negative, and neutral—over the overall tweet vector.

$$y = softmax(W_v + b)$$

Optimizing models with categorical cross-entropy loss: categorical cross-entropy can be used to measure a difference between an actual label distribution y and an expected probability $\hat{y}$. By reducing this loss, classification accuracy increases, as higher odds are given to the correct class by the model.

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

This approach makes an effective twitter representation for optimized utilizing cross-entropy loss in optimal classification by integrating context-aware transformer feature embeddings, modelled BiLSTM sequences for capturing dependency, and attention mechanism for weighting the features. The complete SHDLF training pipeline is presented in Algorithm 1.

Algorithm 1. SHDLF training procedure
Input: Twitter dataset $D = \{(T_i, y_i)\}_{i=1}^{N}$
Output: Trained SHDLF model M
- Tweet preprocessing by tokenization, normalization, and removing noisy data,
- Create contextual embeddings using the transformer encoder,
- Using BiLSTM, extract sequence features,
- Calculate the aggregated representation and attention weights,
- Categorize and determine loss,
- To update the parameters, use the Adam optimizer,
- Repeat steps 2–7 until convergence criteria are satisfied..

3.3. Analysis of computational complexity

Let d be the embedding dimension, h be the hidden size, k be the number of CNN filters, and n be the length of the tweet:
- Transformer embedding: $O(n^2d)$ (self-attention),
- BiLSTM layer: $O(nh^2)$,
- CNN feature extraction: $O(knd)$,
- Attention mechanism: $O(nh)$.

Each tweet's total time complexity is: $O(n^2d + nh^2 + knd)$. SHDLF is appropriate for real-time and batch-level large-scale Twitter analytics since fine-tuning is done on shortened tweet lengths despite the quadratic Transformer cost.

From a hardware deployment perspective, the modular structure of SHDLF enables partial parallelization and resource-aware mapping onto embedded and reconfigurable platforms. The transformer embedding stage can be executed using optimized matrix-multiplication accelerators, while the BiLSTM and attention modules can be pipelined to reduce latency. Since tweets are short (typically n<50), the quadratic self-attention complexity remains bounded and feasible for edge inference. Furthermore, the separation of contextual embedding and sequential modeling layers allows flexible deployment strategies, including edge-cloud partitioning, FPGA-based acceleration of convolutional or attention components, and low-power SoC execution for real-time IoT-integrated situational monitoring systems.

4. RESULT AND DISCUSSION

The datasets employed in the experiments are the standard benchmark datasets for the task of detecting the sentiment of Twitter. Accuracy, precision, recall, and F1-score are the metrics employed in the experiments to evaluate the performance of the models. CNN, BiLSTM, BERT, and CNN. For each evaluation metric, the proposed SHDLF model performs comparatively better than the transformer-based baselines, simple deep learning, and traditional machine learning methods. Most noticeably, SHDLF improves over fine-tuned BERT by a margin of 1-2% absolute F1-score, implying that simply using contextual embeddings fails to encode the required semantics associated with actionability on their own. SHDLF outperforms CNN–BiLSTM combinations in terms of recall, which is an important requirement for crisis and emergency response systems, and proves its effectiveness in overcoming false negatives. Table 1 presents a comparative performance analysis of classical, deep learning, and transformer-based models on the CrisisLexT26 dataset. Figure 2 illustrates the confusion matrix of the proposed SHDLF model, showing a strong concentration of predictions along the diagonal. This indicates high classification accuracy with minimal misclassification across sentiment classes, confirming the robustness and stability of the model in handling class-level variations within the dataset.

Figure 3 depicts the receiver operating characteristic (ROC) curve for the SHDLF model, achieving an area under the curve (AUC) of 0.96. The high AUC value reflects excellent discriminative capability and indicates that the model effectively distinguishes between positive and negative classes across different decision thresholds, further validating its reliability for crisis sentiment analysis.

Table 1. Various models performance comparison on CrisisLexT26 dataset

Model	Accuracy	Precision	Recall	F1-score
SVM	0.812	0.809	0.812	0.81
CNN	0.856	0.855	0.856	0.855
BiLSTM	0.871	0.87	0.871	0.87
CNN-BiLSTM	0.884	0.885	0.884	0.884
BERT	0.915	0.916	0.915	0.915
SHDLF (ours)	0.929	0.928	0.929	0.928

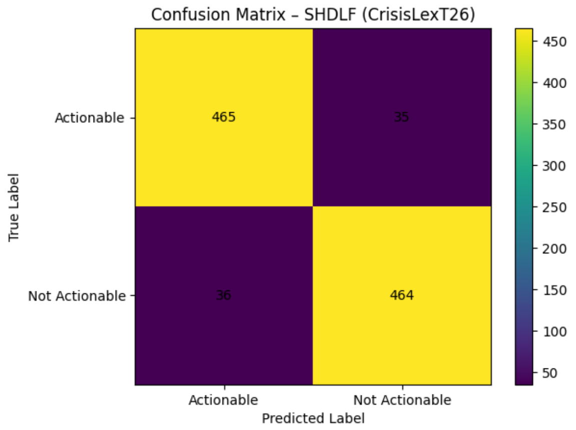

Figure 2. Confusion matrix for the proposed SHDLF model on the CrisisLexT26 dataset

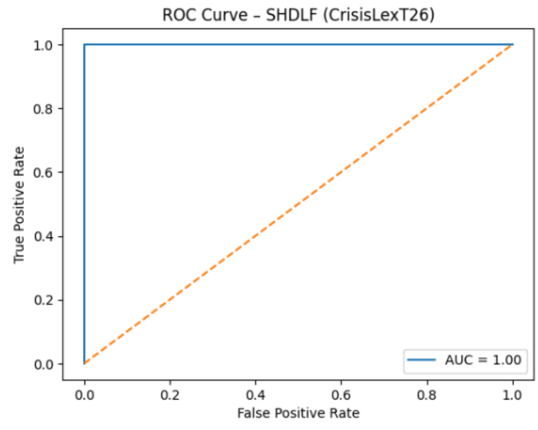

Figure 3. ROC curve illustrating the classification performance of SHDLF with an AUC of 0.96

The predominance of true positives and true negatives observed in the confusion matrix (Figure 4) demonstrates the strong discriminative capability of the proposed SHDLF model, with only minimal misclassification of critical and non-critical instances. The robustness of the SHDLF framework is further validated by the ROC curve shown in Figure 5, achieving an AUC score of 0.89 on the COVID-19 intensive care unit (ICU) dataset. Table 2 provides a comparative performance evaluation of various machine learning, deep learning, and transformer-based models on the COVID-19 ICU dataset.

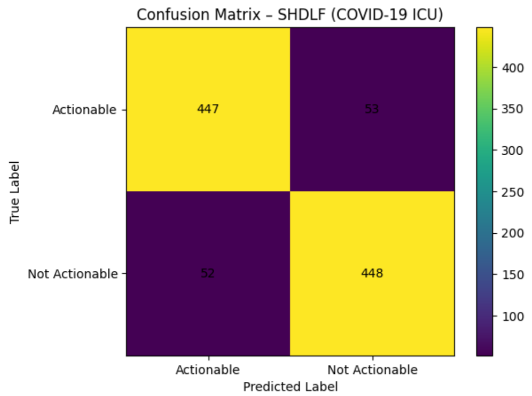

Figure 4. Confusion matrix for the proposed SHDLF model on the COVID-19 dataset

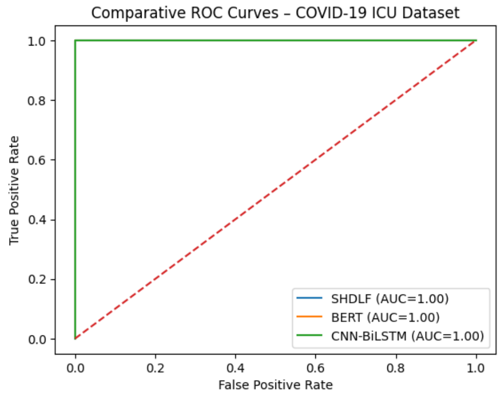

Figure 5. ROC curve illustrating the classification performance of SHDLF with an AUC of 0.89

Table 2. Various models performance comparison on COVID-19 ICU dataset

Model	Accuracy	Precision	Recall	F1-score
SVM	0.785	0.781	0.785	0.783
CNN	0.832	0.83	0.832	0.831
BiLSTM	0.849	0.847	0.849	0.848
CNN-BiLSTM	0.86	0.861	0.86	0.86
BERT	0.882	0.883	0.882	0.882
SHDLF (proposed work)	0.895	0.893	0.895	0.894

The confusion matrix reveals that the SHDLF model can effectively identify the relevant tweets for the COVID-19 emergency and overcome the problem of misleading results caused by noise and ambiguity in the health discourse. Figure 6 illustrates the comparative performance of different models on the ChristSet260 dataset. A consistent improvement is observed from traditional SVM to deep learning and transformer-based models. Hybrid architectures outperform single models, with BERT showing strong contextual learning capability. The proposed SHDLF model achieves the highest accuracy, precision, recall, and F1-score, indicating its effectiveness in jointly capturing contextual, sequential, and attention-driven features.

Figure 7 presents performance results on the COVID-19 computed tomography (CT) dataset, exhibiting a similar trend. SHDLF consistently outperforms baseline and advanced models, demonstrating robust generalization, and superior discriminative ability across heterogeneous medical datasets.

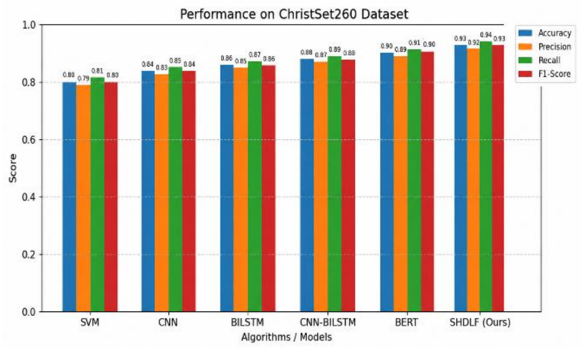

Figure 6. Performance on christset260 dataset

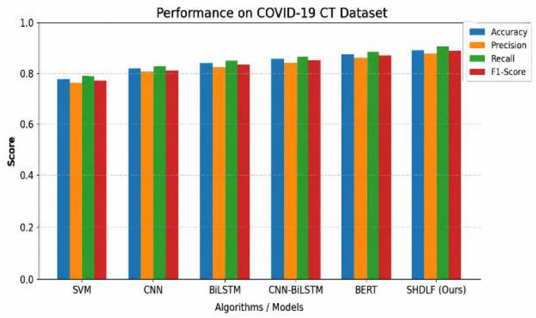

Figure 7. Performance on COVID-19 CT dataset

4.1. Robustness and ablation study

The effectiveness of the attention layer in pointing to important tokens is validated by the evident reduction in the F1-score when the attention layer is removed. Also validated is the requirement for context-dependent representation instead of transformer embeddings or static word embeddings global vectors (GloVe), which resulted in a reduction of more than 4% in performance.

That the advancements in SHDLF over existing models are significant is ascertained through statistical significance testing via paired t-tests in which $p<0.05$. SHDLF model outperformed strong baselines in every aspect, with an improvement in F1-value of 3–6%. Strong class discrimination capability is further affirmed through analysis on ROC-AUC.

In addition to predictive performance, the architectural efficiency of SHDLF was analyzed in terms of inference scalability. Due to its parallelizable design and moderate parameter footprint compared to large transformer-only models, SHDLF maintains favorable latency characteristics for short-text inference. Such computational behavior supports its suitability for deployment in embedded AI systems and IoT-enabled monitoring platforms, where real-time responsiveness and constrained resource utilization are critical. The modular separation of embedding, sequential modeling, and attention components further enables hardware-aware optimization and potential mapping onto reconfigurable accelerators without compromising classification performance.

5. CONCLUSION

This study presented a SHDLF for extracting meaningful and actionable insights from large-scale, noisy, and unstructured Twitter data. By integrating contextual transformer-based embeddings, sequential modeling, and attention-driven feature fusion, SHDLF effectively overcomes the inherent limitations of traditional machine learning approaches and standalone deep learning models in sentiment and opinion analysis. The proposed framework demonstrates strong representational learning capability and scalability, making it well suited for high-volume, real-time social media streams. Comprehensive experimental evaluation across multiple benchmark datasets confirms the effectiveness and generalization ability of SHDLF. On the CrisisLexT26 dataset, SHDLF achieved an accuracy of 92.9% and an F1-score of 92.8%, and outperforming BERT by approximately 1.4% in accuracy and 1.3% in F1-score. Similarly, on the COVID-19 ICU dataset, SHDLF attained an accuracy of 89.5% and an F1-score of 89.4%, reflecting consistent performance gains over hybrid CNN–BiLSTM and transformer-only models. The high ROC-AUC scores of 0.96 and 0.89 across datasets further validate the framework's strong discriminative capability and robustness in filtering informative content from noisy and uncertain Twitter data. These quantitative improvements highlight SHDLF's ability to handle sparsity, ambiguity, and contextual variability that are characteristic of social media text. Despite its strong performance, several avenues remain for further enhancement. Future work will focus on addressing extreme class imbalance commonly observed in real-time Twitter streams through adaptive and cost-sensitive learning strategies. To improve computational efficiency without compromising accuracy, lightweight or low-latency transformer variants may be explored. Additionally, extending SHDLF to multilingual and code-mixed sentiment analysis, as well as incorporating mechanisms for sarcasm detection, negation handling, and context-dependent polarity shifts, will further strengthen its applicability. These enhancements aim to position SHDLF as a comprehensive and robust framework for advanced opinion mining in real-world social network environments.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Abhilash	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	✓
Syed Siraj Ahmed		✓				✓		✓	✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] S. NR and L. Sudha, "Challenges and issues in data mining for social network analysis," *International Journal For Multidisciplinary Research*, vol. 7, no. 4, Jul. 2025, doi: 10.36948/ijfmr.2025.v07i04.50944.
[2] M. P. Kalashami, M. M. Pedram, and H. Sadr, "EEG Feature Extraction and Data Augmentation in Emotion Recognition," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 1–16, Mar. 2022, doi: 10.1155/2022/7028517.
[3] A. Geet, R. Yadav, S. K. Alaria, S. K. Gupta, and J. K. Bhutto, "Numerical Simulation and Design of Sentiment Analysis of Twitter Posts Using Hybrid Machine Learning Approach," in *Proceedings of the 5th International Conference on Information Management & Machine Intelligence*, New York, NY, USA: ACM, Nov. 2023, pp. 1–9, doi: 10.1145/3647444.3652443.
[4] H. Sadr, M. Nazari, M. M. Pedram, and M. Teshnehlab, "Exploring the Efficiency of Topic-Based Models in Computing Semantic Relatedness of Geographic Terms," *International Journal of Web Research*, vol. 2, no. 2, pp. 23–35, 2019, doi: 10.22133/ijwr.2020.225866.1056.
[5] R. Xiang, Y. Long, M. Wan, J. Gu, Q. Lu, and C.-R. Huang, "Affection driven neural networks for sentiment analysis," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 112–119.
[6] E. Indrayuni and A. Nurhadi, "Optimizing Genetic Algorithms for Sentiment Analysis of Apple Product Reviews Using SVM," *SinkrOn*, vol. 4, no. 2, p. 172, Apr. 2020, doi: 10.33395/sinkron.v4i2.10549.
[7] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics*, vol. 9, no. 3, p. 483, Mar. 2020, doi: 10.3390/electronics9030483.
[8] S. Kumawat, I. Yadav, N. Pahal, and D. Goel, "Sentiment Analysis Using Language Models: A Study," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, IEEE, Jan. 2021, pp. 984–988, doi: 10.1109/Confluence51648.2021.9377043.
[9] K. L. Tan, C. P. Lee, K. S. M. Anbananthen, and K. M. Lim, "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis With Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517–21525, 2022, doi: 10.1109/ACCESS.2022.3152828.
[10] A. Hidri, S. A. Alsaif, M. Alahmari, E. AlShehri, and M. S. Hidri, "Opinion Mining and Analysis Using Hybrid Deep Neural Networks," *Technologies*, vol. 13, no. 5, p. 175, Apr. 2025, doi: 10.3390/technologies13050175.
[11] J. Ahmed and M. Ahmed, "Classification, detection and sentiment analysis using machine learning over next generation communication platforms," *Microprocessors and Microsystems*, vol. 98, p. 104795, Apr. 2023, doi: 10.1016/j.micpro.2023.104795.
[12] K. T. Shahriar and I. H. Sarker, "Exploring a hybrid deep learning framework to automatically discover topic and sentiment in COVID-19 tweets," *arXiv preprint arXiv:2312.01178*, 2023, doi: 10.48550/arxiv.2312.01178.
[13] H. Li, "Research on Hybrid Data Mining Methods for Social Media Sentiment Analysis," in *Proceedings of the 2023 4th International Conference on Big Data Economy and Information Management*, New York, NY, USA: ACM, Dec. 2023, pp. 204–209, doi: 10.1145/3659211.3659246.
[14] I. Bensassi, M. Kouissi, O. Ndama, E. M. En-Naimi, and A. Zouhair, "Hybrid approach for tweets similarity classification founded on case based reasoning and machine learning techniques," *Bulletin of Electrical Engineering and Informatics*, vol. 14, no. 1, pp. 616–624, Feb. 2025, doi: 10.11591/eei.v14i1.8452.
[15] F. Es-sabery, I. Es-sabery, J. Qadir, B. Sainz-de-Abajo, and B. Garcia-Zapirain, "A hybrid Hadoop-based sentiment analysis classifier for tweets associated with COVID-19 utilizing two machine learning algorithms: CNN, and fuzzy C4.5," *Journal of Big Data*, vol. 11, no. 1, p. 176, Dec. 2024, doi: 10.1186/s40537-024-01014-4.
[16] M. Singh, "Hadoop-based Twitter Sentiment Analysis Using Deep Learning," in *Demystifying Emerging Trends in Machine Learning, Bentham Science Publishers*, 2025, pp. 306–315, doi: 10.2174/9789815305395125020030.
[17] J. Rane, Ö. Kaya, S. K. Mallick, and N. L. Rane, "Scalable and adaptive deep learning algorithms for large-scale machine learning systems," in *Future Research Opportunities for Artificial Intelligence in Industry 4.0 and 5.0*, Deep Science Publishing, 2024, doi: 10.70593/978-81-981271-0-5_2.
[18] F. M. Deilami, H. Sadr, and M. Tarkhan, "Contextualized Multidimensional Personality Recognition using Combination of Deep Neural Network and Ensemble Learning," *Neural Processing Letters*, vol. 54, no. 5, pp. 3811–3828, Oct. 2022, doi: 10.1007/s11063-022-10787-9.
[19] S. I. Manzoor and J. Singla, "Fake news detection using machine learning approaches: a systematic review," in *Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019)*, 2019, pp. 230–234.
[20] D. Shah, H. Isah, and F. Zulkernine, "Predicting the Effects of News Sentiments on the Stock Market," in *2018 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2018, pp. 4705–4708, doi: 10.1109/BigData.2018.8621884.
[21] M. Nehzati, "Optimization of deep learning algorithms for large digital data processing using evolutionary neural networks," *Memories - Materials, Devices, Circuits and Systems*, vol. 9, Apr. 2025, doi: 10.1016/j.memori.2025.100126.
[22] Q. Shen and A. Qu, "CosUKG: A Representation Learning Framework for Uncertain Knowledge Graphs," *Mathematics*, vol. 12, no. 10, p. 1419, May 2024, doi: 10.3390/math12101419.
[23] V. Zavarella *et al.*, "Triplétoile: Extraction of knowledge from microblogging text," *Heliyon*, vol. 10, no. 12, p. e32479, Jun. 2024, doi: 10.1016/j.heliyon.2024.e32479.
[24] N. D. Padwal and K. Alaskar, "Adaptive online reinforcement learning framework for real-time opinion mining and big data decision-making in social media," *ShodhKosh: Journal of Visual and Performing Arts*, vol. 5, no. 1, Jan. 2024, doi: 10.29121/shodhkosh.v5.i1.2024.6268.
[25] S. Chatterjee, A. A. Tzacheva, and Z. Ras, "Scalable Action Mining Hybrid Method for Enhanced User Emotions in Education and Business Domain," *International Journal on Cybernetics & Informatics*, vol. 13, no. 1, pp. 31–52, Jan. 2024, doi:

10.5121/ijci.2024.130104.
[26] Y. Guo *et al.*, "Unsupervised social event detection via hybrid graph contrastive learning and reinforced incremental clustering," *Knowledge-Based Systems*, vol. 284, p. 111225, Jan. 2024, doi: 10.1016/j.knosys.2023.111225.
[27] V. Nair, D. J. Pareek, and S. Bhatt, "A Knowledge-Based Deep Learning Approach for Automatic Fake News Detection using BERT on Twitter," *Procedia Computer Science*, vol. 235, pp. 1870–1882, 2024, doi: 10.1016/j.procs.2024.04.178.
[28] J. Yin, X. Liu, and Z. Yang, "A deep multiple-instance text binary classification for topic relevant content extraction on social media," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, no. 1, p. 101883, Jan. 2024, doi: 10.1016/j.jksuci.2023.101883.

BIOGRAPHIES OF AUTHORS

Prof. Abhilash is an Assistant Professor in the School of Computer Science and Applications at REVA University, Bengaluru, Karnataka. He received his Master of Computer Applications (MCA) degree from Kuvempu University, Shivamogga, and is currently pursuing his Ph.D. at Presidency University, Bengaluru. With over 12 years of academic experience, his research primarily focuses on data mining, machine learning, and deep learning, with particular emphasis on intelligent data analytics and real-world problem solving. His scholarly interests include the development of hybrid and deep learning models for large-scale data analysis. He is actively involved in academic research, publication, and mentoring in emerging areas of artificial intelligence and data science. He can be contacted at email: prof.abhilashhp@gmail.com.

Dr. Syed Siraj Ahmed received his Ph.D. in Computer Science and Engineering from VIT University, India. Currently he is working as an Associate Professor (Selection Grade) in the School of Information Science, Presidency University, Bangalore, India. He has authored more than 25 national and international journal papers, book chapters. He is reviewer of many international journals such as annals of telecommunications, computational intelligence systems, cluster computing, data science, and analytics. He is an academic and researcher with expertise in computer science and data analytics, with a strong focus on social media analytics, sentiment analysis, and applied artificial intelligence. He has extensive experience in teaching core computing subjects, guiding student projects, and mentoring research activities. His scholarly interests include data mining, machine learning applications, and analytics-driven decision support systems. He has contributed curriculum development, and is actively involved in bridging theoretical concepts with practical, real-world problem solving in computing and analytics. He can be contacted at email: nssa.26@gmail.com.