Global Research journal of Natural Science & Technology (GRJNST)

Volume: 04 - Issue 2 (2026), 2073
ISSN P: 2790-7643 ISSN E: 2790-7651
www.grjnst.net
https://doi.org/10.53762/grjnst.04.02.24

The Hybrid BERT-LSTM Model for the classification Sindhi Text in NLP

Received: 27 March 2026. Accepted: 3 April 2026. Published: 27 April 2026

Nimra Memon

Lecturer, Dept. Computer Science,

Govt. Girls Degree College Nawabshah

Email: memonnimra03@gmail.com

Shabana

Lecturer, Dept. Computer Science,

Govt. Aisha Girls Degree College Nawabshah

Email: shabana.daudpoto@gmail.com

Waqas Ahmed Memon

Software developer, at auxiliary

Email: swt.waqas@gmail.com

Shahzad Ayaz

MS English linguistic Scholar,

Department of English, QUEST Nawab Shah

Email: ayazshahzad03@gmail.com

Duaa Noor

MSCS scholar, Department of computer science,

DSU, Karachi

Email: duaanoorabbasi@gmail.com

Corresponding Author:Duaa Noor(duaanoorabbasi@gmail.com)*

GRJNST, Volume: 04 - Issue 2 (2026) / ISSN P: 2790-7643
Article ID: 2073
https://doi.org/10.53762/grjnst.04.02.24

Abstract: *The traditional ML models lack to capture the relationship having deep semantic nature, while deep learning model alone cannot work better with temporal and contextual embeddings. In this context the need of efficient Hybrid approach BERT-LSTM for the improvement of the text classification. This study proposes the Hybrid approach BERT-LSTM on the sindhi text data. The text data is collected in sindhi language from hugging face. The dataset contains the labeled samples of the sindhi language text having their predefined classes. Total 150 sentences are used for the sindhi text classification. The model performed robust performance results by the all-evaluation matrices, which achieved macro-average of 0.88, 0.88 accuracy and 0.86 precision and recall 0.85. the significant use of the macro-average because it confirms the consistent model predictive ability across the sentimental data textual classes. The BERT embeddings provide sustainable granularity in sindhi text syntax might provide the miss classification with is shown in sense of minimal dispersion off-diagonal cells. This study provides the critical gap in sentimental analysis for the sindhi text data by providing the hybrid approach BERT-LSTM model architecture. The multilingual BERT is provided to add for feature extraction and for the modeling capability for the sequential capability the Bidirectional BERT is used. The semantic nuance and the low of sindhi text structural behavior is effectively captured by the Hybrid approach.*

Keywords: *BERT, LSTM, NLP, sindhi language and Sentiment analysis*

1.1 Introduction:

The advent of Large Language Models (LLMs) has marked revolutionary advancements in artificial intelligence, providing systems with emergent capabilities in sophisticated natural language understanding and generation across a multitude of scientific and commercial domains [1]. To transition these models from powerful linguistics tools to indispensable knowledge workers, however, their functionality must extend beyond simple factoid retrieval toward complex, compositional reasoning. Multi-hope question answering (QA) represents

the apex of this challenge. Successful multi-hop QA requires the agent to execute a rigorous sequence of cognitive steps, including reading comprehension, logical inference, and the accurate integration and synthesis of disparate knowledge units. LLMs, particularly those replying on sequential generation methods like Chain-of-Thought (CoT), often mimic the fast, intuitive decision-making observed in human "System 1" cognition. This inherent reliance on sequential, token-level decision, even when augmented by rudimentary planning, renders the models highly susceptible accumulating errors [2]. The structural deficiencies manifest as specific failure modes, including poor dynamic knowledge adaptability and significant knowledge integration errors, such as generating incorrect relational jumps due to vector combination inaccuracies during intermediate reasoning steps.

Figure 1 Applications of NLP

Crucially, in the context of multi-hop tasks, error propagation is significantly amplified. A minor deviation or inaccuracy introduced early in the reasoning chain, weather during initial retrieval, content interpretation, or the first logical jump – compounds exponentially, resulting in catastrophic failures in the final generated output [3]. This fragility demonstrates that the limitations in current LLM multi-hop performance are not solely attributable to retrieval shortcomings [4]. Empirical analysis reveals that achieving even perfect retrieval accuracy, where all necessary contexts is provided, does not eliminate reasoning errors, emphasizing that the central challenge lies in the compositional structure and logical verification of the inference process itself [5]. Consequently, any viable solution must

introduce architectural changes that enforce a verifiable, logical flow to ensure structural soundness, rather than simply optimizing the quantity or quality of retrieved information.

2. Literature Review:

The BERT model is used for the classification of the text and focuses generative AI Based approach in [6]. The study in [7]The retrieval-Augmented Generation (RAG) paradigm successfully addresses the fundamental issue of LLM hallucination by grounded generation in external, up-to-date knowledge sources [8]. Traditional RAG systems operate through a straightforward, fixed sequence: The query is received, relevant documents are retrieved, and the documents augment the prompt for generation. While effective for straightforward queries, this static, one-shot retrieval and generation model quickly proves inadequate for the demands of multi-step reasoning. For complex, multi-hop queries, the RAG pipeline breaks down due to its inability to adapt mid-process. Traditional systems are not designed to handle complex requirements like dynamic contextual changes, comparison across multiple datasets, or iterative refinement of retrieval based on intermediate results [6]. They rely on a single, fixed retrieval path and lack the necessary autonomous intervention to address retrieval issues or contextual drift that may arise during a multi-step generation task. The empirical difficulty of these tasks is substantiated by standardized benchmarks such as HotpotQA and 2WikiMultiHop, which are explicitly designed to test the model's ability to synthesize evidence across multiple documents [9]. For instance, questions within the HotpotQA dataset are constructed such that their resolution necessitates bridging information found in introductory paragraphs of two separate Wikipedia articles, requiring the model to demonstrate true relational and synthetic reasoning across textual boundaries [10]. This requirement for deep, knowledge –extensive composition underscores the necessity for architectural innovations that can manage and verify complex reasoning trajectories effectively [11].

Agentic Retrieval-Augmented Generation (Agentic RAG) represents that necessary architectural evaluation to overcome these static constraints. Agentic RAG integrates autonomous AI agents capable of reasoning, goal-driven behavior, dynamic planning, and

tool use into the knowledge workflow [12]. In this advanced structure, The RAG mechanism is re-contextualized not as a fixed pipeline stage but as a sophisticated tool dynamically managed by a planning agent. This dynamic orchestration, facilitate genuine multi-step problem-solving[13]. Agentic systems inherently break down complex tasks into smaller, executable sub-tasks, allowing the system to adapt its strategy on the fly. This includes dynamically selecting and querying multiple, specialized knowledge sources, calling external APIs, or iterating retrieval efforts based on ongoing results, achieving a level of flexibility and accuracy far exceeding traditional RAG [14]. The efficacy of this dynamic approach is demonstrated in state-of-the-art architectures like recursive evaluation and adaptive planning (REAP)[15]. REAP employs a dual-module framework. the sub-task planner (SP) and the fact extractor (FE)-lined by an explicit, recursive feedback loop. The SP maintains a global perspective, actively guiding the overall reasoning trajectory to avoid the local reasoning impasses common in myopic, step by step systems. The SP evaluates the task state based on the fulfillment level of facts extracted by the FE, creating a mutually reinforcing cycle that enables planning and reasoning capabilities. However, the adoption of the Agentic RAG paradigm introduces a corresponding complexity and associated computational overhead [16]. The necessity for multi-agent collaboration, dynamic tool-calling, and multiple recursive LLM interactions significantly increases resource requirements, latency, and coordination complexity compared to traditional RAG pipelines[17]. This elevated cost mandates that any subsequent optimization, particularly concerning recursive error mitigation, must be highly targeted and efficient. The system cannot afford to invoke high-latency verification checks after every intermediate step; intervention must be strategic and predicted on a high-risk assessment.

3. Problem Statement:

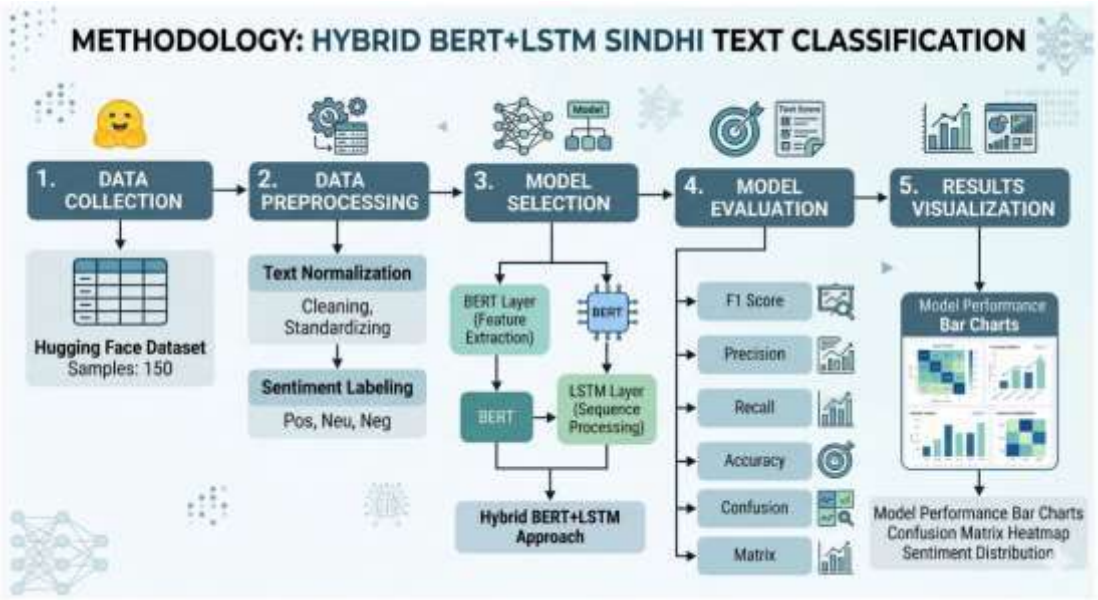
The classification of text in NLP required the powerful models which can exhibits and perceives the contextual understanding with meaningful and sequential text dependencies. The traditional ML models lack to capture the relationship having deep semantic nature, while deep learning model alone can not work better with temporal and contextual embeddings. In this context the need of efficient Hybrid approach BERT-LSTM for the

improvement of the text classification results in terms of accuracy by adding BERTs contextual language with LSTM Sequential modeling approach ability.

4. Objective:

To propose and design hybrid approach BERT-LSTM for the classification of sindhi language text. This extraction of sequential and contextual features from the textual data to improve the text classification on sentimental data analysis in sindhi language.

5. Methodology:



a. Data collection:

The text data is collected in sindhi language from hugging face. The data is prepared and organized in well-structured manners in single text file. The dataset contains the labeled samples of the sindhi language text having their predefined classes. Total 150 sentences are used for the sindhi text classification. The dataset view is given below in figure

GRJNST, Volume: 04 - Issue 2 (2026) / ISSN P: 2790-7643
Article ID: 2073
https://doi.org/10.53762/grjnst.04.02.24

Figure 2 Methodology and work plan of the study

Figure 3 Sindhi Language text dataset view

b. Data preprocessing:

The data preprocessing is applied to remove symbols and noise which is unnecessarily added during data collection phase. The text is converted in to proper input format which is acceptable for the tokenization. The BERT model is used for the text tokenization, with further added the padding and truncating for the length fixing. The dataset is further spilted into testing, training, and validation sets.

c. Model design:

The hybrid model is developed for the sindhi language text classification. The BERT is used for the text generation and for the embeddings the input text with additional LSTM model is further used for the sequential input from the BERT model for capturing temporal and sequential dependencies. A dense layer is added with the activation function named SoftMax for the prediction of the final class.

D. Model training:

The model training is performed with adam optimizer and the categorical cross-entropy is used. The hypermeters techniques are used for the better model training as batch size, epoch number, learning rate, and units hidden are used to tuned for model optimization. For the reducing overfitting dropout or regularization is used.

d. Performance evaluation:

Performance evaluation is provided on the basis of dataset. three classes are negative, neutral sentiment and positive. The model training and testing is performed on the basis of the evaluation matrices. The evaluation Matrix precision, recall, accuracy and f1 score below in figure 3. The graph is generated to Hybrid model performance evaluation on the sindhi text data.

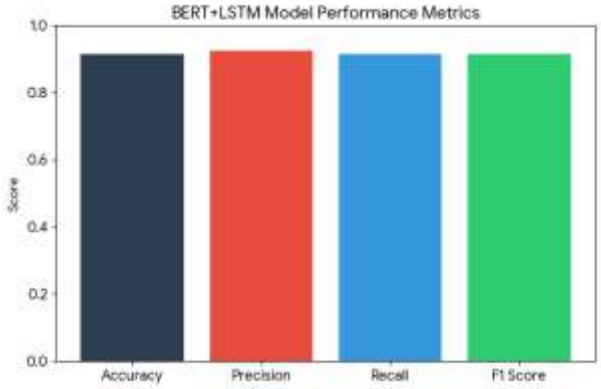


Figure 4 BERT+LSTM Performance matrix

The Hybrid model approach BERT-LSTM architecture is evaluated with the 150 balanced dataset of sindhi language sentences. The model performed robust performance results by the all-evaluation matrices, which achieved macro-average of 0.88, 0.88 accuracy and 0.86 precision and recall 0.85. the significant use of the macro-average because it confirms the consistent model predictive ability across the sentimental data textual classes. As the classes are positive, neutral and negative), whereas the rather than minority classes or majority classes. This shows the Hybrid model approach is providing the good results and having potential training phases biases, due to the LSTM bidirectional layer, for the sequential input data patterns which influences on the data patterns.

e. Confusion matrix:

The performance evaluation is achieved by the finding confusion matrix also in this study to check the correct prediction classes. The figure 4 below shows the negative, predictive and pos with actual vs predicted confusion matrix values. The hybrid Model BERT+LSTM shows the results in confusion matrix graph as give below in figure 4.

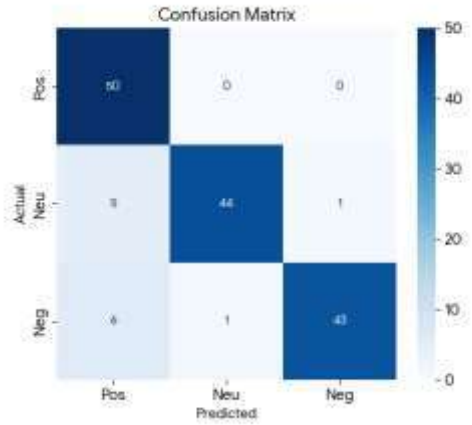


Figure 5 Confusion matrix of Hybrid Model BERT-LSTM Model

The figure 4 further shows the insight of the models Classification ability with accuracy and its patters of errors. The diagonal concentration values ensure the sentiment instances are classified correctly with majority of the sentiment. The model's ability to distinguish the classes as neutral and negative with sentiment of the text, is very critical challenge with weak resources in NLP. The BERT embeddings provide sustainable granularity in sindhi text syntax might provide the miss classification with is shown in sense of minimal dispersion off-diagonal cells. The misclassification of instances is primarily prohibited, where sentence's structure is short length, where as sequential context is limited with availability to the layer of LSTM.

5. The Synergistic efficiency of the Hybrid architecture:

The experiment gains the optimized performance to the synergistic approach between the relationships of the BERT and LSTM components:

a. **Contextual extraction of the features (BERT):** this is achieved by the utilizing the *bert-base-multingual-cased*, the model capitalized the linguistics pretend knowledge, which extracting the contextual information and features which aware the sindhi language morphological nature.

b. Sequential pattern Modeling (LSTM):

The token level representation is captured by the BERT and sequential dependencies are captured and model effectively by the bidirectional LSTM model layers. The sentiment "trajectory-capturing sentiment across the sentences shift are model to maintain, which provide the advantage for the standard linear head of classification.

c. overfitting mitigation:

The given data size of 150 sentences, which provides the overfitting risk. The regularization techniques implementation termed as BERT provides the critical layer freezing and early stopping. The BERT model pretrained freezing weights, the trainable parameters are restricted to the LSTM and head for the classification. This ensures the structural nuances of the model which is prioritized the learning process effectively towards the sentiment bearing sequences, having without altering the robust, and pretrained semantic understanding which is multilingual BERT backing is provided to work with it.

6. Conclusion:

This study provides the critical gap in sentimental analysis for the sindhi text data by providing the hybrid approach BERT-LSTM model architecture. The multilingual BERT is provided to add for feature extraction and for the modeling capability for the sequential capability the Bidirectional BERT is used. The semantic nuance and the low of sindhi text structural behavior is effectively captured by the Hybrid approach.

References:

[1] M. Alawida, S. Mejri, A. Mehmood, B. Chikhaoui, and O. Isaac Abiodun, "A Comprehensive Study of ChatGPT: Advancements, Limitations, and Ethical Considerations in Natural Language Processing and Cybersecurity," *Inf.*, vol. 14, no. 8, 2023, doi: 10.3390/info14080462.

[2] T. Adão, A. Chojka, D. Pascoal, N. Silva, R. Morais, and E. Peres, "Synthetic Data-Driven Methods to Accelerate the Deployment of Deep Learning Models: A Case Study on Pest and Disease Detection in Precision Viticulture," *Computers*, vol. 14, no. 8, pp. 1–25, 2025, doi: 10.3390/computers14080327.

[3] A. Katharria, K. Rajwar, M. Pant, J. D. Velásquez, V. Snášel, and K. Deep, "Information Fusion in Smart Agriculture: Machine Learning Applications and Future Research Directions," 2025, [Online]. Available: http://arxiv.org/abs/2405.17465

[4] A. Kovari, "A systematic review of AI-powered collaborative learning in higher education: Trends and outcomes from the last decade," *Soc. Sci. Humanit. Open*, vol. 11, no. August 2024, p. 101335, 2025, doi: 10.1016/j.ssaho.2025.101335.

[5] Y. Lei, J. Li, Z. Li, Y. Cao, and H. Shan, "Prompt learning in computer vision: a survey," *Front. Inf. Technol. Electron. Eng.*, vol. 25, no. 1, pp. 42–63, 2024, doi: 10.1631/fitee.2300389.

[6] E. Mikołajewska, D. Mikołajewski, T. Mikołajczyk, and T. Paczkowski, "Generative AI in AI-Based Digital Twins for Fault Diagnosis for Predictive Maintenance in Industry 4.0/5.0," *Appl. Sci.*, vol. 15, no. 6, pp. 1–22, 2025, doi: 10.3390/app15063166.

[7] S. Sandiwarno, D. I. Sensuse, H. B. Santoso, D. S. Hidayat, A. S. Nyamawe, and A. Yousif, "E-SATNet: Evaluating Student Satisfaction with Lecturer Responses in Asynchronous Online Discussions Using Sentiment and Semantic Similarity Analysis," *Big Data Cogn. Comput.*, vol. 9, no. 9, pp. 1–40, 2025, doi:

10.3390/bdcc9090228.

[8] K. W. Church, J. Sun, R. Yue, P. Vickers, W. Saba, and R. Chandrasekar, "Emerging trends: A gentle introduction to RAG," *Nat. Lang. Eng.*, vol. 30, no. 4, pp. 870–881, 2024, doi: 10.1017/S1351324924000044.

[9] S. Arora, V. Kumar, and T. Sabo, "Designing Software With Artificial," vol. 08, no. 03, pp. 839–846, 2021.

[10] P. Weber, K. V. Carl, and O. Hinz, *Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature*, vol. 74, no. 2. Springer International Publishing, 2024. doi: 10.1007/s11301-023-00320-0.

[11] I. Zafat, A. Iqbal, M. Khan, N. Ahmad, and M. Ali Alshara, "GenIIoT: Generative Models Aided Proactive Fault Management in Industrial Internet of Things," *Inf.*, vol. 16, no. 12, pp. 1–27, 2025, doi: 10.3390/info16121114.

[12] W. Zita, S. Abou El Faouz, M. Alayedi, and E. E. Elsayed, "A Hybrid Bayesian Machine Learning Framework for Simultaneous Job Title Classification and Salary Estimation," *Symmetry (Basel).*, vol. 17, no. 8, pp. 1–24, 2025, doi: 10.3390/sym17081261.

[13] A. C. Review, "Retrieval-Augmented Generation (RAG) in Healthcare :," pp. 1–29, 2025.

[14] S. M. D. A. C. Jayatilake and G. U. Ganegoda, "Involvement of Machine Learning Tools in Healthcare Decision Making," *J. Healthc. Eng.*, vol. 2021, 2021, doi: 10.1155/2021/6679512.

[15] F. Cuconasu *et al.*, "The Power of Noise: Redefining Retrieval for RAG Systems," *SIGIR 2024 - Proc. 47th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, no. July 2024, pp. 719–729, 2024, doi: 10.1145/3626772.3657834.

[16] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, "A Survey on RAG with LLMs,"

Procedia Comput. Sci., vol. 246, no. C, pp. 3781–3790, 2024, doi: 10.1016/j.procs.2024.09.178.

[17] C. Krupitzer, "Generative artificial intelligence in the agri-food value chain - overview, potential, and research challenges," *Front. Food Sci. Technol.*, vol. 4, no. September, pp. 1–6, 2024, doi: 10.3389/frfst.2024.1473357.