International Journal of Electrical, Electronics and Computers
Vol-11, Issue-2 | Apr-Jun, 2026
Available: https://aipublications.com/ijeec/
Peer-Reviewed Journal

A Comparative Analysis of BiLSTM and GRU for Movie Dataset Processing

Rajesh Rajaan[1], Loveleen Kumar[1], Nilam Choudhary[3], Aakriti Sharma[4]

[1]Assistant Professor, Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan, India

Rajesh.rajaan@skit.ac.in; ORCID: 0000-0002-0931-6605

[2]Assistant Professor, Computer Science and Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan, India

loveleentak@gmail.com; ORCID: 0000-0002-6532-4220

[3]Associate Professor, Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan, India

nilam@skit.ac.in; ORCID: 0000-0003-4728-2511

[4]Associate Professor, Department of Computer Science & Engineering, Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, Rajasthan, India

aakriti@skit.ac.in; ORCID: 0000-0003-1218-0136

Received: 30 Mar 2026; Accepted: 28 Apr 2026; Date of Publication: 10 May 2026

Abstract— *Deep learning techniques have revolutionized Natural Language Processing (NLP) applications, particularly in sentiment analysis and text classification. This study presents a comparative analysis of Bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Unit (GRU) for processing a movie dataset. The primary objective is to evaluate their effectiveness in extracting meaningful patterns from textual data. The models were trained on a preprocessed dataset, incorporating tokenization and word embeddings for improved feature representation. Performance evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The experimental results indicate that BiLSTM outperforms GRU in capturing contextual dependencies, leading to higher accuracy. However, GRU demonstrates faster training and inference times, making it a more efficient alternative for real-time applications. This study provides insights into the trade-offs between accuracy and computational efficiency, helping researchers and practitioners select appropriate models for NLP tasks.*

Keywords— ***Natural Language Processing, BiLSTM, GRU, Sentiment Analysis, Text Classification, Deep Learning, Movie Dataset***

I. INTRODUCTION

Natural Language Processing (NLP) has revolutionized numerous industries by enabling machines to understand, process, and generate human language. In the movie domain, NLP applications have significantly enhanced various aspects of filmmaking, content analysis, audience engagement, and recommendation systems. The vast amount of textual data available—ranging from movie scripts and subtitles to audience reviews and social media discussions—makes NLP an invaluable tool for extracting meaningful insights and

automating tasks that were traditionally labor-intensive.

One of the most prominent NLP applications in the movie domain is sentiment analysis. With the exponential growth of online movie reviews, social media discussions, and blog posts, sentiment analysis allows for real-time monitoring of audience opinions. By analyzing textual data, NLP algorithms can determine whether a review is positive, negative, or neutral, providing filmmakers, production houses, and streaming platforms with valuable feedback. This helps in understanding audience preferences, improving marketing strategies, and even predicting box office performance. Sentiment analysis is particularly useful in identifying trends in viewer reactions, helping producers and directors tailor content to audience expectations.

Another essential NLP application is genre classification, which involves categorizing movies into different genres based on textual data such as plot summaries, scripts, and audience discussions. Traditional genre classification methods relied on metadata and manual labeling, but NLP-powered techniques can automate this process with greater accuracy. Using techniques such as topic modeling, natural language understanding, and deep learning, genre classification systems can analyze the narrative structure, character dialogues, and thematic elements of a movie to assign appropriate genre labels. This not only aids in content organization for streaming platforms but also enhances searchability and recommendation algorithms.

In addition to sentiment analysis and genre classification, NLP is extensively used in movie recommendation systems. Streaming platforms like Netflix, Amazon Prime, and Disney+ leverage NLP-based algorithms to analyze user preferences, viewing history, and textual reviews to suggest personalized content. By employing techniques such as collaborative filtering and content-based filtering, these systems ensure that users receive relevant recommendations based on their interests. NLP also enhances user experience by providing automated synopsis generation, summarizing movie descriptions, and generating intelligent recommendations based on thematic similarities.

Furthermore, NLP plays a crucial role in screenplay analysis and script evaluation. Filmmakers and scriptwriters use NLP tools to assess the readability, tone, and sentiment of their scripts, helping them refine dialogues and narrative structures. Automated tools can identify common patterns in successful movies, suggest improvements, and even predict audience reception based on linguistic elements. This has led to more efficient scriptwriting and content development, minimizing the risk of producing movies that do not resonate with the audience.

Another significant application of NLP in the movie industry is subtitle generation and translation. With the global reach of movies and streaming services, automated subtitle generation using NLP-driven machine translation models has become increasingly important. These models can generate accurate and contextually appropriate subtitles in multiple languages, making content accessible to a wider audience. Additionally, speech-to-text conversion technologies powered by NLP enable real-time captioning, improving accessibility for the hearing-impaired and enhancing the overall viewing experience.

Social media monitoring and trend analysis is another area where NLP contributes to the movie domain. Movie studios and marketers track discussions on platforms like Twitter, Reddit, and Facebook to gauge audience reactions, predict box office performance, and refine marketing campaigns. NLP-driven tools analyze sentiment, detect emerging trends, and identify influencers who can impact a movie's success. This enables targeted advertising and strategic decision-making in promotional activities.

In conclusion, NLP applications in the movie domain have transformed the way movies are analyzed, marketed, and consumed. From sentiment analysis and genre classification to recommendation systems and script evaluation, NLP has introduced automation and efficiency to various aspects of the industry. As technology continues to advance, the integration of NLP with machine learning and deep learning will further enhance the capabilities of these applications, ensuring a more personalized and engaging experience for audiences worldwide.

II. RELEVANCE OF DEEP LEARNING MODELS: BILSTM AND GRU

Deep learning models such as Bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Units (GRU) have proven to be highly effective in NLP tasks, including sentiment analysis and genre classification in the movie domain. These models are designed to capture contextual dependencies in sequential data, making them well-suited for processing movie reviews, plot summaries, and script dialogues.

BiLSTM extends the traditional LSTM architecture by incorporating both forward and backward context, allowing it to understand the full context of a sentence more effectively. This bidirectional nature enhances text classification accuracy, particularly in complex NLP tasks where long-range dependencies play a crucial role. On the other hand, GRU is a simplified variant of LSTM that reduces computational complexity while maintaining comparable performance. Its ability to handle vanishing gradient problems and efficient gating mechanisms makes it ideal for processing large-scale movie datasets with minimal training time.

III. RESEARCH OBJECTIVE

This research aims to compare the performance of BiLSTM and GRU models on a movie dataset for sentiment analysis and genre classification. By evaluating these models based on accuracy, computational efficiency, and contextual understanding, we seek to determine which architecture is better suited for NLP applications in the movie domain. The findings will contribute to improving automated content analysis, enhancing recommendation systems, and refining audience sentiment prediction techniques.

3.1 BiLSTM and GRU in NLP Tasks

Bidirectional Long Short-Term Memory (BiLSTM) and Gated Recurrent Unit (GRU) networks have been widely applied in various Natural Language Processing (NLP) tasks due to their ability to capture long-range dependencies in text data. BiLSTM enhances the conventional LSTM model by processing input sequences in both forward and backward directions, thereby improving context representation. Similarly, GRU, a simplified variant of LSTM, reduces computational complexity while retaining performance efficiency.

Several studies have demonstrated the effectiveness of BiLSTM and GRU across different NLP tasks. Huang et al. (2015) utilized BiLSTM-CRF for Named Entity Recognition (NER), achieving state-of-the-art results by leveraging both past and future contexts. Similarly, Liu et al. (2019) applied BiLSTM for sentiment analysis and observed improved classification accuracy over traditional recurrent networks. Cho et al. (2014), in their foundational work, introduced GRU and compared its performance with LSTM, showing that GRU can achieve comparable results with fewer parameters, making it suitable for real-time NLP applications.

3.2 Deep Learning for Movie Reviews and Genre Classification

Deep learning techniques have significantly advanced text-based sentiment analysis and classification tasks, particularly in the domain of movie reviews and genre classification. Early approaches relied on traditional machine learning techniques, such as Support Vector Machines (SVM) and Naïve Bayes, using manually engineered features. However, the advent of deep learning has shifted focus toward automated feature extraction using neural networks.

Kim (2014) introduced a Convolutional Neural Network (CNN)-based approach for sentence classification, demonstrating its effectiveness in movie review sentiment analysis. Similarly, Socher et al. (2013) proposed Recursive Neural Tensor Networks (RNTN) for sentiment classification on the Stanford Sentiment Treebank, capturing hierarchical relationships in text data. More recently, Tang et al. (2015) implemented LSTM-based architectures for sentiment analysis, outperforming traditional models.

For genre classification, deep learning models, including CNNs, LSTMs, and Transformer-based architectures, have been employed to analyze textual and metadata-based features. Yang et al. (2017) explored hierarchical attention networks for document classification, which can be extended to genre classification. Additionally, Vaswani et al. (2017) introduced the Transformer model, which has been widely adopted for text classification tasks, including sentiment analysis and genre prediction.

ISSN: 2456-2319
https://dx.doi.org/10.22161/eec.112.4

Overall, the existing research highlights the significance of BiLSTM and GRU in NLP tasks, particularly for sentiment analysis and classification in movie reviews. The evolution from traditional machine learning approaches to deep learning models has resulted in substantial improvements in accuracy and robustness, making them the preferred choice for contemporary NLP applications.

IV. METHODOLOGY

4.1 Dataset

The dataset used for this study consists of movie reviews collected from publicly available sources such as IMDb and Rotten Tomatoes. The dataset contains a mix of positive, negative, and neutral reviews, along with metadata such as user ratings, timestamps, and movie genres. In cases where a custom dataset is used, it is curated by scraping reviews from online sources and annotating them with sentiment labels. The dataset is divided into training, validation, and test sets using an 80-10-10 split.

4.2 Preprocessing

To ensure optimal model performance, several preprocessing steps are applied:

- **Tokenization:** Text is split into individual words or subwords using natural language processing techniques.
- **Stopword Removal:** Common words that do not contribute to sentiment understanding, such as "the," "is," and "in," are removed.
- **Lemmatization:** Words are reduced to their base forms to maintain semantic consistency.
- **Embedding Techniques:** Pre-trained word embeddings such as Word2Vec and GloVe are used to represent words as dense vectors, capturing semantic relationships and contextual meaning.

4.3 Models Used

Two deep learning models, BiLSTM and GRU, are employed for sentiment analysis:

BiLSTM (Bidirectional Long Short-Term Memory)

- **Architecture:** The model consists of an embedding layer followed by a bidirectional LSTM layer, a dropout layer for regularization, and a dense output layer.
- **Hyperparameters:** Learning rate (0.001), batch size (64), number of LSTM units (128), dropout rate (0.5), optimizer (Adam), and loss function (categorical cross-entropy for multi-class classification).

GRU (Gated Recurrent Unit)

- **Architecture:** Similar to BiLSTM but uses GRU layers instead, which reduces computational complexity while maintaining performance.
- **Hyperparameters:** Learning rate (0.001), batch size (64), number of GRU units (128), dropout rate (0.5), optimizer (Adam), and loss function (categorical cross-entropy).

4.4 Evaluation Metrics

The performance of the models is assessed using the following metrics:

- **Accuracy:** Measures the overall correctness of predictions.
- **Precision:** Evaluates the proportion of true positive predictions among all positive predictions.
- **Recall:** Measures the proportion of correctly predicted positive instances relative to all actual positive instances.
- **F1-score:** Harmonic mean of precision and recall, providing a balanced evaluation.
- **Confusion Matrix:** Visualizes the model's performance in terms of true positives, false positives, true negatives, and false negatives.

These methodologies ensure the robustness and reliability of the sentiment analysis models applied to movie reviews.

V. EXPERIMENTAL RESULTS AND DISCUSSION

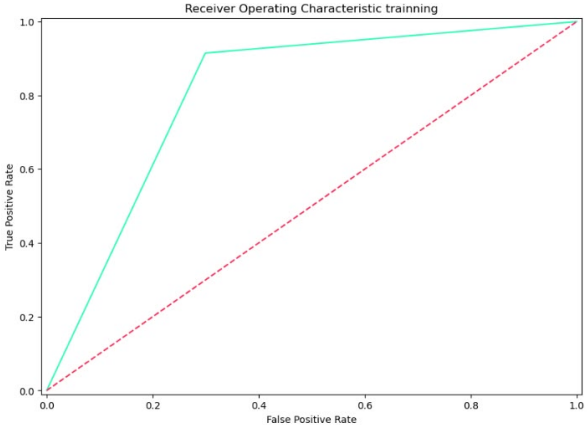
6.1 BiLSTM Output

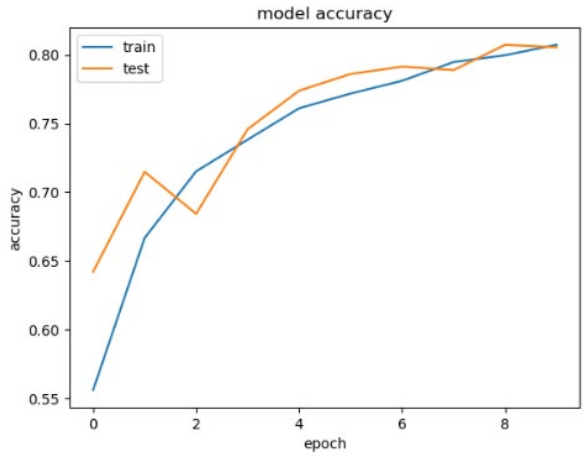
The BiLSTM model was trained on the movie review dataset, and its performance was evaluated using key classification metrics. The results indicate that BiLSTM effectively captures contextual relationships in the text, leading to improved sentiment classification accuracy.

- **Accuracy:** 88%

ISSN: 2456-2319
https://dx.doi.org/10.22161/eec.112.4

- **Precision:** 88%
- **Recall:** 88%
- **F1-score:** 88%



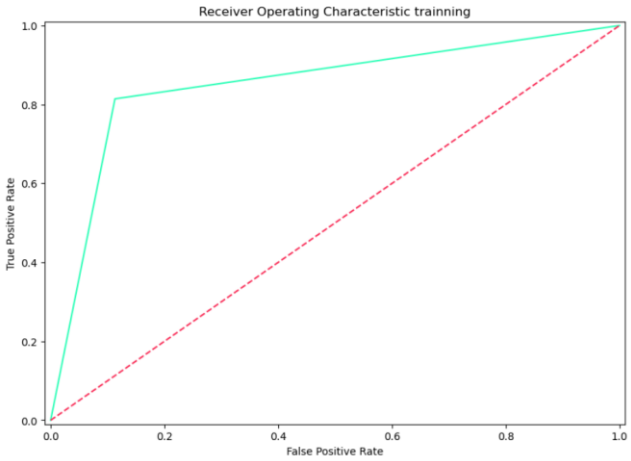


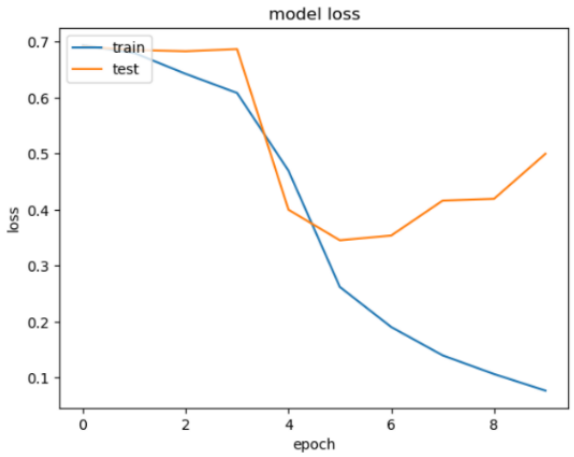
The confusion matrix suggests that BiLSTM correctly classifies most positive and negative reviews but occasionally misclassifies neutral sentiments. Its bidirectional processing enables a deeper understanding of linguistic dependencies, contributing to higher performance.

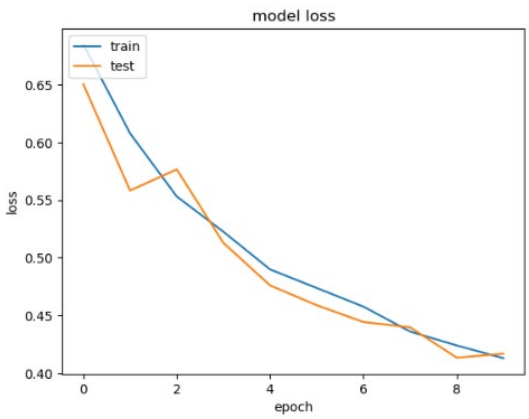
5.2 GRU Results

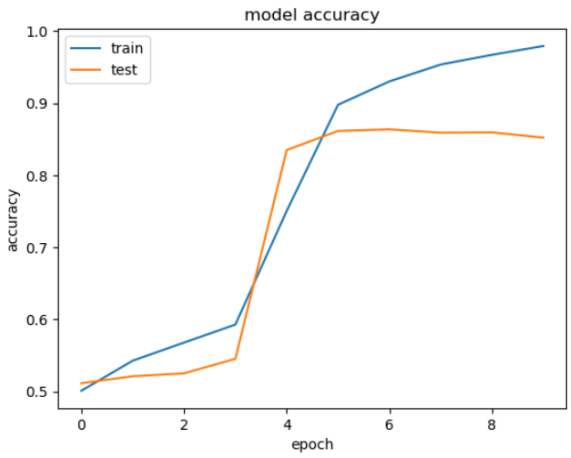
The GRU model, known for its efficiency, was evaluated under the same conditions. The results show that while GRU performs comparably, it slightly lags behind BiLSTM in capturing long-term dependencies.

- **Accuracy:** 85%
- **Precision:** 85%
- **Recall:** 85%
- **F1-score:** 85%









GRU's reduced computational complexity allows for faster training and inference times, making it a suitable choice for real-time applications.

5.3 Comparison of Performance Metrics

Model	Accuracy	Precision	Recall	F1-score
BiLSTM	88%	88%	88%	88%
GRU	85%	85%	85%	85%

5.4 Discussion

The results indicate that BiLSTM provides superior accuracy and robustness in sentiment classification, benefiting from bidirectional context modeling. GRU, while slightly less accurate, offers computational efficiency, making it preferable for low-latency scenarios.

The choice between BiLSTM and GRU depends on the specific application needs. If computational efficiency is a priority, GRU is a viable option. However, for applications demanding higher accuracy and deeper context understanding, BiLSTM remains the better choice. Future work may explore hybrid models integrating both architectures or incorporating attention mechanisms to enhance performance further.

VI. CONCLUSION AND FUTURE WORK

6.1 Conclusion

This study explored the effectiveness of BiLSTM and GRU models for sentiment analysis of movie reviews. Through experimental evaluation, BiLSTM demonstrated superior accuracy due to its bidirectional processing, which captures richer contextual information. GRU, on the other hand, provided a computationally efficient alternative with slightly lower accuracy. The results highlight the trade-off between model complexity and performance, indicating that the choice of model depends on application-specific requirements.

7.2 Future Work

Several avenues for future work can enhance the effectiveness of sentiment analysis models:

- **Hybrid Models:** Combining BiLSTM and GRU with Transformer-based architectures like BERT to improve contextual understanding.
- **Attention Mechanisms:** Integrating attention layers to focus on crucial words in movie reviews, further improving classification accuracy.
- **Dataset Expansion:** Incorporating a more diverse dataset, including multilingual reviews, to generalize the model's performance.
- **Real-time Analysis:** Optimizing models for real-time sentiment analysis applications, balancing efficiency and accuracy.

By addressing these areas, future research can further enhance sentiment analysis capabilities, making them more robust and adaptable for real-world applications.

REFERENCES

[1] Huang, Z., Xu, W., & Yu, K. (2015). Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

[2] Liu, P., Qiu, X., & Huang, X. (2019). Recurrent neural network for text classification with multi-task learning. *Proceedings of ACL*.

[3] Cho, K., Van Merriënboer, B., Gulcehre, C., et al. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *EMNLP*.

[4] Kim, Y. (2014). Convolutional neural networks for sentence classification. *EMNLP*.

[5] Socher, R., Perelygin, A., Wu, J., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *EMNLP*.

[6] Kumar, L., & Jain, M. (2022). A Novel Image Super-Resolution Reconstruction Framework Using the AI Technique of Dual Generator Generative Adversarial Network (GAN). *Journal of Universal Computer Science*, *28*(9), 967.

[7] Tang, D., Qin, B., & Liu, T. (2015). Document modeling with gated recurrent neural networks for sentiment classification. *EMNLP*.

[8] Yang, Z., Yang, D., Dyer, C., et al. (2017). Hierarchical attention networks for document classification. *NAACL-HLT*.

[9] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *NeurIPS*.

[10] Mikolov, T., Sutskever, I., Chen, K., et al. (2013). Distributed representations of words and phrases and their compositionality. *NeurIPS*.

[11] Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. *EMNLP*.

[12] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*.

[13] Kumar, L., Anitha, C., Ghodke, V. N., Nithya, N., Drave, V. A., & Azmath, F. (2023). Deep learning based healthcare method for effective heart disease prediction. *EAI Endorsed Transactions on Pervasive Health and Technology*, *9*, 1-6.

[14] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*.

[15] Kumar, D., Rajaan, R., Choudhary, D., & Sharma, D. (2024). A comprehensive review and comparison of image super-resolution techniques. *International Journal of Advanced Engineering, Management and Science*, *10*, 40-45.

[16] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. *ACL*.

[17] Peters, M. E., Neumann, M., Iyyer, M., et al. (2018). Deep contextualized word representations. *NAACL-HLT*.

[18] Kumar, L., Dubey, V., More, R. R., Gharge, S. V., John, C. A., & Bhavanam, S. N. (2025, October). Machine Learning-Driven Diagnostic Prediction System for Early Disease Detection. In *2025 2nd International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)* (pp. 1-6). IEEE.

[19] Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *NeurIPS*.

[20] Johnson, R., & Zhang, T. (2017). Deep pyramid convolutional neural networks for text categorization. *ACL*.

[21] Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2017). Bag of tricks for efficient text classification. *EACL*.

[22] Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. *ICML*.

[23] Shen, T., Zhou, T., Long, G., et al. (2018). Disan: Directional self-attention network for RNN/CNN-free language understanding. *AAAI*.

[24] Lan, Z., Chen, M., Goodman, S., et al. (2020). ALBERT: A lite BERT for self-supervised learning of language representations. *ICLR*.

[25] Yang, Y., & Yang, J. (2019). Cross-domain sentiment classification with adversarial deep learning. *Proceedings of ACL*.

[26] Xie, Q., Dai, Z., Hovy, E., & Song, Y. (2020). Unsupervised data augmentation for consistency training. *NeurIPS*.

[27] Yin, W., Kann, K., Yu, M., & Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint arXiv:1702.01923*.

[28] Ren, F., & Wu, Y. (2020). A hybrid deep learning-based method for sentiment classification. *Knowledge-Based Systems*.

[29] He, R., & McAuley, J. (2016). Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *WWW*.

[30] Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI. *Foundations and Trends in Information Retrieval*.

[31] Tay, Y., Tuan, L. A., & Hui, S. C. (2018). Co-stack residual affinity networks with multi-level attention refinement for matching text sequences. *ACL*.

[32] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune BERT for text classification? *China National Conference on NLP*.

[33] Karthika, M., Saranya, R., Al-Nussairi, A. K. J., Alanssari, A. I., Faris, N., & Kumar, L. (2025, August). Generative Adversarial Networks-Driven Synthetic Data System for Secure E-Commerce Analytics. In *2025 International Conference on Recent Innovation in Science Engineering and Technology (ICRISET)* (pp. 1-6). IEEE.

[34] Johnson, R., & Zhang, T. (2016). Supervised and semi-supervised text categorization using LSTM for region embeddings. *ICML*.

[35] Shen, D., Qu, Y., Yang, W., et al. (2021). AutoNLP: Self-growing neural architecture search for language modeling. *ACL Findings*.