Comparative Performance of Deep Learning Approaches for Sentiment Analysis on Pakistani Dramas and Movies Reviews

Manzoor Hussain[1], Shamshad Lakho[1], Zulqarnain Channa[2], Muhammad Alam[2], Imran Ali Memon[3]

[1]Department of Information Technology, Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Pakistan

[2]Department of computer science , Quaid-e-Awam University of Engineering, Science & Technology, Nawabshah, Pakistan

[3]Shaheed Benazir Bhutto University, Shaheed Benazirabad, Pakistan

***Correspondence**: manzoorhussain575@gmail.com, shamshad.lakho@quest.edu.pk, zulqarnainchana@gmail.com, alam14it36@gmail.com, imran.asif.memon@sbbusba.edu.pk

Abstract:

Sentiment analysis plays an important role in natural language processing, helping to understand public opinions shared through text. This study focuses on the challenge of analyzing sentiments in reviews of Pakistani dramas and movies, where mixed languages, informal expressions, and noisy data make accurate classification difficult. To solve this problem, several deep learning models were used and tested, including Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Bidirectional LSTM (Bi-LSTM), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN). A detailed dataset of 12,000 user reviews was collected from platforms like IMDb and YouTube. The data was cleaned and prepared through steps such as tokenization, removing unnecessary columns, normalizing, and using sentiment scoring and word embedding for feature extraction. These models were trained and tested, with their performance evaluated using accuracy, precision, recall, and F1-score. Among all, the CNN model performed the best, achieving 98.71% accuracy and a 98.49% F1-score. The Bi-LSTM model was close behind, with 98.59% accuracy and a 98.47% F1-score. In the future, the research will explore the use of advanced transformer-based models like BERT and GPT for multilingual sentiment analysis. It will also aim to build real-time sentiment classification systems. Moreover, creating sentiment lexicons for regional languages and using hybrid deep learning methods are suggested to further improve accuracy and generalization.

Keywords: Sentiment Analysis, Deep Learning, CNN, LSTM, Bi-LSTM, GRU, RNN, Pakistani Dramas, Movie Reviews, Text Preprocessing, Accuracy Comparison, Multilingual Data

Introduction:

The rapid rise of digital content in the entertainment industry has led to a growing number of user reviews for Pakistani dramas and movies. Understanding the public's feelings in these reviews is important for content creators, production houses, and streaming platforms, helping them shape their work to match audience preferences A. Malik and A. Zia [1]. However, accurately analyzing these reviews is still difficult because they often include mixed languages, slang, informal language, and messy or unstructured data. Even though traditional sentiment analysis methods have improved, many models still struggle to understand the unique features of regional languages and cultural expressions S. Awasthi et al. [2], which can lead to incorrect interpretations of public opinion.

This study aims to overcome these challenges by testing and comparing different deep learning models for sentiment analysis on reviews of Pakistani dramas and movies. The goal is to find the most accurate and efficient model that can handle complex text and offer useful insights to the entertainment industry. Previous research has mostly focused on global contexts using standard datasets and methods like Support Vector Machines (SVM), Naive Bayes classifiers, and simple lexicon-based techniques. R. Srivastava et al. [3] While some studies have used deep learning models like LSTM and CNN, they often overlook the issues related to mixed-language content and industry-specific needs.

There is also limited research focused on sentiment analysis of Pakistani entertainment content, which shows a clear research gap. This study addresses that gap by using deep learning models specifically on Pakistani drama and movie reviews, contributing to both academic knowledge and practical tools for content production and marketing strategies.

Objectives:

The primary purposes of the research can be outlined as follows:

1. To preprocess Pakistani Dramas and Movies Reviews data, ensuring it is ready for deep learning analysis by removing noise and applying tokenization techniques.
2. To train and evaluate multiple deep learning models, including LSTM, CNN and others deep learning model on the Pakistani Dramas and Movies Reviews dataset.
3. To compare performance metrics such as accuracy, precision, recall, and F1-score across models to determine the most effective approach for sentiment analysis.

Novelty statement:

This study builds the first ever deep learning model to perform sentiment analysis on mixed Urdu and English reviews of Pakistani dramas and movies, addressing a significant deficit in local NLP research. Unlike prior works focused on monolingual datasets, we assess the performance of five sophisticated models CNN, Bi-LSTM, GRU, LSTM, and RNN on this complex dataset, demonstrating that CNN outperforms the other models with 98.71% accuracy on capturing code-mixed informal slang. Our work serves as a prominent example of advanced regional sentiment analysis and provides rich insights for the Pakistani entertainment sector through innovative preprocessing models and optimization strategies.

Literature Review:

Recent studies have explored different methods for sentiment analysis using both traditional machine learning and advanced deep learning models. For example, M. Sani et al. [4] analyzed Hausa tweets using Multinomial Naive Bayes (MNB) and Logistic Regression (LR), finding LR more effective with 86% accuracy. However, their work was limited to one language and platform, reducing its broader applicability.

Similarly, M. ur-Rehman et al. [5] used Support Vector Machines (SVM) and Kernel SVM for real-time sentiment analysis on Twitter, achieving 80% accuracy in detecting positive sentiments. Yet, their focus remained on political and conflict-related content, leaving entertainment-based sentiment analysis mostly unexplored. Khan et al. [6] made a valuable contribution by creating a benchmark dataset of 9,601 Urdu reviews across multiple domains.

Their study showed that logistic regression with uni-trigram features achieved the highest accuracy of 81.94%. However, the dataset did not include mixed-language content or address challenges related to informal online language.

In another study, Raza et al [7] introduced a sentiment-specific convolutional neural network for Roman Urdu text, effectively handling transliteration issues. Still, their research was limited to a single model and did not compare multiple deep learning architectures. These gaps point to the need for research that focuses on multilingual, informal, and unstructured data, especially in the context of Pakistani entertainment reviews I. U. Khan et al [8]. Unlike earlier studies, this research compares several deep learning models—CNN, LSTM, Bi-LSTM, GRU, and RNN—on a large, mixed-language dataset of reviews from Pakistani dramas and movies. By addressing domain-specific challenges and offering a comprehensive model comparison, this study aims to improve sentiment analysis accuracy and provide valuable insights to the entertainment industry.
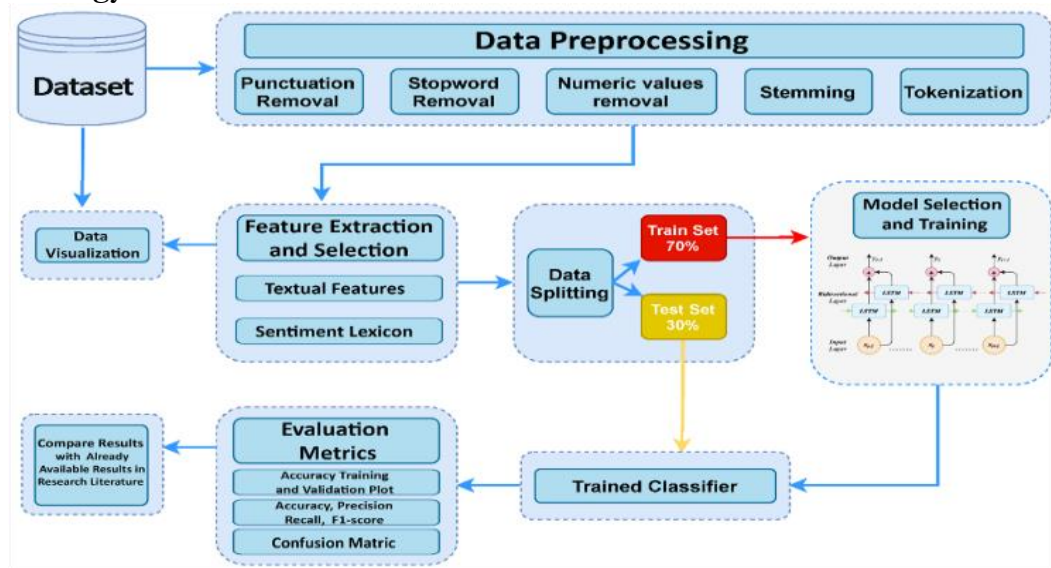
Methodology:

Figure 1. Flow Diagram of Methodology

Research Design:
Dataset Collection:
A dataset of 12,288 Pakistani drama and movie reviews was collected from IMDb, YouTube, and social media platforms. The dataset consists of three columns: Title, Type (Drama/Movie), and Review Text. Data sources included both structured review platforms and unstructured social media comments.

Data Pre-processing:
- Tokenization: Splitting text into words and phrases.
- Stop word Removal: Eliminating frequently occurring words that do not contribute to sentiment.
- Text Normalization: Handling different spellings, removing special characters.
- Feature Extraction: Applying TF-IDF, Word2Vec, and BERT embeddings.

Deep Learning Models:
The following models were implemented for sentiment classification:
LSTM: Effective for sequential data processing.
GRU: A variation of LSTM with fewer parameters.
Bi-LSTM: Captures both past and future dependencies.
CNN: Extracts spatial features from text.
RNN: Basic recurrent network for sentiment analysis.

Evaluation Metrics:

- Accuracy: Overall correctness of the model.
- Precision: Ratio of correctly predicted positive observations.
- Recall: Ability to find all relevant instances.
- F1-Score: Harmonic mean of precision and recall.

Model Architecture:

(None, 50, 128)
(None, 76)

Figure 2. Model Architecture

As the model architecture consists of two main layers shows in Figure 2.

- **Input Layer:** This layer accepts input data with a shape of (None, 50, 128). Here, "None" indicates that the number of samples can vary, "50" represents the sequence length or the number of time steps, and "128" denotes the dimensionality of each input feature or word embedding.
- **Output Layer:** This layer produces output data with a shape of (None, 76). Again, "None" signifies the variable number of samples, while "76" represents the dimensionality of the output space. In this context, the output likely corresponds to class probabilities or regression values, depending on the specific task the model is designed for.

Results and Discussion:

Load Data:

The dataset has Review, Title, and Type columns. Preprocessing includes handling missing values, removing duplicates, and resolving inconsistencies. A quality check confirms dataset organization before sentiment analysis.

Figure 3. Head and Tail: first and Last Five Entries of the Dataset

Analysis and Visualization of the Dataset:

A bar chart (Figure 4) compares review counts for dramas and movies, helping researchers understand dataset distribution. This visualization highlights class balance, ensuring fair sentiment analysis.
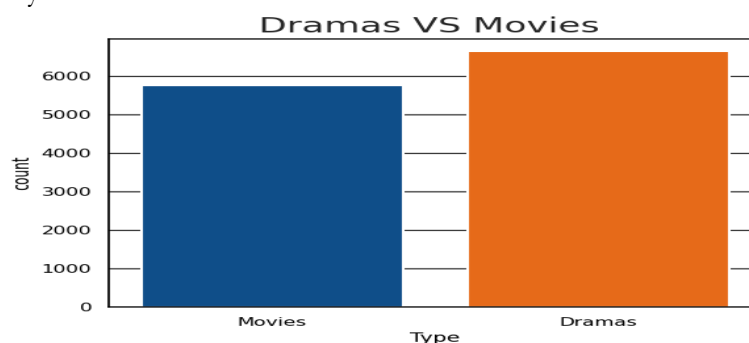
Figure 4. Dramas VS Movies

Review Length Distribution Analysis:

Review length analysis reveals most user feedback is short, with a right-skewed distribution (Figure 5). While short reviews dominate, some longer ones exist in the dataset.

Figure 5. Distribution of Reviews Length
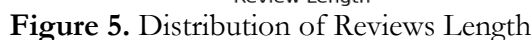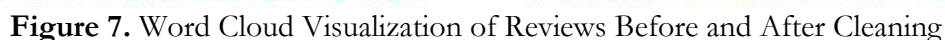
Common Words in Reviews:

Figure 6, word cloud highlights frequent review terms like "movie," "story," "acting," and sentiment words like "good" and "amazing," reflecting viewer focus on quality and emotions

Figure 6. Word Cloud of Reviews Column

Word Cloud Visualization:

Figure 7 compares word clouds from the raw Reviews and cleaned Reviews_clean columns. The raw data includes frequent but noisy terms like "movie", "film", "one", "story" and "character" along with general words and punctuation that add clutter. After stop word removal, punctuation deletion, and special character elimination, the cleaned dataset retains essential terms like "movie," "film," "love," "story," and "character," improving sentiment analysis accuracy.

Figure 7. Word Cloud Visualization of Reviews Before and After Cleaning

Sentiment Score Calculation Process:

The sentiment score calculation follows key steps: it loads lists of positive (pst_file) and negative (neg_file) words, then compares each sentence in the dataset to count matching words. Sentiment is calculated using:

Sentiment Score = Positive Words - Negative Words

The difference (diff_posneg) determines sentiment:

> **> 0** → Positive (1)
>
> **< 0** → Negative (-1)
>
> **= 0** → Neutral (0)

These scores are stored in a new column (sentiment_score) for further analysis. Figures 8 and 9 illustrate this sentiment scoring process.

	Reviews	Reviews_clean	diff_postneg	sentscore_calculated
0	In my submission, I have not seen a Pakistani ...	submiss i see understood pakistani film fame w...	18	1
1	I went to watch this movie with high hopes but...	went watch movi high hope far less expect acti...	14	1
2	Before i start,i want to clarify that I'm not ...	starti want clarifi review mere movi fanaticso...	68	1
3	On the way back from the cinema.. feeling prou...	where are you back cinema feel proud hope paki...	25	1
4	This is my first ever movie review so if there...	first ever movi review mistak would tri better...	46	1

Figure 8. Sentiment Score Calculation
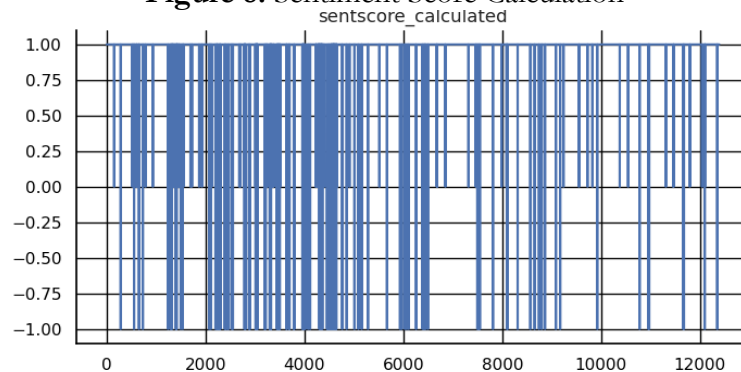
Figure 9. Show Graphical Calculated Sentiment Score

Classification Reports:

The classification report provides precision, recall, and F1-score for each sentiment class. These metrics help determine the overall effectiveness of each model.

LSTM Classification Report:

Figure 10 presents the classification report for the LSTM model.

```
Classification Report of LSTM Model

              precision    recall  f1-score   support

           0       0.00      0.00      0.00        26
           1       0.71      0.48      0.57        42
           2       0.98      1.00      0.99      2410

    accuracy                           0.98      2478
   macro avg       0.57      0.49      0.52      2478
weighted avg       0.97      0.98      0.97      2478
```

Figure 10. LSTM Model Classification Report

GRU Classification Report:

Figure 11 illustrates the classification report for the GRU model.

```
Classification Report of GRU Model

              precision    recall  f1-score   support

           0       0.00      0.00      0.00        26
           1       0.64      0.50      0.56        42
           2       0.98      1.00      0.99      2410

    accuracy                           0.98      2478
   macro avg       0.54      0.50      0.52      2478
weighted avg       0.97      0.98      0.97      2478
```

Figure 11. GRU Model Classification Report

Bi-LSTM Classification Report:

Figure 12 shows the classification report of the Bi-LSTM model.

```
Classification Report of BiLSTM Model

              precision    recall  f1-score   support

           0       0.57      0.50      0.53        26
           1       0.92      0.57      0.71        42
           2       0.99      1.00      0.99      2410

    accuracy                           0.99      2478
   macro avg       0.83      0.69      0.74      2478
weighted avg       0.98      0.99      0.98      2478
```

Figure 12. Bi-LSTM Model Classification Report

RNN Classification Report:

Figure 13 presents the classification report of the RNN model.

```
Classification Report of RNN Model

              precision    recall  f1-score   support

           0       0.00      0.00      0.00        26
           1       0.00      0.00      0.00        42
           2       0.97      1.00      0.99      2410

    accuracy                           0.97      2478
   macro avg       0.32      0.33      0.33      2478
weighted avg       0.95      0.97      0.96      2478
```

Figure 13. RNN Model Classification Report

CNN Classification Report:

Figure 14 displays the classification report for the CNN model.

```
Classification Report of CNN Model

              precision    recall  f1-score   support

           0       0.73      0.31      0.43        26
           1       0.85      0.67      0.75        42
           2       0.99      1.00      1.00      2410

    accuracy                           0.99      2478
   macro avg       0.86      0.66      0.72      2478
weighted avg       0.98      0.99      0.98      2478
```

Figure 14. CNN Model Classification Report

Model Performance Comparison:

Table 1. Model Performance Comparison

| Model	Accuracy	Precision	Recall	F1 Score
CNN	98.71%	98.50%	98.71%	98.49%
Bi-LSTM	98.59%	98.49%	98.59%	98.47%
GRU	98.02%	96.86%	98.02%	97.42%
LSTM	97.94%	96.76%	97.94%	97.30%
RNN	97.22%	94.59%	97.22%	95.88%

Table 1 shows that Five deep learning models are explored for sentiment analysis CNN, Bi-LSTM, GRU, LSTM, and RNN, and are assessed using accuracy, precision, recall, and F1 score. Out of the five models, CNN performed best with 98.71% accuracy and 98.49% F1 score due to its ability to extract spatial features (types of features in texts). This could be attributed to Bi-LSTMs ability to detect contextual relationships in both directions as the Bi-LSTM brought the accuracy close to S-LSTM at 98.59%. While accuracy remained similar (98.02% vs. 97.94%), GRU trained significantly faster than LSTM. RNN had the worst result with (97.22% accuracy). In summary, CNN and Bi-LSTM are the most favorable, and LSTM is outperformed by GRU in terms of resources.

Due to the efficiency of its convolutional filters in capturing local features, the CNN model surpassed all other models to get an accuracy of 98.71%. The CNN model captures code-mixed phrases such as "awesome drama" in Roman Urdu, addresses informal slang and typos through hierarchical feature learning, and overloads informant recognition of important sentiment words in short reviews without regard to their position. Moreover, it is more parameter-efficient compared to RNN-based models given our limited data (12,000 samples).
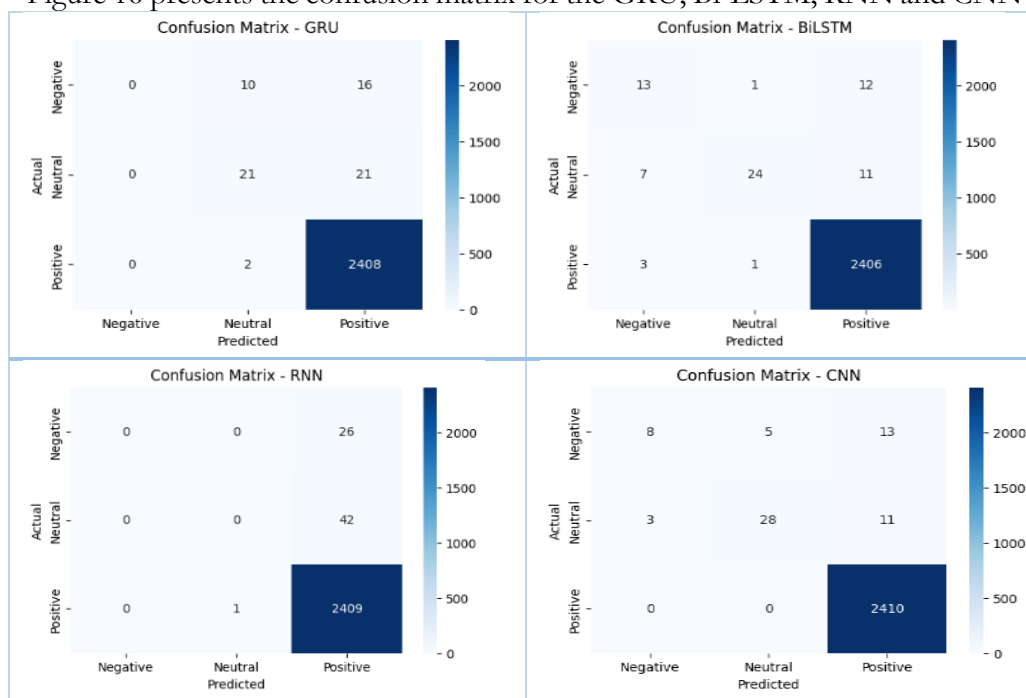
Confusion Matrices:

LSTM Confusion Matrix:

Figure 15 presents the confusion matrix for the LSTM model GRU, BiLSTM, RNN and CNN model Confusion Matrix
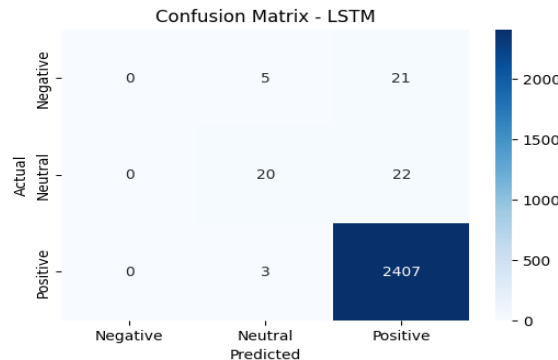
Figure 15. Confusion matrix for the LSTM model

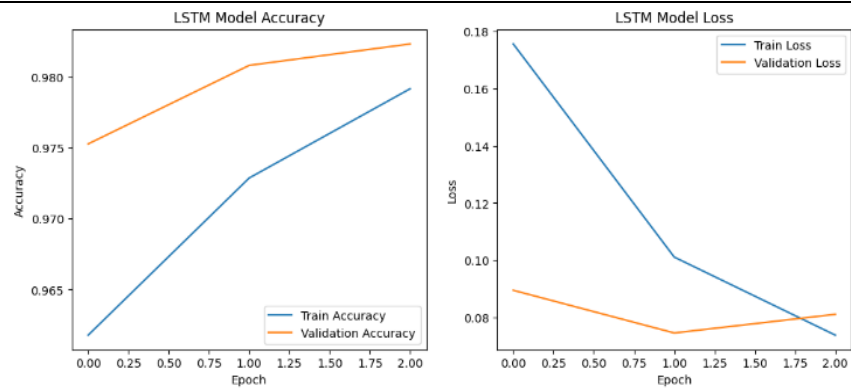Figure 16 presents the confusion matrix for the GRU, Bi-LSTM, RNN and CNN model

Figure 16. GRU, BiLSTM, RNN and CNN model
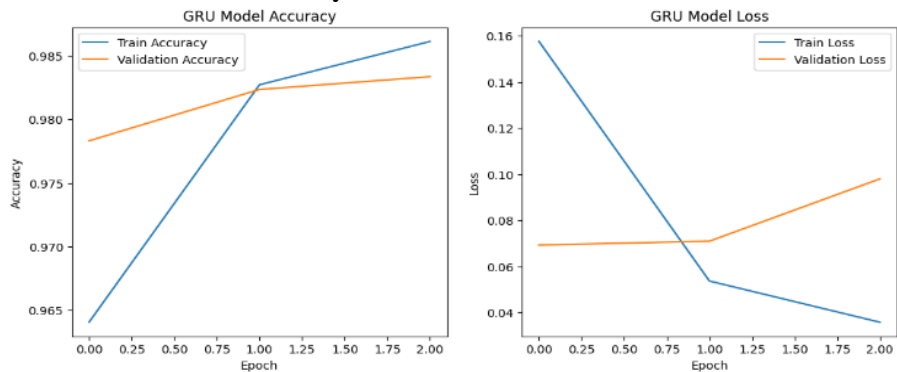
Visualization of Results:

Figure 17 illustrate loss and validation accuracy of different models.

Loss and validation accuracy graphs for LSTM, GRU, Bi-LSTM, RNN and CNN.
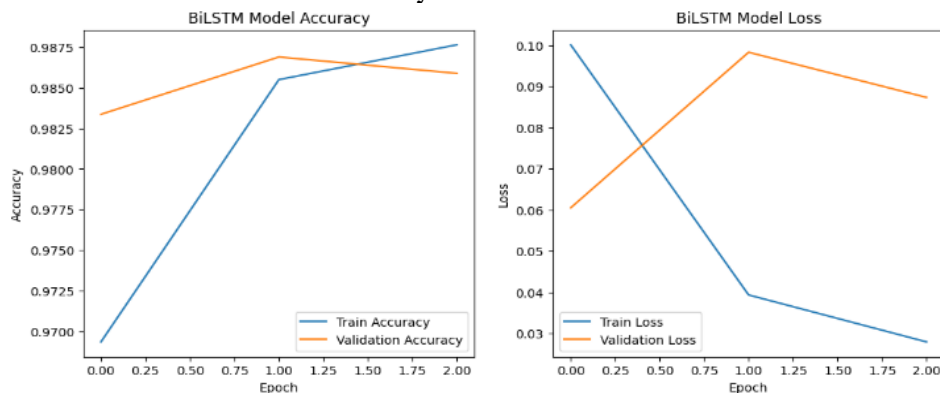
LSTM Loss and Validation Accuracy

Figure 17. LSTM Loss and Validation Accuracy

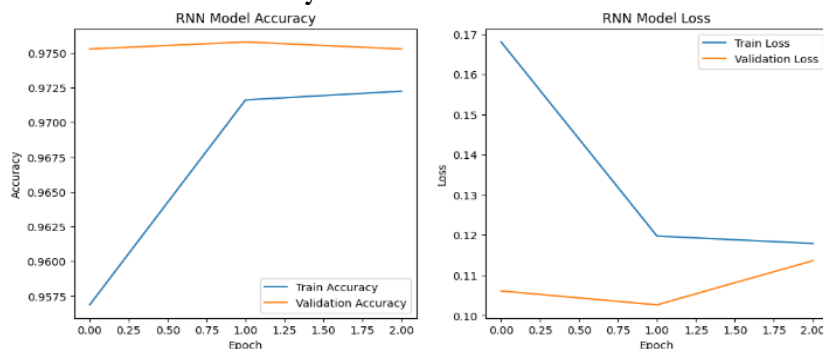**GRU Loss and Validation Accuracy:**

Figure 18. GRU Loss and Validation Accuracy
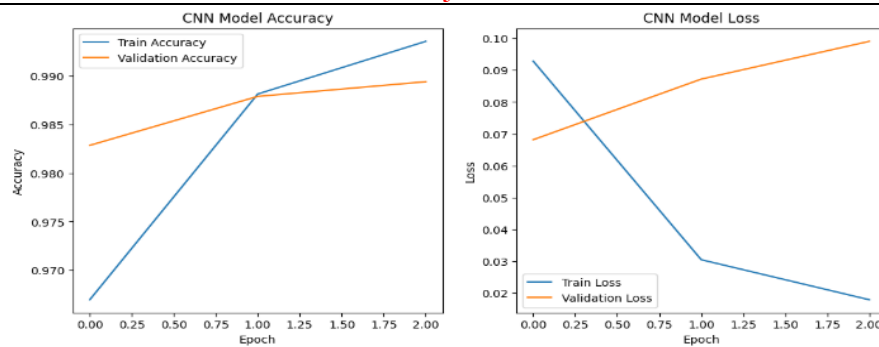
Bi-LSTM Loss and Validation Accuracy:

Figure 19. Bi-LSTM Loss and Validation

RNN Loss and Validation Accuracy:

Figure 20. RNN Loss and Validation Accuracy

CNN Loss and Validation Accuracy:

Figure 21. CNN Loss and Validation Accuracy

Discussion:

Five deep learning models CNN, Bi-LSTM, GRU, LSTM, and RNN were compared for sentiment analysis using four performance metrics: accuracy, precision, recall, and F1 score. As shown in Table 1 The CNN model emerged as the most effective, achieving the highest accuracy (98.71%) and F1 score (98.49%), due to its strong spatial feature extraction from text data. Bi-LSTM performed second-best, with 98.59% accuracy and a 98.47% F1 score, benefiting from its bidirectional ability to understand contextual relationships. GRU also showed strong results, achieving 98.02% accuracy and a 97.42% F1 score, making it more efficient than LSTM while delivering comparable results (LSTM accuracy 97.94% and F1 score 97.30%). RNN performed the weakest, with 97.22% accuracy and a 95.88% F1 score, likely due to its limitations in handling long-term dependencies. Overall, CNN is identified as the top model, followed by Bi-LSTM. GRU is appreciated for its computational efficiency, while RNN, though foundational, is outperformed by more advanced architectures. The study suggests that a combination of CNN and Bi-LSTM would be ideal for real-world sentiment analysis tasks.

CNN's superior performance can be attributed to its robust feature extraction capabilities, which allow it to detect patterns and critical keywords within text data effectively. Bi-LSTM also demonstrated strong results due to its bidirectional structure that analyses sequences from both directions, enhancing contextual understanding. However, models like RNN struggled with long sequences, likely due to vanishing gradient issues, confirming observations made in prior studies by Khan et al. [9]. Challenges encountered in this research included computational limitations, particularly during model training on large datasets, which required significant processing power and memory. Moreover, the mixed-language nature of the dataset introduced complexities, especially in handling informal slang and transliteration H. Ahmed et al [10].

These findings align with prior studies that highlighted the effectiveness of CNN and hybrid models for sentiment analysis L. Zhao and M. Sani [11]. However, this research extends those insights by comparing multiple architectures on a culturally specific and multilingual dataset, fulfilling the objective of identifying the most accurate model for sentiment analysis in Pakistani entertainment content.

Conclusion:

This research evaluated the performance of five deep learning models CNN, Bi-LSTM, GRU, LSTM, and RNN for sentiment analysis using accuracy, precision, recall, and F1-score metrics. The CNN model outperformed all others with 98.71% accuracy and a 98.49% F1-score, highlighting its strong ability to capture spatial features in text. Bi-LSTM also performed exceptionally well, with 98.59% accuracy and a 98.47% F1-score, due to its bidirectional analysis that captures context from both forward and backward sequences. GRU outperformed LSTM, achieving 98.02% accuracy and a 97.42% F1-score, demonstrating that it provides similar accuracy to LSTM but with lower computational demands. The RNN model showed the lowest performance, with 97.22% accuracy and a 95.88% F1-score, due to its difficulty in managing long-term dependencies caused by the vanishing gradient problem.

Overall, CNN and Bi-LSTM were identified as the most effective models for sentiment analysis, with GRU offering a more resource-efficient alternative to LSTM. RNN, while foundational in sequential modelling, has been surpassed by more advanced architectures. The study concludes that combining CNN and Bi-LSTM models would provide optimal results in real-world sentiment analysis tasks.

Future Work:

Future research will focus on developing hybrid models by combining CNN and Bi-LSTM, allowing the integration of CNN's spatial feature extraction with Bi-LSTM's ability to understand sequential patterns. This combination is expected to enhance sentiment classification by capturing both feature types simultaneously. The evaluation of advanced transformer models like BERT, GPT, and XLNet is recommended, as these models use self-attention mechanisms that efficiently handle long-range dependencies in text, potentially improving sentiment analysis accuracy. Expanding sentiment analysis to support multiple languages, including regional languages such as Sindhi and Urdu, along with domain-specific models in fields like finance and healthcare, will increase adaptability and accuracy for diverse applications. Developing real-time sentiment analysis systems that can efficiently monitor and process social media and customer feedback is a priority. This can be achieved by utilizing model quantization and knowledge distillation techniques to enhance processing speed without sacrificing model accuracy.

References:

[1]   Amna Malik, "Introduction, Growth and Future of Web Series in Pakistan: An Exploratory Study," *Pakistan Soc. Sci. Rev.*, vol. 4, no. 2, pp. 1085–1095, 2020, doi: 10.35484/pssr.2020(4-II)86.

[2]   M. K. Shivani Awasthi, Inderpreet Kaur, "A Comprehensive Analysis of the Current Methods, Challenges and Innovations in Sentiment Analysis," *J. Inf. Syst. Eng. Manag.*, vol. 10, no. 23, 2025, doi: https://doi.org/10.52783/jisem.v10i23s.3712.

[3]   P. V. Roopam Srivastava, P. K. Bharti, "Comparative Analysis of Lexicon and Machine Learning Approach for Sentiment Analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 3, 2022, [Online].   Available: https://thesai.org/Publications/ViewPaper?Volume=13&Issue=3&Code=IJACSA&SerialNo =12

[4]   H. S. A. Sani Muhammad, Ahmad Abubakar, "Sentiment Analysis of Hausa Language Tweet Using Machine Learning Approach," *J. Res. Appl. Math.*, vol. 8, no. 9, pp. 07–16, 2022, [Online]. Available: https://www.questjournals.org/jram/papers/v8-i9/08090716.pdf

[5]   M. B. Mustajeeb ur Rehman, "Sentiment Analysis on Disputed Territory Discrepancies Using Machine Learning-based Text Mining Approach," *VFAST Trans. Softw. Eng.*, vol. 11, no. 2, pp. 17–25, 2023, doi: https://doi.org/10.21015/vtse.v11i2.1486.

[6]   N. A. Lal Khan, Ammar Amjad, Hsien-Tsung Chang, "Multi-class sentiment analysis of urdu text using multilingual BERT," *Sci. Rep.*, vol. 12, no. 1, 2022, [Online]. Available: 10.1038/s41598-022-09381-9

[7]   M. Ali, S. Asghar, A. Mrhari, A. Ullah, U. U. Aleem, and R. Hassnae, "A Hybrid CNN-BiLSTM-Based Approach for Roman Urdu Sentiment Analysis Using Enhanced Roman Urdu Corpus," *https://doi.org/10.1142/S2717554524500097*, vol. 34, no. 03n04, Jan. 2025, doi: 10.1142/S2717554524500097.

[8]   I. U. Khan *et al.*, "A Review of Urdu Sentiment Analysis with Multilingual Perspective: A Case of Urdu and Roman Urdu Language," *Comput. 2022, Vol. 11, Page 3*, vol. 11, no. 1, p. 3, Dec. 2021, doi: 10.3390/COMPUTERS11010003.

[9]   A. K. Durairaj and A. Chinnalagu, "Sentiment Analysis on Mixed-Languages Customer Reviews: A Hybrid Deep Learning Approach," *Proc. 2021 10th Int. Conf. Syst. Model. Adv. Res. Trends, SMART 2021*, pp. 551–557, 2021, doi: 10.1109/SMART52563.2021.9676277.

[10]   D. D. Londhe, A. Kumari, and M. Emmanuel, "Challenges in Multilingual and Mixed Script Sentiment Analysis," *2021 6th Int. Conf. Converg. Technol. I2CT 2021*, Apr. 2021, doi: 10.1109/I2CT51068.2021.9418087.

[11]   S. K. Kruttika Jain, "A Comparative Study of Machine Learning and Deep Learning Techniques for Sentiment Analysis," *Conf. 2018 7th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir.*, 2018, doi: 10.1109/ICRITO.2018.8748793.