The International Innovations Journal of Applied Science (IIJAS) Vol. 3 No.2, 15-09-2026
DOI: https://doi.org/10.61856/s64z9e54

The International Innovations Journal of Applied Science

Journal homepage: https://iijas.eventsgate.org/iijas

ISSN: 3009-1853 (Online)

Systematic Review of Deep Learning Models in Sentiment Analysis (2020-2025)

Shaima Mahdi Orebi [1*]

[1] Information Technology Department ,University of Thi- Qar, Nasiriyah, Iraq

ARTICLE INFO

Article history:
Received 11Apr. 2026,
Revised 13 Aug. 2026,
Accepted 15 Aug. 2026,
Available online 15 Sep. 2026.

Keywords:
Sentiment Analysis
Natural Language Processing
Deep Learning
Long Short-Term Memory
Bidirectional Long Short-Term Memory
Transformer

ABSTRACT

This review examines deep learning approaches to sentiment analysis published during 2020–2025, with an emphasis on recurrent and hybrid architectures. Twenty study entries are summarized by model architecture, dataset, application task, and reported performance. The selected evidence includes LSTM, BiLSTM, convolutional–recurrent models, and architectures incorporating BERT or RoBERTa. Reported results vary substantially across datasets and tasks, including sentiment polarity classification, emotion classification, aspect-based analysis, and recommendation applications. Because evaluation protocols and outcome measures differ, numerical scores are interpreted descriptively rather than pooled or used to establish universal model superiority. The synthesis identifies a need for clearer reporting of evaluation conditions and further assessment of domain transfer, linguistic variation, and model robustness.

1. Introduction

Sentiment analysis examines opinions expressed in text and commonly classifies their polarity as positive, negative, or neutral. It supports applications such as product feedback analysis and public opinion monitoring (Devika et al., 2016; Wankhade et al., 2022).

Social media provides large volumes of user-generated opinions. Twitter sentiment analysis is challenging because posts may contain informal language, abbreviations, sarcasm, and limited context. Deep learning offers methods for learning useful textual representations from these data (Etaiwi et al., 2021).

Opinion mining can be applied to reviews, social media posts, and other textual feedback. Its interpretation depends on the task, the annotation scheme, and the context in which the text was produced (Liu, 2012).

Sentiment polarity classification and emotion classification are related but distinct tasks. Polarity classification identifies evaluative orientation, whereas emotion classification assigns categories such as anger, sadness, or joy. Their results should therefore be distinguished when synthesizing evidence (Mao et al., 2024).

Many machine learning algorithm models have been used in sentiment analysis. In this study, we highlight the use of neural networks in sentiment analysis and their ability to detect polarity in data.

Artificial neural networks learn representations through interconnected computational units. Deep learning uses multiple processing layers to learn increasingly complex representations; neural networks and

* Corresponding author.
E-mail address: shiama_mahdi@utq.edu.iq

deep learning are therefore not interchangeable terms (Etaiwi et al., 2021; Guo et al., 2016).

Deep learning architectures used in text analysis include convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and bidirectional LSTM (BiLSTM) networks (Etaiwi et al., 2021; Mehta & Pandya, 2020).

This review examines deep learning approaches to sentiment analysis published during 2020–2025, with particular attention to recurrent and hybrid architectures. Traditional classifiers such as support vector machines and naive Bayes commonly use explicitly engineered textual features. Neural models can instead learn representations from data, although their performance depends on training data, model configuration, and evaluation design.

Deep learning also has applications outside text analysis, including medical image analysis (Wang et al., 2025). Evidence from those applications provides background context rather than direct evidence of sentiment classification performance.

For sentiment analysis, the value of a model must be assessed using an evaluation design appropriate to the intended language, domain, and prediction task.

Recent sentiment analysis research examines recurrent models, contextual representations, and combinations of multiple architectures (Dang et al., 2021; Khan & Srivastava, 2024). Such combinations motivate a review that distinguishes model design from differences in datasets and evaluation protocols.

RNNs process sequential inputs while maintaining a hidden state that carries information from earlier positions. LSTM and BiLSTM are recurrent architectures relevant to the studies examined here (Etaiwi et al., 2021; Priyadarshini & Cotton, 2021).

LSTM networks use gating mechanisms to retain and update information across a sequence. They were developed to address difficulties in learning long-range dependencies with conventional RNNs, including vanishing gradients (Mehta & Pandya, 2020; Priyadarshini & Cotton, 2021).

BiLSTM networks process an input sequence in both forward and backward directions and combine the resulting representations. This allows a token representation to incorporate preceding and following context when the complete sequence is available (Etaiwi et al., 2021).

Existing reviews address sentiment analysis methods, applications, and challenges, including developments in deep learning (Alahmadi et al., 2025; Etaiwi et al., 2021; Mao et al., 2024; Wankhade et al., 2022). This review focuses on a selected set of recurrent, transformer-based, and hybrid studies from 2020–2025. Its contribution is a task-aware comparison of the architectures, datasets, and reported evaluation measures represented in that set.

The review addresses two questions: (1) Which recurrent, transformer-based, and hybrid architectures are represented in the selected studies published during 2020–2025? (2) How do task definitions, datasets, and evaluation measures affect the interpretation of their reported performance?

2. Methodology

The review considered studies of deep learning for sentiment analysis, with an emphasis on LSTM, BiLSTM, and related hybrid architectures. The reported search sources were Google Scholar, IEEE Xplore, Scopus, ScienceDirect, and Wiley Online Library.

3.1 Research strategy:

The search concepts included sentiment analysis, opinion mining, text mining, sentiment classification, deep learning, RNN, LSTM, BiLSTM, natural language processing, and word embedding. The topic combinations are listed below.

- "sentiment analysis" AND "deep learning"
- "sentiment classification" AND (LSTM OR BiLSTM)
- "text classification" AND "word embedding"

DOI: https://doi.org/10.61856/s64z9e54

- "deep learning" AND "natural language processing"
- "opinion mining" AND (LSTM OR BiLSTM)
- LSTM AND Twitter
- BiLSTM AND "social media"

The stated inclusion criteria were English-language conference papers and peer-reviewed journal articles published during 2020–2025, inclusive, that investigated deep learning for sentiment analysis and reported experimental evaluation. This interval covers six calendar years.

Studies limited to traditional machine learning without a deep learning component, or without clear experimental results and performance measures, were excluded under the stated criteria. Screening comprised title and abstract review followed by full-text assessment. Table 1 contains 20 study entries, summarized by architecture, dataset, and reported evaluation results.

Study selection was organized around identification, screening, and full-text eligibility assessment. Figure 1 presents the study-selection flow. Sentiment polarity, emotion classification, aspect-based analysis, and recommendation-related applications are distinguished in interpreting the evidence because they do not define equivalent prediction tasks.

3.2 Characteristics of the Research:

The synthesis compares architectures, application tasks, datasets, and evaluation measures. Particular attention is given to whether apparently different performance results can be explained by differences in task definition or evaluation conditions.

3.3 Evidence interpretation

The synthesis is descriptive and does not assign numerical quality scores to the included studies. Differences in methodological reporting and evaluation conditions limit confidence in direct comparisons of reported performance.

3.4 The Studies Selection Process:

The selection process comprised record collection, duplicate removal, title and abstract screening, and full-text eligibility assessment.

3.5 Data Collection:

Data were summarized qualitatively by publication year, model architecture, application task, dataset, and reported performance. Accuracy and F1 scores were not pooled. Comparisons are descriptive because the studies use different datasets, class definitions, and evaluation protocols.

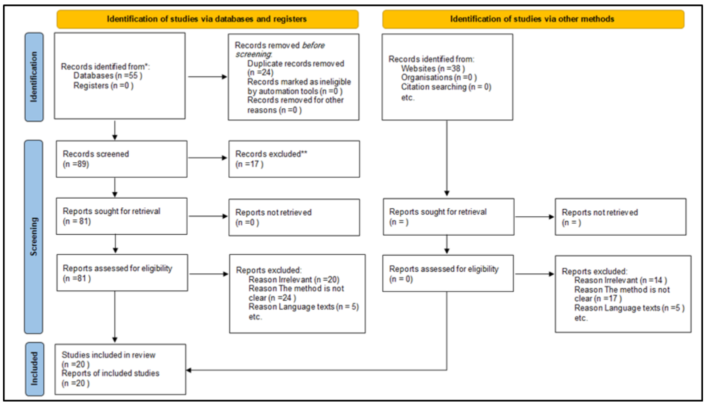

Figure 1. Flowchart of the study selection process according to the PRISMA methodology

Table 1. Characteristics and reported performance of the selected studies

Source	Year	Method	Reported performance	Dataset
(Murthy et al., 2020)	2020	Sentiment analysis on text reviews by using Long Short-Term Memory (LSTM)	99.56%	IMDB & Amazon
(Srinivas et al., 2021)	2021	Simple Neural Network, Long Short-Term Memory (LSTM), and Convolutional Neural Network (CNN) methods	87%	Analysed 1.6 million Twitter sentiment dataset is collected from a Kaggle repositories
(Mahto et al., 2022)	2022	ConvBidirectional-LSTM	93.25%	US airline dataset
(Xiaoyan et al., 2022)	2022	GloVe-CNN-BiLSTM	0.9565; 0.9509; 0.9560	Twitter's comment data on COVID-19
(Tan et al., 2022)	2022	Integrates Robustly optimized BERT approach and Long Short-Term Memory	93%, 91%, and 90%	IMDb dataset, Twitter US Airline Sentiment dataset, and Sentiment140 dataset
(Zhang et al., 2023)	2023	Bidirectional LSTM (Bi-LSTM) sentiment analysis method	77%	open-source comment sentiment140 data set
(Mahadevaswamy & Swathi, 2023)	2023	Sentiment Analysis using Bidirectional LSTM Network	91.4%	Amazon product review dataset
(Glenn et al., 2023)	2023	Single-layer bidirectional long short-term memory model over a two-layer stacked bidirectional long short-term memory mode	70.83%	4401 Indonesian tweets labeled into five emotion classes
(Shlkamy, 2023)	2023	Sentiment Analysis Using Bidirectional LSTM Network	91.79%	A Russia-Ukraine Conflict Tweets
(Greeshma & Simon, 2024)	2024	Bidirectional GRU layer	98%	IMDb dataset.
(Xu et al., 2024)	2024	Text Sentiment Analysis and Classification BasedonBidirectional Gated Recurrent Units (GRUs) Model	94.8%	461810 text messages and their corresponding label classifications
(Orebi & Naser, 2025)	2025	Word embedding Global Vectors for Word	89% and 90%	ChatGPT tweets from the first month after

DOI: https://doi.org/10.61856/s64z9e54

Source	Year	Method	Reported performance	Dataset
		Representation (GLOVE) with Bidirectional LSTM and Long Short-Term Memory (LSTM)		launch
(Varadharajan et al., 2025)	2025	Long Short-Term Memory	80%.	IMDb Review
(Lu et al., 2025)	2025	LSTM combined with the Word2vec	92%	IMDB dataset
(Rahman et al., 2025)	2025	RoBERTa–BiLSTM hybrid	92.35%	IMDb, Twitter US Airline, and Sentiment140
(S & Kabadi, 2025)	2025	combines the Robustly Optimized BERT Pretraining Approach (RoBERTa) with Bidirectional Long Short-Term Memory (BiLSTM) networks.	80.73%, 92.35%, and 82.25%	Twitter US Airline, IMDb, and Sentiment140 datasets,
(Li et al., 2025)	2025	K-means + BERT + MLP	0.832	real-world e-commerce
(Taj et al., 2025)	2025	Fine-tuned BERT with LIME explanations for aspect-based requirements elicitation	F1: 0.83 (ACD); 0.94 (ACP) And 0.94	AWARE app reviews
(Talukder et al., 2025)	2025	BERT +(LSTM)	92.10%, 92.10%, and 91.50%	COVID-19-related sentiments
(Bouassida & Mezali, 2025)	2025	BERT-BiLSTM, Roberta-CNN-BiLSTM, and DistilBERT-BiLSTM.	Accuracy: 79%, 77%, and 81%, respectively	Sentiment140 dataset

3. Results and Discussion

Table 1 summarizes 20 study entries published during 2020–2025. The evidence includes sentiment polarity classification, emotion classification, aspect-based analysis, and a recommendation application incorporating sentiment features. These tasks are considered separately when interpreting reported scores.

The selected studies include recurrent architectures, convolutional–recurrent combinations, and transformer–recurrent hybrids. Recurrent approaches remain represented throughout the period, while several later entries combine BERT or RoBERTa with sequence models. This pattern describes the selected sample and does not establish a field-wide replacement of recurrent models.

The table includes both review datasets, such as IMDb and Amazon, and social media datasets, such as Sentiment140 and Twitter US Airline. Differences between their reported scores cannot be attributed to text length or domain alone because model settings, class definitions, data splits, and preprocessing also vary.

The percentage-labelled performance values in Table 1 range from 70.83% to 99.56%. This range spans heterogeneous tasks and is not a pooled estimate or a ranking of model families. Taj et al. (2025), for example, report F1 scores for aspect category detection and aspect category polarity, which are distinct from classification accuracy.

Transformer-based and hybrid models report promising results in several selected studies, but the table does not establish their consistent superiority over all recurrent alternatives. Such a conclusion would require matched datasets, comparable splits, consistent metrics, and uncertainty estimates.

The included evaluations do not collectively establish robustness to sarcasm, irony, dialect variation, or domain shift. Higher benchmark scores alone should therefore not be interpreted as evidence of deeper linguistic understanding or reliable deployment performance.

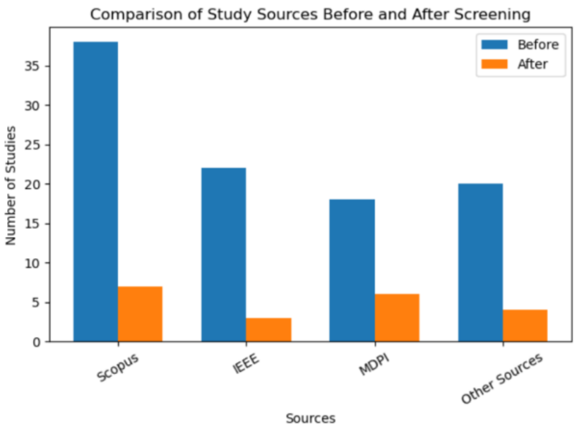

Figure 2. Reported study counts by source before and after screening

Interpretation of the research questions

For the first research question, the selected evidence demonstrates continued use of recurrent architectures alongside transformer-based and hybrid approaches. Because the search emphasizes LSTM and BiLSTM and the sample is small, the architecture distribution should not be generalized to all sentiment analysis research.

For the second research question, task and evaluation heterogeneity are central. A five-class emotion task, binary sentiment classification, aspect category detection, and recommendation prediction assess different capabilities. Their numerical scores cannot serve as interchangeable evidence of sentiment classification quality.

The absence of harmonized reporting of data splits, class balance, preprocessing, and uncertainty limits direct comparisons. Computational cost is also not reported consistently enough in Table 1 to support quantitative claims about efficiency.

The review has limitations arising from the restricted language and publication window, its emphasis on recurrent architectures, and the heterogeneous tasks represented in the selected sample. These constraints limit coverage and prevent a pooled estimate of model effectiveness.

Future research should report reproducible evaluation protocols, model and data configurations, uncertainty measures, and resource requirements. Evaluation across domains, languages, dialects, and challenging linguistic phenomena would provide a stronger basis for assessing robustness and practical usefulness.

4. Conclusion

The selected studies illustrate a range of recurrent, transformer-based, and hybrid approaches to sentiment-related applications during 2020–2025. Their reported results vary across datasets, tasks, and evaluation protocols. The synthesis supports task-specific interpretation of performance rather than a universal ranking of architectures. Reliable conclusions require accurate source extraction and comparable evaluations.

Large language models and multimodal sentiment analysis are possible directions for future investigation. Their effectiveness is not established by the present study set and should be evaluated separately using explicit task definitions and reproducible benchmarks.

References

Alahmadi, K., Alharbi, S., Chen, J., & Wang, X. (2025). Generalizing sentiment analysis: A review of progress, challenges, and emerging directions. *Social Network Analysis and Mining, 15*(1), Article 45. https://doi.org/10.1007/s13278-025-01461-8

Bouassida, Y., & Mezali, H. (2025). Enhancing Twitter sentiment analysis using hybrid transformer and sequence models. *Japan Journal of Research, 6*(1), Article 089. https://doi.org/10.33425/2690-8077.1163

Dang, C. N., Moreno-García, M. N., & De La Prieta, F. (2021). Hybrid deep learning models for sentiment analysis. *Complexity, 2021*, Article 9986920. https://doi.org/10.1155/2021/9986920

Devika, M. D., Sunitha, C., & Ganesh, A. (2016). Sentiment analysis: A comparative study on different approaches. *Procedia Computer Science, 87*, 44–49. https://doi.org/10.1016/j.procs.2016.05.124

Etaiwi, W., Suleiman, D., & Awajan, A. (2021). Deep learning based techniques for sentiment analysis: A survey. *Informatica, 45*(7). https://doi.org/10.31449/inf.v45i7.3674

Glenn, A., LaCasse, P., & Cox, B. (2023). Emotion classification of Indonesian tweets using bidirectional LSTM. *Neural Computing and Applications, 35*(13), 9567–9578. https://doi.org/10.1007/s00521-022-08186-1

Greeshma, M., & Simon, P. (2024). Bidirectional gated recurrent unit with GloVe embedding and attention mechanism for movie review classification. *Procedia Computer Science, 233*, 528–536. https://doi.org/10.1016/j.procs.2024.03.242

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing, 187*, 27–48. https://doi.org/10.1016/j.neucom.2015.09.116

Khan, M., & Srivastava, A. (2024). *A review on sentiment analysis of Twitter data using machine learning techniques*. Zenodo. https://doi.org/10.5281/ZENODO.10791471

Li, S., Liu, K., & Chen, X. (2025). A context-aware personalized recommendation framework integrating user clustering and BERT-based sentiment analysis. *Journal of Computer, Signal, and System Research, 2*(6), 100–108. https://doi.org/10.71222/1cgq9333

Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool Publishers. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Lu, J., Fan, H., & Zhang, Y. (2025). Sentiment analysis of IMDb movie reviews based on LSTM. *Journal of Advanced Engineering and Technology, 2*(2). https://doi.org/10.62177/jaet.v2i2.429

Mahadevaswamy, U. B., & Swathi, P. (2023). Sentiment analysis using bidirectional LSTM network. *Procedia Computer Science, 218*, 45–56. https://doi.org/10.1016/j.procs.2022.12.400

Mahto, D., Yadav, S. C., & Lalotra, G. S. (2022). Sentiment prediction of textual data using hybrid convbidirectional-LSTM model. *Mobile Information Systems, 2022*, Article 1068554. https://doi.org/10.1155/2022/1068554

Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University - Computer and Information Sciences, 36*(4), Article 102048. https://doi.org/10.1016/j.jksuci.2024.102048

Mehta, P., & Pandya, S. (2020). A review on sentiment analysis methodologies, practices and applications. *International Journal of Scientific and Technology Research, 9*(2), 601–609.

Murthy, G. S. N., Allu, S. R., Andhavarapu, B., Bagadi, M., & Belusonti, M. (2020). Text based sentiment analysis using LSTM. *International Journal of Engineering Research & Technology, 9*(5), Article IJERTV9IS050290. https://doi.org/10.17577/IJERTV9IS050290

Orebi, S. M., & Naser, A. M. (2025). Opinion mining in text short by using word embedding and deep learning. *Journal of Applied Data Sciences, 6*(1), 526–636. https://doi.org/10.47738/jads.v6i1.438

Priyadarshini, I., & Cotton, C. (2021). A novel LSTM–CNN–grid search-based deep neural network for sentiment analysis. *The Journal of Supercomputing, 77*(12), 13911–13932. https://doi.org/10.1007/s11227-021-03838-w

Rahman, M. M., Shiplu, A. I., Watanobe, Y., & Alam, M. A. (2025). RoBERTa-BiLSTM: A context-aware hybrid model for sentiment analysis. *IEEE Transactions on Emerging Topics in Computational Intelligence*. Advance online publication. https://doi.org/10.1109/TETCI.2025.3572150

S, N., & Kabadi, M. (2025). A sentiment analysis framework using RoBERTa with bi-LSTM classification on dynamic Twitter data. *ETRI Journal*. Advance online publication. https://doi.org/10.4218/etrij.2024-0217

Shlkamy, E. S. I. (2023). A Russia-Ukraine conflict tweets sentiment analysis using bidirectional LSTM network. *International Journal of Science and Research, 12*(2), 522–530. https://doi.org/10.21275/SR23209172025

Srinivas, A. C. M. V., Satyanarayana, Ch., Divakar, Ch., & Sirisha, K. P. (2021). Sentiment analysis using neural network and LSTM. *IOP Conference Series: Materials Science and Engineering, 1074*(1), Article 012007. https://doi.org/10.1088/1757-899X/1074/1/012007

Taj, S., Daudpota, S. M., Imran, A. S., & Kastrati, Z. (2025). Aspect-based sentiment analysis for software requirements elicitation using fine-tuned Bidirectional Encoder Representations from Transformers and Explainable Artificial Intelligence. *Engineering Applications of Artificial Intelligence, 151*, Article 110632. https://doi.org/10.1016/j.engappai.2025.110632

Talukder, M. A., Uddin, M. A., Roy, S., Ghose, P., Sarker, S., Khraisat, A., Kazi, M., Rahman, M. M., & Hakimi, M. (2025). A hybrid deep learning model for sentiment analysis of COVID-19 tweets with class balancing. *Scientific Reports, 15*(1), Article 27788. https://doi.org/10.1038/s41598-025-97778-7

Tan, K. L., Lee, C. P., Anbananthen, K. S. M., & Lim, K. M. (2022). RoBERTa-LSTM: A hybrid model for sentiment analysis with transformer and recurrent neural network. *IEEE Access, 10*, 21517–21525. https://doi.org/10.1109/ACCESS.2022.3152828

Varadharajan, V., Smith, N., Kalla, D., Samaah, F., & Mandala, V. (2025). Deep learning-based sentiment analysis: Enhancing IMDb review classification with LSTM models. *Universal Journal of Computer Sciences and Communications, 4*(1), 1–14. https://doi.org/10.31586/ujcsc.2025.1249

Wang, J., Wang, S., & Zhang, Y. (2025). Deep learning on medical image analysis. *CAAI Transactions on Intelligence Technology, 10*(1), 1–35. https://doi.org/10.1049/cit2.12356

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*(7), 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Xiaoyan, L., Raga, R. C., & Xuemei, S. (2022). GloVe-CNN-BiLSTM model for sentiment analysis on text reviews. *Journal of Sensors, 2022*, Article 7212366. https://doi.org/10.1155/2022/7212366

Xu, W., Chen, J., Ding, Z., & Wang, J. (2024). Text sentiment analysis and classification based on bidirectional gated recurrent units (GRUs) model. *Applied and Computational Engineering, 77*(1), 132–137. https://doi.org/10.54254/2755-2721/77/20240670

Zhang, H., Xu, J., Lei, L., Jianlin, Q., & Alshalabi, R. (2023). A sentiment analysis method based on bidirectional long short-term memory networks. *Applied Mathematics and Nonlinear Sciences, 8*(1), 55–68. https://doi.org/10.2478/amns.2022.1.00015