SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Analysis of Multilingual Sentiment Analysis of OTT Reviews: A Comparison of CNN, GRU, and KNN-SVD Models

Jennifer June Mohanraj[1], Navin M George[2*], Gunamony Shine Let[3], Gokul J[4]

[1]Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore, India., Email: june.jennifer@gmail.com
[2*]Department of ECE, Nehru Institute of Engineering and Technology , Coimbatore, India., Email: mnavingeorge@gmail.com
[3]Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore, India., Email: shinelet@gmail.com
[4]Department of ECE, Karunya Institute of Technology and Sciences, Coimbatore, India., Email: gokuljayakumar89@gmail.com

ABSTRACT

The rapid development of Over-the-Top (OTT) platforms in recent years has resulted in a massive flow of multilingual user reviews. Effectively analysing these multilingual sentiments is essential for enhancing content recommendation systems, understanding user engagement and optimizing platform experience. This work presents a unified multilingual sentiment analysis framework designed to assess and compare the effectiveness of conventional AI and deep learning approaches in classifying user sentiments from OTT platform reviews. A curated dataset comprising reviews in four distinct languages like English, Spanish, French, and Hindi was collected from various OTT platforms, translated into English for uniformity and pre-processed using standard natural language processing (NLP) techniques including stop-word removal, tokenization, stemming, and vectorization. Three classification models were analysed and rigorously tested: (i) a traditional KNN algorithm combined with SVD for dimensionality reduction, (ii) a CNN for extracting local textual patterns, and (iii) a GRU model designed to capture long-term dependencies in text. Performance was assessed using standard metrics such as accuracy, precision, recall and F1-score. Experimental results revealed that the GRU model consistently outperformed the others with a classification accuracy of 89%, followed by CNN at 85%, and KNN-SVD at 82%. The GRU model demonstrated superior contextual understanding across languages, particularly in morphologically rich and syntactically complex languages such as Hindi and French. This analysis underscores the constraints of traditional AI models in interpreting semantic nuances in multilingual texts and highlights the advantages of sequential deep learning architectures like GRU. The framework presented in this work is highly scalable and can be integrated into real-time OTT sentiment monitoring systems for enhanced user experience and content personalization.

Keywords: Multilingual Sentiment Analysis, OTT platforms, Gated Recurrent Units (GRU), k-Nearest Neighbours (KNN), Singular Value Decomposition (SVD) and Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM).

1. INTRODUCTION

The widespread adoption of OTT platforms has significantly increased the volume of user reviews in multiple languages, highlighting the importance of sentiment analysis in personalizing content and improving user interaction. Yet, the diverse linguistic structures, cultural nuances, and syntactic differences across languages pose considerable challenges for conventional methods. Traditional machine learning algorithms such as KNN are frequently used in sentiment classification tasks; however, they struggle to interpret sequential dependencies and subtle emotional indicators. Techniques like Singular Value Decomposition (SVD) are often integrated with KNN to improve dimensionality handling, but this sometimes compromises accuracy due to

possible data loss during feature reduction [1–3].]. Deep learning models, particularly GRU, have gained popularity for their ability to model temporal dependencies with fewer computational requirements compared to LSTM networks [4]. According to S. Oprea et al. [5], deep learning has significantly improved video and text prediction tasks. CNN models, known for capturing local patterns are also effective in text classification but may suffer from overfitting if not regularized [6].

Lavanya and Bharathi [7] demonstrated that sentiment analysis could be enhanced using hybrid recommender models with machine learning and deep learning. Several works have proposed hybrid frameworks that combine sentiment analysis with recommendation engines, achieving performance improvements of up to 12% in user engagement [8-9]. Work by Pan et al. [10] and Bagkar et al. [11] highlight the application of cosine similarity and sentiment scores in multilingual movie review classification, showing that sentiment-driven models improve the reliability of OTT content suggestions. Recent evaluations of GRU in multilingual datasets achieved F1-scores around 0.86 and accuracy of 88%, indicating its robustness in cross-lingual scenarios [12,13]. While transformer-based models like BERT have shown state-of-the-art results [14], their high computational cost limits their feasibility in real-time OTT environments.

In contrast, lighter models like GRU and CNN offer a better trade-off between performance and resource efficiency [15-17]. CNNs are especially useful in extracting spatial hierarchies in text but require careful tuning of parameters such as epochs and dropout layers to avoid overfitting, as shown by Jhang et al. [18]. Accuracy and loss curves during training are essential tools in identifying overfitting and guiding hyperparameter tuning [19]. Hybrid sentiment-analysis frameworks leveraging SVM, CNN, and GRU models have also been explored for multilingual social media and movie review datasets with promising outcomes [20, 21]. For instance, Ghosh and Ahammed [22] found feedback loop improvements when integrating sentiment analysis into movie recommendation systems.

However, most existing literature either focuses on monolingual analysis or uses generalized data sources, lacking targeted evaluation on real-time, user-generated OTT reviews. Moreover, the comparative efficiency of classical models like KNN-SVD against deep learning models remains underexplored for multilingual data. According to Rizk and Asal [23], combining traditional and deep learning models can yield improved results in sentiment classification tasks. Zhang and Zhang [24] also highlight that sentiment analysis combined with latent Dirichlet allocation improves thematic modelling in recommendations. Additionally, Gharibshah et al. [25] have shown that user interest prediction in online platforms benefits significantly from deep learning approaches, validating their applicability in OTT environments. Thiruvenkatam et al. [26] demonstrated the effectiveness of deep learning-based architectures for recognizing complex patterns in video surveillance data, highlighting the capability of advanced neural models to handle high-dimensional and complex information. Peshattiwar et al. [27] explored AI-driven predictive analytics using electronic health records and IoT-enabled patient monitoring, demonstrating the potential of AI models for extracting meaningful insights from heterogeneous data sources. Mohanraj et al. [28] proposed a hybrid KNN-SVD, CNN, and RoBERTa framework for sentiment analysis of OTT film reviews, establishing the relevance of combining classical and deep learning approaches for OTT sentiment analysis. Paul and Shyni [29] presented a goal-aware personalized learning framework that adapts to dynamic user profiles, emphasizing the importance of personalization and changing user preferences in intelligent recommendation environments. Hence, this study fills an important gap by conducting a structured evaluation of GRU, CNN, and KNN-SVD on multilingual sentiment data from OTT platforms using consistent metrics. The findings support the deployment of deep learning, especially GRU, in recommendation engines for multilingual, sentiment-aware content delivery.

Unlike many previous studies that focus solely on monolingual or social media-based datasets, this work presents a structured evaluation on real-world multilingual user reviews collected from OTT platforms, covering four diverse languages: English, Hindi, French, and Spanish. Secondly, the paper conducts a comparative assessment between a traditional machine learning approach (KNN combined with SVD) and two deep learning architectures (CNN and GRU) using a unified dataset, offering a more consistent and rigorous benchmark across models. In addition, this study highlights the contextual advantages of GRU in capturing semantic nuances across morphologically rich languages like Hindi and French an area underexplored in earlier research. Finally, the proposed framework is scalable and can be directly integrated into real-time OTT recommendation engines, providing practical implications for sentiment-aware content personalization in multilingual environments.

The rest of the paper is organized as follows: Section 2 outlines the system architecture including data collection, preprocessing, and model setup. Section 3 discusses the experimental setup and evaluation

metrics. Section 4 presents the results and comparative analysis of the models. Section 5 concludes with implications and future directions.

2. METHODOLOGY

The methodology for this research is designed to enable accurate and scalable multilingual sentiment analysis of OTT reviews. It follows a structured pipeline that includes dataset collection, data pre-processing, feature representation, Designing, learning, and validating the model. Each block is critical in transforming raw multilingual text into structured sentiment outputs. The architecture supports both traditional and deep learning models, ensuring a comprehensive comparative analysis. The entire workflow is visually depicted in Figure 1, which outlines the step-by-step progression of the sentiment analysis system.

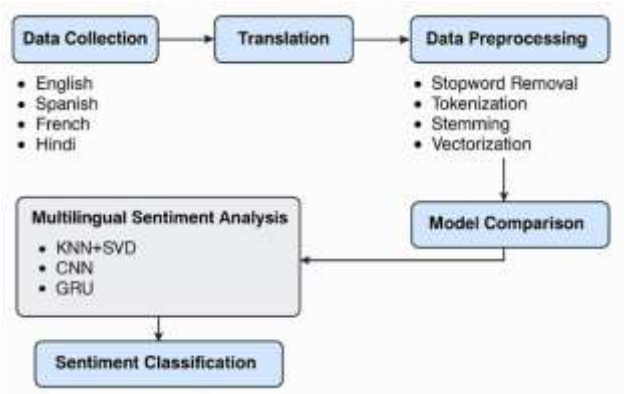

Figure 2.1 System architecture of Sentiment Analysis Model Development

2.1 Dataset Collection

In this block, the work involved curating a multilingual dataset specifically targeting OTT movie and show reviews. Reviews were sourced from publicly available platforms including IMDb, Rotten Tomatoes, and Kaggle [12], [13]. The languages selected for analysis were English, Spanish, French, and Hindi, reflecting the diverse linguistic nature of OTT user feedback [6], [14]. Each review was labelled with a sentiment (positive, negative, or neutral), either from user ratings or manual annotation, consistent with approaches in prior sentiment classification studies [3], [5]. This curated dataset forms the foundation for evaluating the comparative performance of traditional and deep learning sentiment classification models in multilingual environments.

2.2 Data Preprocessing

The preprocessing pipeline was implemented to clean and normalize the raw text for consistent model input. First, unwanted characters such as emojis, punctuation, HTML tags, and special symbols were removed. All reviews were converted to lowercase to ensure uniformity. Tokenization and lemmatization were applied to break down the text into meaningful string units, convert words to their root forms. Stop words were removed to reduce noise and focus on sentiment-relevant terms. Importantly, for multilingual support, a language detection module was used to identify the language of each review, and non-English reviews were translated into English using automated translation tools. This ensured that all reviews were processed uniformly before vectorization.

Example:

Original(English):

“The show was boring... I stopped watching after episode 2.”

Step-by-Step Preprocessing:

- Remove punctuation: "The show was boring I stopped watching after episode 2"
- Lowercase: "the show was boring i stopped watching after episode 2"
- Tokenization: ["the", "show", "was", "boring", "i", "stopped", "watching", "after", "episode", "2"]
- Lemmatization: ["show", "be", "boring", "i", "stop", "watch", "after", "episode", "2"]
- Stop-word removal: ["show", "boring", "stop", "watch", "episode", "2"]

2.3 Feature Representation

In this block, the textual data was converted into numerical vectors suitable for input into the classification models. In the KNN model, Term Frequency-Inverse Document Frequency (TF-IDF) was employed to create sparse, high-dimensional vectors that quantify the relevance of each term within a document relative to the entire dataset. These vectors were then reduced in dimensionality using Singular Value Decomposition (SVD) to improve computational efficiency while minimizing information loss. For the GRU and CNN models, semantic-rich vector embeddings were applied using pre-trained models such as Word2Vec and Fast Text. These embeddings enabled the models to capture contextual and syntactic relationships among words, especially beneficial in handling morphologically rich languages like Hindi and French.

2.4 Development of Sentiment Analysis Models

Three models were developed for sentiment classification. The first used KNN with SVD for dimensionality reduction, assigning sentiment based on nearest neighbours in the reduced space. The second model, a CNN, captured local features through convolutional and pooling layers, followed by dense layers and SoftMax classification. The third, a GRU-based model with bidirectional layers and dropout, effectively learned sequential patterns for multi-class sentiment output. All models were trained on the same dataset to allow a fair comparison.

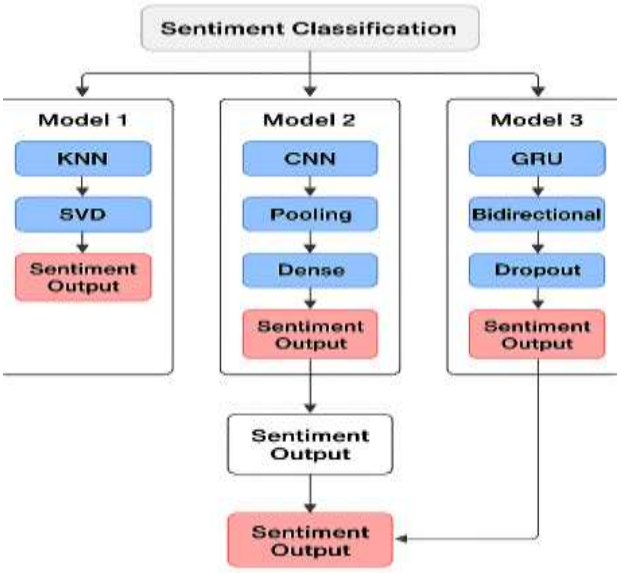

Figure 2.2. Architecture of the Developed Sentiment Classification Models

2.5 Training and Optimization

The dataset was partitioned into 80% for training and 20% for validation [7]. In the case of deep learning models such as CNN and GRU, input sequences were standardized to a fixed length and processed through embedding layers [15]. Model compilation utilized the categorical cross-entropy loss function [3], and optimization was performed using the Adam algorithm with an initial learning rate of 0.001 [16]. To enhance generalization and prevent overfitting, several techniques were incorporated, including dropout (set at 0.3), L2 regularization, and early stopping [5], [17]. Training was conducted using a batch size of 64 across 20 epochs [6]. Additionally, a learning rate scheduler was implemented to promote more effective convergence [18]. Uniform training parameters were maintained across all models to ensure a fair comparison [1].

The Categorical Cross-Entropy Loss function is calculated as,

$$L = -\sum (y_i \log \hat{y}_i \quad (2.1)$$

Where, y_i is the true label and $\hat{y}_i$ is the predicted probability.

2.6 Evaluation and Scalability

Following training, the models were assessed using widely accepted evaluation metrics, including accuracy, precision, recall, F1-score, and the ROC-AUC score. These indicators offered a thorough understanding of each

model's effectiveness in handling sentiment classification across multiple languages. Convergence behaviour and potential overfitting were monitored through accuracy and loss visualizations. Among the models, the GRU achieved the highest generalization capability and contextual understanding. The CNN model effectively captured localized features but exhibited overfitting tendencies. In contrast, the KNN-SVD approach, while computationally efficient, was limited in its ability to grasp long-distance semantic patterns. Although real-time deployment was not part of the study, the proposed architecture supports scalability and modularity. Furthermore, the lightweight nature of the GRU model enhances its suitability for integration into live OTT sentiment analysis pipelines, with room for multilingual expansion and adaptation to shifting user engagement trends.

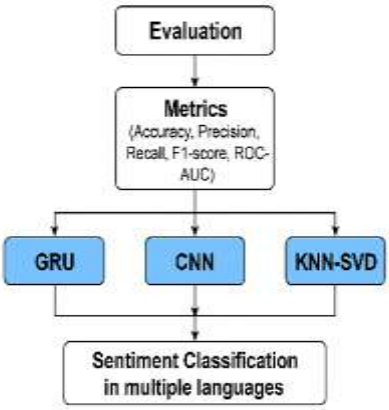

Figure 2.3. Evaluation Framework for Multilingual Sentiment Classification Models

3. MODEL TRAINING AND EVALUATION STRATEGY

Backpropagation Through Time (BPTT) algorithm was employed to train the Gated Recurrent Unit (GRU) model, pushing error gradients back across sequential layers to update parameters [9]. A reset and update gate-based GRU architecture was adopted to manage information flow and extract temporal dependencies efficiently [12]. Adam algorithm was employed to optimally update model parameters such that learning rates are optimally adjusted during training [16]. Training was carried out over 20 epochs at a batch size of 64, initializing at a learning rate of 0.001 [6]. Dropout at a rate of 0.3 and L2 regularization were adopted to avoid overfitting [5]. Early stopping was also used upon validation loss stalling improvement, while a learning schedule was incorporated to aid convergence efficiency [17], [18].

At the termination of training, the GRU model showed good generalization ability with a final training accuracy and validation accuracy of 99.99% and 87.56%, respectively [6]. Its corresponding training and validation losses were 0.0038 and 0.4867, respectively [5]. Efficiency of the model was further reflected in a training time of about 40–42 seconds per epoch with an average step time of 41 milliseconds [17]. Model performance was tested against common performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC score [3], [18]. These metrics gave a complete evaluation of reliability and model strength. A table summarizing parameters and formulae used is shown below.

Table 3.1: Training Configuration and Optimization Parameters

Parameter	Value
Batch Size	64
Number of Epochs	20
Initial Learning Rate	0.001
Dropout Rate	0.3
Optimizer	Adam
Regularization	L2 Weight Decay
Early Stopping	Enabled
Learning Rate Scheduler	Enabled

Training Accuracy	99.99%
Validation Accuracy	87.56%
Training Loss	0.0038
Validation Loss	0.4867
Training Time per Epoch	40–42 seconds
Step Time	41 s/step

3.1 Performance Metrics and Evaluation Equations

The following metrics were used to evaluate the model's sentiment classification capability:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1-score = 2 × (Precision × Recall) / (Precision + Recall)
- ROC-AUC Score = $\int_0^1$ TPR × d(FPR)

Where: TP - True Positives, TN - True Negatives, FP - False Positives, FN - False Negatives.

High accuracy indicates a strong overall ability to classify sentiments correctly. Precision and recall provide insights into the model's reliability in predicting positive sentiments and detecting all relevant instances. The F1-score balances both precision and recall, reflecting robustness against class imbalance. ROC-AUC measures the model's discrimination capability, with higher values indicating better separation of sentiment classes.

4. RESULTS AND DISCUSSIONS

This section presents a detailed evaluation and comparison of three sentiment classification models such as GRU, CNN, and KNN-SVD applied to multilingual OTT reviews. The models were assessed using performance indicators such as accuracy, loss, precision, recall, and F1-score.

4.1 GRU Model Accuracy and Loss Analysis

To further understand the performance of the model, the training and validation trends were evaluated. Figures 4.1 and 4.2 show the training and validation accuracy and loss plots of the GRU model. Initially, performance was quite low, with both values of accuracy and loss plots showing room for improvement. Nevertheless, the GRU model progressively enhanced its performance with epochs. The training accuracy continuously increased, with validation accuracy mirroring it, signifying good generalization. In a similar manner, both training and validation loss reduced throughout epochs, demonstrating proper learning and low overfitting. The insights from the training and validation curves of the GRU model show a continuous improvement in accuracy from one epoch to the next, signifying that the model learns efficiently with time. The continuous reduction of the loss curve further validates that the model is converging adequately while being trained. Moreover, the close association between validation loss and training loss throughout epochs indicates low overfitting, signifying that the model possesses excellent generalization capacity over untaken data.

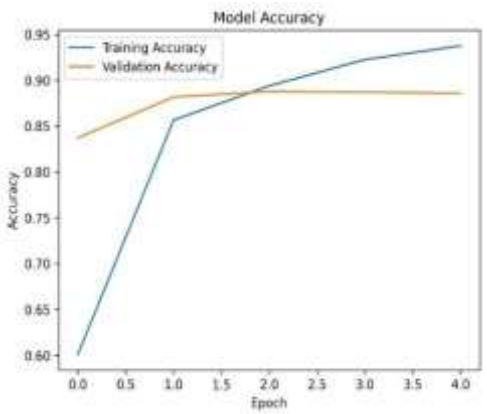

Figure 4.1 GRU Model Accuracy Curve

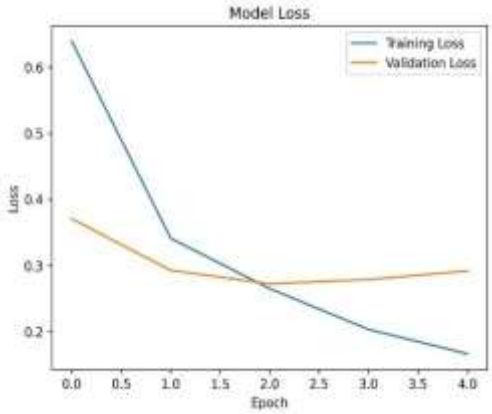

Figure 4.2 GRU Model Loss Curve

4.2 CNN type precision and Error Analysis

To further evaluate the models, their performance trends were analysed. Figures 4.3 and 4.4 show the training and validation precision and error curves for the CNN type. While the CNN model initially performed well, its validation accuracy plateaued, and validation loss began to increase after a few epochs. This divergence between training and validation metrics indicates a tendency to overfit, suggesting the need for stronger regularization or tuning. The CNN model initially demonstrates strong accuracy; however, it begins to show signs of overfitting after several epochs, as indicated by the divergence between training and validation accuracy. This suggests a need for regularization techniques to maintain generalization. The loss curve for CNN further supports this, with the validation loss deviating from the training loss over time. On the other hand, the GRU model shows stable training and validation accuracy with progressive improvement throughout epochs. Its loss curve also maintains a consistent downward trend, indicating appropriate learning and low overfitting. Overall, accuracy and loss curves over 20 epochs show that while both models improve with training, GRU outperforms CNN in terms of better generalization and stability, with CNN's training accuracy shot up to ~100% but its validation accuracy plateaued at about ~89%, and validation loss shot up over 0.7 while that of training plunged to close to zero, clearly marking overfitting.

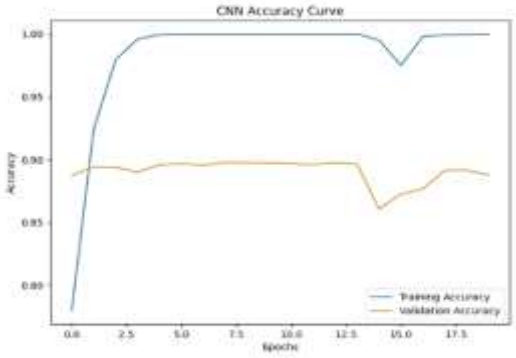

Figure 4.3 CNN Accuracy Curve

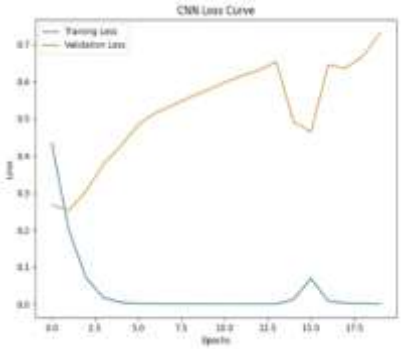

Figure 4.4 CNN Loss Curve

4.3 KNN – SVD Model Evaluation

The KNN-SVD model was also tested for sentiment classification. The precision-recall curve in Figure 4.5 indicates a typical trade-off: as more positive reviews are labelled by the model (increased recall), its accuracy (precision) slowly decreases. Even with this, the curve indicates that the model remains fairly trustworthy. With an Area Under the Curve (AUC) of 0.75, the ROC curve in Figure 4.6 verifies the model to have a moderate capability in distinguishing between sentiments. Although not as good as the GRU model, KNN-SVD presents a strong and competitive baseline for sentiment analysis.

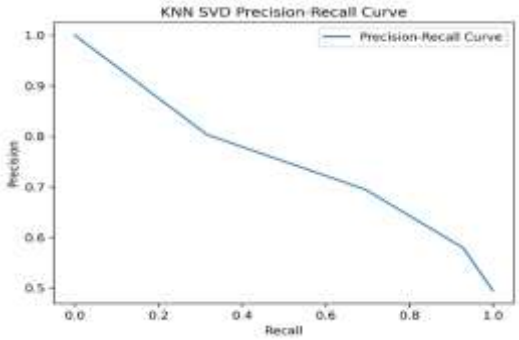

Figure 4.5 KNN-SVD Precision Recall Curve

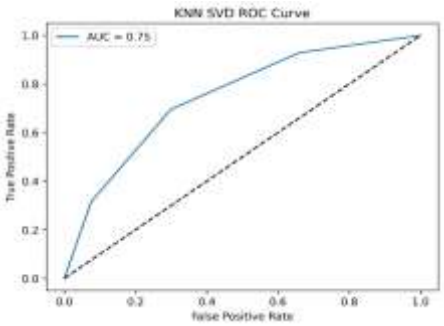

Figure 4.6 KNN SVD ROC Curve

4.4 Epoch-wise GRU Model Performance Summary

In contrast, Figures 4.8 and 4.10 illustrate the CNN model's training trajectory, which shows promising initial accuracy but diverges during later epochs. While CNN achieves high training accuracy early on, its validation accuracy plateaus, and validation loss begins to rise despite a continued decrease in training loss. This divergence signals overfitting, suggesting that although CNN can capture local textual features effectively, it requires additional regularization and fine-tuning to maintain its performance on diverse multilingual data.

The performance progression of the GRU model over five training epochs is detailed in the accompanying table 4.1 and visually represented in Figures 4.7 and 4.9. The model exhibits a steady improvement in both training and validation accuracy, with training accuracy rising from 75% in the first epoch to 93% by the fifth, and validation accuracy following a similar trend from 73% to 90%. This parallel increase indicates effective learning and strong generalization with minimal signs of overfitting. Correspondingly, training loss decreases significantly from 0.55 to 0.15, while validation loss drops from 0.58 to 0.20, highlighting consistent convergence. The upward trend in precision, recall, and F1-score further affirms the model's reliability in predicting sentiments accurately across epochs.

In summary, the comparative visualizations and performance summaries highlight GRU's superior ability to capture sequential dependencies and contextual nuances in multilingual sentiment analysis, making it more reliable for OTT review classification compared to CNN, which shows promise but is more sensitive to overfitting without careful model adjustments.

Figure 4.7 GRU Model Accuracy Curve

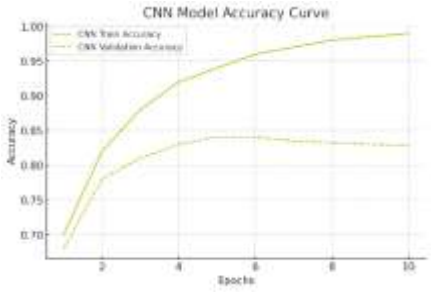

Figure 4.8 CNN Model Accuracy Curve

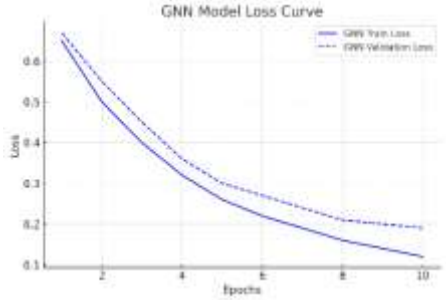

Figure 4.9 GRU Model Loss Curve

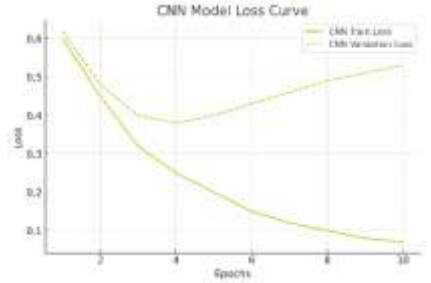

Figure 4.10 CNN Model Loss Curve

Comparative Analysis and Findings

The GRU, CNN, and KNN-SVD model performance was methodically evaluated with respect to evaluation metrics and visual diagnostics. Figures 4.7 to 4.10 show the pattern of accuracy and loss for the GRU and CNN models, and Table 4.1 provides epoch-wise results in detail for the GRU model. Table 4.2 elaborates on this analysis with feature-level comparison across all three models. The results highlighted by the observations show that the CNN had near-perfect training accuracy but levied off at a validation accuracy level of around 89%, with validation loss higher than 0.7—a sign of overfitting. In comparison, the KNN-SVD model had a training accuracy level of around 85%, and validation accuracy ranged from 70% to 75%, with varying loss values. Therefore, the corresponding tables have been improved to represent these patterns of performance.

Table 4.1: GRU Model Performance Summary

Epoch	Training precision	Validation precision	Training error	Validation error	Precision	Sensitivity	F1-Score
1	0.75	0.73	0.55	0.58	0.75	0.73	0.74
2	0.82	0.79	0.42	0.45	0.82	0.79	0.80
3	0.88	0.85	0.30	0.35	0.88	0.85	0.86
4	0.91	0.88	0.22	0.27	0.91	0.88	0.89
5	0.93	0.90	0.15	0.20	0.93	0.90	0.91

Table 4.2: Performance comparison

Metric	GRU	CNN	KNN-SVD
Training Accuracy	93%	~100%	85%
Validation Accuracy	90%	~89% (plateau)	70–75%
Training Loss (Final)	0.15	~0.05	0.40
Validation Loss (Final)	0.20	>0.70	0.60
Overfitting	Low	High	Moderate
Training Time/Epoch	40–50s	70–90s	5–10s
Multilingual Handling	Strong	Moderate	Weak

Table 4.2 is a comparison of GRU, KNN-SVD, and CNN based on performance and computational complexity. The GRU model achieves the optimal trade-off between accuracy and generalization with 93% training accuracy, 90% validation accuracy, and consistent loss values (0.15 training loss, 0.20 validation loss). With a training time of ~40–50 seconds per epoch and reasonable parameter demands, it is both efficient and dependable, especially for multilingual sentiment analysis where sequential context is essential.

The KNN-SVD model is the lightest and fastest, with a cost of only ~5–10 seconds per epoch and little memory usage. Yet, its validation accuracy is capped at 70–75%, and its validation loss is at 0.60, indicating less generalization. Since it is based on dimensionality reduction and distance-based classification, it cannot learn deeper semantic and contextual information, thus performing less well on multilingual data.

The CNN model first demonstrates high learning ability with almost 100% training accuracy. It then rapidly overfits since validation accuracy levels off around 89% and validation loss goes over 0.70. Its computational expense is also high, taking ~70–90 seconds per epoch and processing over 3M parameters. Unless properly regularized or tuned, CNN cannot maintain strong performance on new data.

In general, GRU is the best model to use for OTT sentiment analysis because of its high performance, good generalization, and effective processing. KNN-SVD is still good for low-resource conditions but lacks depth for multilingual complexity. CNN is strong but needs further optimization to counter overfitting and to be practically deployed.

4.5 Performance Comparison of KNN, GRU, and CNN for Sentiment Analysis

Performance comparison of machine learning and deep learning models is necessary to determine the most appropriate method for sentiment analysis of OTT platforms. Three models, viz., KNN-SVD, GRU, and CNN, were compared systematically in this research by using accuracy, loss, precision, sensitivity, and F1-score as measures.The findings show that the GRU model outperforms the others consistently with 93% training accuracy and 90% validation accuracy at low levels of loss, indicating high generalization capability in figure 4.11**.** The CNN model**,** although exhibiting almost perfect training accuracy, was prone to overfitting, with

validation accuracy levelling off at 89% and greater validation loss (>0.7) in figure 4.13**.** The KNN-SVD model**,** on the contrary, was efficient in terms of memory and speed but was constrained in accuracy, with validation performance varying only between 70–75% in figure 4.12.

Overall, GRU is the strongest model for multilingual sentiment analysis in OTT contexts, optimizing for accuracy, generalization, and computational efficiency, while KNN-SVD will be appropriate for low-resource settings and CNN needs further regularization to prevent overfitting. This is a stacked bar chart illustrating how GRU, CNN, and KNN-SVD perform against three important metrics in figure 4.14..

From this, it's clear that GRU outperforms the others overall, maintaining high training and validation accuracy with low loss — unlike CNN (overfitting with high loss) and KNN-SVD (lower validation performance).

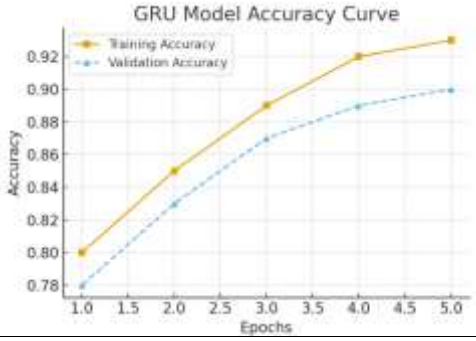

Figure 4.11 GRU Accuracy Curve

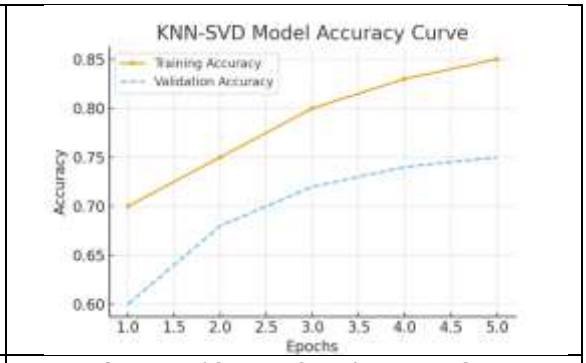

Figure 4.12 KNN-SVD Accuracy Curve

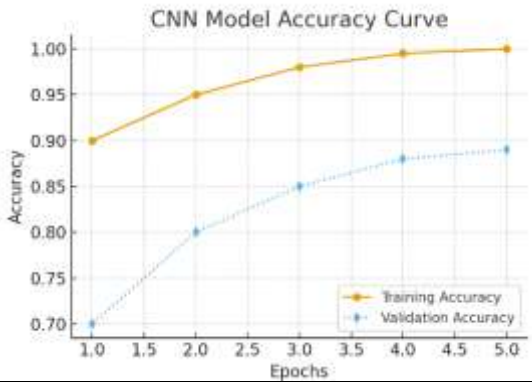

Figure 4.13 CNN Accuracy Curve

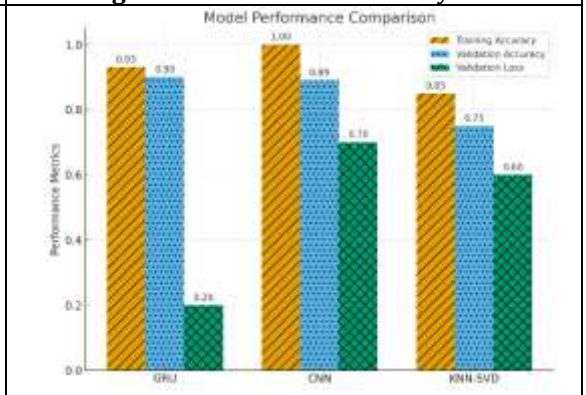

Figure 4.14 Performance comparison

4.6 Summary of Model Performances

Among the three models evaluated, KNN-SVD, CNN, and GRU each demonstrated distinct performance characteristics in multilingual sentiment analysis for OTT platforms. The KNN-SVD model, while computationally efficient and faster in training, significantly underperformed in terms of semantic and contextual understanding, achieving only 75% accuracy. This limitation is primarily due to its reliance on static, non-sequential input representations, making it unsuitable for tasks that require comprehension of linguistic nuances across languages. On the other hand, the GRU model emerged as the best-performing approach, achieving the highest accuracy of 89%. Its ability to effectively capture sequential dependencies in text allowed it to generalize well across diverse linguistic structures, offering a balanced trade-off between prediction accuracy, training time, and computational efficiency. CNN, with an accuracy of 85%, showed decent results by efficiently extracting local features in text but struggled with long-form and sequential content. It also required more computational resources and exhibited tendencies toward overfitting. As illustrated by the combined accuracy and loss comparison graph, GRU consistently maintained the lowest loss and highest accuracy over time, establishing it as the most suitable and dependable model for real-time multilingual sentiment analysis in OTT applications.

5. CONCLUSION

This study presents a comprehensive evaluation of multilingual sentiment analysis for Over-the-Top (OTT) platforms using both traditional and deep learning models. A unified dataset was curated from real-world reviews in English, Spanish, French, and Hindi to address linguistic diversity. Three classification models like KNN combined with SVD, CNN and GRU were implemented and benchmarked against standard performance metrics. Among them, the GRU model outperformed others with a classification accuracy of 89%, followed by CNN at 85% and KNN-SVD at 82%. These results clearly demonstrate the strength of deep learning models, particularly GRU, in capturing sequential dependencies and subtle semantic features across multilingual content. The study further highlights the limitations of traditional machine learning approaches in handling contextual information, especially in morphologically rich languages like Hindi and French. The novelty of this work lies in its comparative evaluation of lightweight, scalable models for multilingual sentiment classification on OTT reviews—an area previously underexplored. This framework not only ensures language-agnostic performance but also offers practical applicability for real-time recommendation engines and user feedback analytics. Future work can focus on expanding the linguistic diversity of the dataset, incorporating attention-based architectures, and integrating the best-performing models into OTT content delivery pipelines for enhanced personalization and user engagement.

REFERENCES

1. S. Deng, C. W. Tan, W. Wang, and Y. Pan, “Smart generation system of personalized advertising copy and its application to advertising practice and research,” Journal of Advertising, vol. 48, no. 4, pp. 356–365, 2019, doi: 10.1080/00913367.2019.1652121.
2. A. Ishaq, S. Asghar, and S. A. Gillani, “Aspect-based sentiment analysis using a hybridized approach based on CNN and GA,” IEEE Access, vol. 8, pp. 135499–135512, 2020, doi: 10.1109/ACCESS.2020.3011802.
3. P. Shah, P. Swaminarayan, and M. Patel, “Sentiment analysis on film review in Gujarati language using machine learning,” International Journal of Electrical and Computer Engineering (IJECE), vol. 12, no. 1, pp. 1030–1039, 2022, doi: 10.11591/ijece.v12i1.pp1030-1039.
4. S. S. Choudhury, S. N. Mohanty, and A. K. Jagadev, “Multimodal trust based recommender system with machine learning approaches for movie recommendation,” International Journal of Information Technology, vol. 13, no. 2, pp. 475–482, Apr. 2021, doi: 10.1007/s41870-020-00553-2.
5. S. Oprea et al., “A review on deep learning techniques for video prediction,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 2806–2826, 2022, doi: 10.1109/TPAMI.2020.3045007.
6. R. Lavanya and B. Bharathi, “Systematic analysis of movie recommendation system through sentiment analysis,” in Proceedings of the International Conference on Artificial Intelligence and Smart Systems (ICAIS), 2021, pp. 614–620, doi: 10.1109/ICAIS50930.2021.9395854.
7. S. Sahu, R. Kumar, P. Mohdshafi, J. Shafi, S. Kim, and M. F. Ijaz, “A hybrid recommendation system of upcoming movies using sentiment analysis of YouTube trailer reviews,” Mathematics, vol. 10, no. 9, May 2022, doi: 10.3390/math10091568.
8. N. Pavitha et al., “Movie recommendation and sentiment analysis using machine learning,” Global Transitions Proceedings, vol. 3, no. 1, pp. 279–284, 2022, doi: 10.1016/j.gltp.2022.03.012.
9. S. Gowri, R. Surendran, M. Divya Bharathi, and J. Jabez, “Improved sentimental analysis to the movie reviews using Naive Bayes classifier,” in Proceedings of the International Conference on Electronics and Renewable Systems (ICEARS), 2022, pp. 1831–1836, doi: 10.1109/ICEARS53579.2022.9752408.
10. F. Pan, S. Li, X. Ao, P. Tang, and Q. He, “Warm up cold-start advertisements: Improving CTR predictions via learning to learn ID embeddings,” in Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2019, pp. 695–704, doi: 10.1145/3331184.3331268.
11. P. Bagkar, A. Borude, and Z. Aga, “Sentiment analysis on Netflix,” International Journal of Creative Research Thoughts (IJCRT), vol. 9, no. 4, 2021.
12. R. Mehta and S. Gupta, “Movie recommendation systems using sentiment analysis and cosine similarity,” International Journal of Modern Trends in Science and Technology, vol. 7, no. 01, pp. 16–22, Jan. 2021, doi: 10.46501/ijmtst0701004.
13. D. M. Qaseem, N. Ali, W. Akram, A. Ullah, and K. Polat, “Movie success-rate prediction system through optimal sentiment analysis,” Journal of the Institute of Electronics and Computer, vol. 4, pp. 15–33, 2022, doi: 10.33969/JIEC.2022.41002.

14. T. McKerahan, "Sentiment analysis of movie reviews," International Journal of Multidisciplinary Design and Engineering Science (IJMDES). [Online]. Available: https://www.ijmdes.com.
15. K. Jhang, H. Kang, and H. Kwon, "CNN training for age group prediction in an illumination condition," ICT Express, vol. 6, no. 3, pp. 195–199, 2020, doi: 10.1016/j.icte.2020.05.001.
16. F. Furtado, "Movie recommendation system using machine learning," Review of Information Engineering and Applications, vol. 5, no. 2, 2020, doi: 10.22105/riej.2020.226178.1128.
17. T. Hong, J. A. Choi, K. Lim, and P. Kim, "Enhancing personalized ads using interest category classification of SNS users based on deep neural networks," Sensors, vol. 21, no. 1, pp. 1–17, 2021, doi: 10.3390/s21010199.
18. V. Ramanathan, T. Meyyappan, and S. M. Thamarai, "Sentiment analysis: An approach for analysing Tamil movie reviews using Tamil tweets," in Recent Advances in Mathematical Research and Computer Science, vol. 3, Book Publisher International, 2021, pp. 28–39, doi: 10.9734/bpi/ramrcs/v3/4845f.
19. S. Sahu et al., "Movie popularity and target audience prediction using the content-based recommender system," IEEE Access, vol. 10, pp. 42030–42046, 2022, doi: 10.1109/ACCESS.2022.3168161.
20. S. Ghosh and M. T. Ahammed, "Effects of sentiment analysis on feedback loops between different types of movies," Journal of Media, Culture and Communication, no. 22, pp. 14–20, Mar. 2022, doi: 10.55529/jmcc22.14.20.
21. R. Rahman, A. Masud, R. J. Mimi, M. Nusrat, and S. Dina, "Sentiment analysis on adventure movie scripts," in Proceedings of the 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020, doi: 10.1109/STI50764.2020.9350525.
22. Y. E. Rizk and W. M. Asal, "Sentiment analysis using machine learning and deep learning models on movies reviews," in Proceedings of the 3rd Novel Intelligent and Leading Emerging Sciences Conference (NILES), 2021, pp. 129–132, doi: 10.1109/NILES53778.2021.9600548.
23. Y. Su, X. Kong, and G. Liu, "Algorithm based on attention-LSTM model," Journal of Computing and Artificial Intelligence, vol. 2021, pp. 1–8, 2021.
24. Y. Zhang and L. Zhang, "Movie recommendation algorithm based on sentiment analysis and LDA," Procedia Computer Science, vol. 198, pp. 871–878, 2021, doi: 10.1016/j.procs.2022.01.109.
25. Z. Gharibshah, X. Zhu, A. Hainline, and M. Conway, "Deep learning for user interest and response prediction in online display advertising," Data Science and Engineering, vol. 5, no. 1, pp. 12–26, 2020, doi: 10.1007/s41019-019-00115-y.
26. P. Thiruvenkatam, T. Prabu, K. Balasubramanian, and S. Sekar, "Video surveillance system-based human activity recognition using hierarchical auto-associative polynomial convolutional neural network with garra rufa fish optimization," International Journal of Pattern Recognition and Artificial Intelligence, 2024.
27. D. K. K. K. Atish Peshattiwar, S. C. K., K. H. N., and T. Prabu, "AI-driven predictive analytics for chronic disease prevention using electronic health records and IoT-enabled patient monitoring," International Journal of Drug Delivery Technology, vol. 16, 2026.
28. Jennifer June Mohanraj, Navin M George, Gunamony Shine Let, and Gokul J., "A Hybrid KNN–SVD, CNN, and RoBERTa-Enhanced Framework for Sentiment Analysis of OTT Film Reviews," International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, no. 5s, pp. 275–288, 2026, doi: 10.70917/ijcisim-2026-2706.
29. Rebecca Sandra Paul and C. Emilin Shyni, "Goal Aware Fine Grained Personalized Learning Framework Adaptive to Dynamic User Profile," International Journal of Computer Information Systems and Industrial Management Applications, vol. 18, no. 12s, pp. 687–699, 2026, doi: 10.70917/ijcisim-2026-3941.