[1*]Kanak Pandit
[2]Harshali Patil
[3]Drashti Shrimal
[4]Lydia Suganya
[5]Pratiksha Deshmukh

Regular paper
Comparative Analysis of Deep Learning Models for Sentiment Analysis on IMDB Reviews

Journal of Electrical Systems

Abstract: **-** Within the domain of natural language processing, sentiment analysis assumes a fundamental position, facilitating the comprehension of societal perspectives and opinions, thereby playing a pivotal role in understanding public sentiment. In this study, an examination and comparison of deep learning architectures was conducted on IMDB movie reviews. We evaluated the performance of Basic Recurrent Neural Networks, Long Short-Term Memory (LSTM), Gated Recurrent Neural Unit (GRU), Bidirectional LSTM, Bidirectional GRU and 1D Convolutional Neural Networks (Conv1D) based on their training, validation, and testing accuracies. Our results indicate that while LSTM achieved the highest accuracy of 99.91% on the training data, GRU demonstrated superior performance (88.28%) on the validation dataset. Interestingly, Bidirectional GRU emerged as the top-performing model (87.54%) on the testing data, showcasing its robustness in generalizing to unseen instances. These findings highlight the importance of evaluating model performance across multiple datasets to assess their real-world effectiveness. Furthermore, our comparative analysis provides valuable understanding of the advantages and limitations of each model, offering practical guidance for selecting the optimal framework for sentiment analysis endeavors. Overall, this research contributes to the progress of such methodologies and deep learning approaches in natural language processing.

Keywords: LSTM, Deep Learning Models, IMDB Reviews, Comparative Analysis, Natural Language Processing.

I. INTRODUCTION

In an era dominated by vast amounts of textual data generated from various online sources, understanding the sentiments expressed within this data has become increasingly crucial. Sentiment analysis, alternatively referred to as opinion mining, involves the collection and examination of individuals' viewpoints, perspectives, and evaluations concerning diverse subjects, items, themes, and offerings [1]. As social media continues to gain widespread adoption, opinion mining has risen as a significant domain of investigation. Within this domain, sentiment analysis holds significant importance, aiming to uncover public sentiment and attitudes towards particular topics or events by analyzing user-generated text data on digital platforms. Nevertheless, the abundance of opinion data often lacks sufficient annotation, posing challenges for the development and training of opinion models. Hence, attention is directed towards addressing the issue of limited labeled data in opinion analysis [2]. The significance of sentiment analysis spans across diverse fields, including product evaluations, social media assessments, and more. This analytical approach holds pivotal importance in understanding consumer sentiment towards products, services, and brands, aiding businesses in making informed decisions and enhancing customer satisfaction. Additionally, sentiment analysis plays a crucial role in gauging public opinion on social media platforms, enabling organizations to monitor trends, assess feedback, and manage online reputation effectively. As for the objective of this study, it involves an evaluation comparing different deep learning architectures for sentiment analysis using IMDB movie reviews. Through assessing the effectiveness of various model structures, the aim is to determine the most efficient method for accurately categorizing reviews as positive or negative sentiment, contributing to advancements in sentiment analysis methodologies.

Deep learning (DL) is experiencing rapid growth within the realm of materials data science, with an expanding array of applications across various data types including atomistic, image-based, spectral, and textual data[3]. The objective of this research is to assess the efficacy of different deep learning architectures, such as Simple Recurrent

[1] *Corresponding author: Kanak Pandit , Computer Engineering Department, Thakur College of Engineering, Mumbai, India, kanakpandit17@gmail.com
[2] Harshali Patil , Computer Engineering Department, Thakur College of Engineering, Mumbai, India, harshali.patil@thakureducation.org
[3]Drashti Shrimal, Computer Engineering Department, Thakur College of Engineering, Mumbai, India, drashti.shrimal@thakureducation.org
[4]Lydia Suganya, Computer Engineering Department, Thakur College of Engineering, Mumbai, India, lydia.suganya@thakureducation.org
[5]Pratiksha Deshmukh, Computer Engineering Department, Thakur College of Engineering, Mumbai, India, pratiksha.deshmukh@thakureducation.org

Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Neural Unit (GRU), Bidirectional LSTM and GRU, and 1D Convolutional Neural Networks (1D ConvNets), for analyzing customer sentiments on IMDB reviews, ultimately identifying the most effective model for binary classification of positive and negative sentiment.

II. RELATED WORKS

Various terms are used interchangeably to refer to sentiment analysis, depending on its specific application domains, including analysis based on specific aspects, opinion-based assessment and impact evaluation. Additionally, it is often referred to as idea mining, and the terms sentiment, opinion, and impact are frequently used interchangeably. Emotion classification can be categorized into a pair of main strategies: machine learning-based and dictionary-based methods [4].

Initially, sentiment analysis studies predominantly relied on word embedding methods based on frequency like Bag of Words and TF-IDF, followed by the application of different machine learning algorithms like Stochastic Gradient Descent, Decision Trees for emotion classification. Subsequently, prediction-based approaches such as FastText, Doc2Vec, ELMo, Word2Vec and Global Vectors (GloVe) gained traction in sentiment analysis tasks. Moreover, "Bidirectional Encoder Representations from Transformers (BERT)" has emerged as a pivotal tool in various NLP tasks, including sentiment analysis, as evident in pre-trained models like CNN. This transition reflects a shift towards more advanced and nuanced methodologies for sentiment analysis, leveraging both traditional machine learning techniques and cutting-edge deep learning architectures.

Table 1: Studies on Sentimental Analysis on IMDB Dataset

Reference	Model	Accuracy
[5]	LSTM	89.9%
[6]	Artificial neural network having six layers (four hidden, one input, and one output)	91.9%
[7]	Bi-LSTM, Bi-GRU	91.98%
[8]	Bi-LSTM -CNN	92.05%
[9]	Multi-Layer Perceptron	86.6%

As indicated in Table 1, different neural network architectures based on deep learning were developed employing either identical or diverse neural network structures with multiple layers. Upon reviewing the studies, it becomes evident that employing multiple layers results in greater accuracy compared to single-layer architectures.

III. METHODOLOGY

a. Dataset

The dataset comprises 50,000 reviews sourced from the IMDB website, evenly divided between positive and negative sentiments, totaling 25,000 positive tweets and 25,000 negative tweets. The distribution of sentiment classes within the dataset is illustrated in Figure 1.

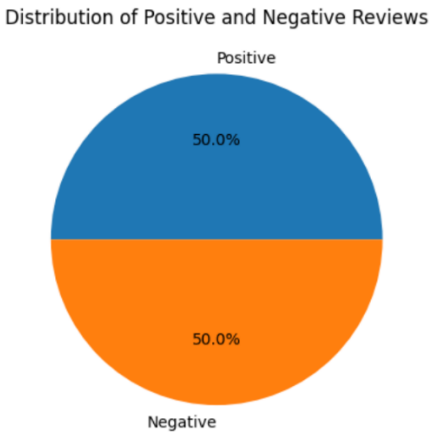

Fig. 1 Positive and Negative Reviews

Table 1 presents details regarding the features of the IMDB dataset.

Table 1 Description of IMDB Dataset

Feature	**Feature Type**	**Description**
Review	Text	User Comments
Sentiment	Text	Sentiment Classes of the Text(Positive, Negative)

Table 2 provides a more comprehensive overview of the dataset.

Table 2 Detailed Data of IMDB Review Dataset

Attribute	**Values**
Class Count	2
Review Count	50,000
Favourable Reviews	25,000
Unfavourable Reviews	25,000
Longest Review Size	13,704
Shortest Review Size	6
Number of Unique Words	1,24,252

Figure 2 displays the word cloud depicting the prevalence of positive and negative terms, derived from the 50,000 comments within the dataset. Notably, the positive-tagged data predominantly features words like “great”, while negative-tagged data contains terms such as “even”,”bad”, and “though”. Additionally, terms like “film” and “movie” are observed in both positive and negative word clouds.

Fig. 2 Word Clouds

b. Data Preprocessing

The text preprocessing steps were implemented using the TensorFlow Keras library. Firstly, a maximum review length of 200 words was established, along with specifications for training, validation, and testing sample sizes. The analysis focused on the top 100,000 words within the dataset to ensure comprehensive coverage. Subsequently, the Tokenizer class was employed to tokenize each review, converting textual data into sequences of integers. The tokenizer from 'Keras' was trained on the entirety of the available reviews stored in the variable `all_reviews`, facilitating the conversion of text into sequences of tokens.

A dictionary, `word_index`, was then created to map words to their corresponding integer indices, providing insight into the unique token count within the dataset. The total number of unique tokens was printed to provide a reference point for the analysis. Padding sequences was conducted to ensure uniformity in sequence length across all samples, with a predetermined maximum length of 200 words. This step is crucial for compatibility with deep learning models that necessitate fixed input sizes. Additionally, the text preprocessing operations encompassed the removal of HTML tags, special characters, URLs, email addresses, numbers, and punctuations. Tokenization facilitated the segmentation of text into individual tokens, followed by the conversion of all text to lowercase to standardize word forms. Removal of stopwords, commonly occurring but semantically insignificant words, was implemented to enhance the quality of the textual data.

Figure 3 illustrates the framework of the study, encompassing preprocessing of the dataset, techniques for word embedding, partitioning the dataset into training, testing, and validation sets, inputting these subsets into the designated classification models, and assessing the outcomes. In the experimentation phase aimed at assessing the efficacy of the models, the dataset was partitioned into 80 percent and 20 percent for training-testing and 10 percent for validation within the training set.

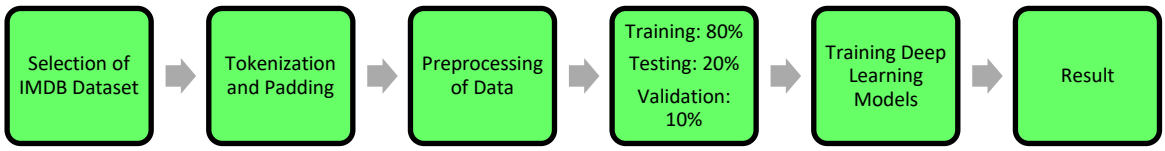

Fig. 3 Process Flow of our Experimentation

These preprocessing procedures were instrumental in cleansing and preparing the textual data for subsequent analysis within the realm of language understanding technology, ultimately contributing to the reliability and accuracy of the classification results obtained.

c. Word Embedding

It is a technique that assigns mapping each word in a sentence to a multi-dimensional vector, ensuring those terms sharing similar underlying meanings are positioned in proximity within this vector space. This method offers a closer sophisticated approach compared to the traditional bag-of-words model, as it encapsulates richer contextual information within a lower-dimensional space. Given that neural networks require numerical inputs, text must undergo a transformation known as text vectorization, converting it into a numeric representation for further processing. Word embedding finds application across a spectrum of Natural language processing (NLP) assignments, comprising tasks such as text categorization and sentiment assessment, among others[10].

In our experimentation, we have used word embedding as an embedding layer which acts as the first layer in the deep learning models. During training, the neural network learns the embeddings along with the other parameters

of the model using backpropagation. The embeddings are adjusted in a way that similar words in the input text corpus are mapped to nearby points in the embedding space, capturing semantic relationships between words. After training, the learned embeddings can be used to represent words in a meaningful way, where similar words have similar embedding vectors. These embeddings can subsequently serve as attributes for subsequent tasks like sentiment analysis, text classification, or any other natural language processing task.

d. Deep Learning

It represents a machine learning approach, leveraging algorithms to derive new insights from existing data. Employing at least one neural network, deep learning techniques can operate in supervised, semi-supervised, or unsupervised modes. Widely adopted across diverse fields, spanning computer vision, audio understanding, processing human language, and machine translation, deep learning has demonstrated remarkable success in tasks such as machine vision, voice recognition, social network filtering, and beyond[11].

i. Simple Recurrent Neural Networks

Also known as feedforward neural networks, they are fundamental architectures in deep learning. In a simple neural network, the data flows sequentially through layers of interconnected neurons, each layer transforming the input data until it produces the final output. It consists of an embedding layer followed by global average pooling and several dense layers with ReLU activation functions. The embedding layer converts input text data into dense vectors, while the subsequent layers learn hierarchical representations of the data to perform classification tasks. The network is trained utilizing “binary cross-entropy loss” and optimized using the “Adam optimizer” to minimize the difference between predicted and actual labels. Through the process of forward and backward propagation, the network learns to accurately predict outcomes on data not previously encountered by adjusting the weights of connections between neurons during training.

ii. Gated Recurrent Unit Networks

A variant of Recurrent Neural Networks, it is a slightly more simplified variation of the LSTM [12]. It comprises of two important gates: the update gate and the reset gate. These gates have critical functions in determining which data to retain and discard, as well as determining the extent to which past information should be discarded. One notable advantage of GRU over other RNN variants is its slightly faster processing speed due to its reduced number of vector operations. GRUs, while widely used and effective in recurrent neural networks, still pose challenges in understanding their intricate dynamics and how well they can capture underlying patterns. This lack of insight makes it challenging to predict their performance on different tasks and their ability to mirror the complex behaviours observed in biological systems[13].

iii. LSTM

It represents a tailored form of Recurrent Neural Networks (RNNs) crafted to tackle the issue of the vanishing gradient issue and capture extended correlations in successive data. Unlike traditional RNNs, LSTM units incorporate a sophisticated memory mechanism with three gates: the input mechanism, the forget mechanism, and the output mechanism. These mechanisms control the data flow through the cell state, allowing LSTMs to choose specifically what to keep or discard content over multiple time steps. By enabling the network to remember relevant information for extended periods while preventing the degradation of gradients during training, LSTMs have proven highly effective in different activities, covering areas like understanding human language, forecasting sequences over time, and recognizing spoken language. It is mainly designed for processing sequential data of varying lengths [14]. The LSTM's capability to retain sequences of information distinguishes it as a unique variant of recurrent neural networks[15].

iv. Bidirectional LSTM

Bidirectional LSTMs are a variation of the LSTM architecture that enhances its ability to capture temporal relationships in ordered data. Unlike conventional LSTMs, which handle data sequentially unidirectionally, bidirectional LSTMs process the input data in in both forward and reverse directions concurrently[16]. This bidirectional processing permits the model to grasp not only past but also subsequent circumstance, enabling it to make more informed predictions. By leveraging information from both directions, bidirectional LSTMs can better understand the context of each input token, resulting in enhanced effectiveness in tasks like natural language understanding, voice identification, and forecasting sequential data. Nonetheless, Bidirectional Long Short-Term

Memory (BLSTM) networks require considerable time and effort to tune, potentially more so than for feedforward networks[17].

v. Bidirectional GRU

Bidirectional Gated Recurrent Units (Bi-GRU) are a variant of recurrent neural networks that process input sequences in both forward and backward directions simultaneously. This dual-directional processing allows Bi-GRU networks to capture dependencies from past and future context, enhancing their ability to understand and model sequential data[18]. By combining information from both directions, Bi-GRU networks can learn more comprehensive representations, making them particularly effective for tasks involving sequence modeling. Unlike traditional GRU networks, which only consider past context, Bi-GRU networks provide a more holistic view of the input sequence, leading to improved performance in various applications.

vi. Conv1D

It is a potent category of deep learning architectures extensively employed in image identification and natural language processing tasks. Conv1D is specifically designed for processing one-dimensional progressive information, like chronological data or textual arrangements. It helps in situations when training data is less [19]. By applying convolutional filters across the input sequence, Conv1D captures local patterns and dependencies, enabling it to extract relevant features from the data efficiently. The use of pooling layers further reduces the dimensionality of the features while retaining important information.

IV. RESULT AND DISCUSSION

The outcomes of the conducted experiments on the various deep neural networks models showcased promising performance across different tasks.

Table 3 Performance of Models

Models	**Training Accuracy**	**Validation Accuracy**	**Testing Accuracy**
Simple Recurrent Neural Networks	0.997	0.8730	0.8642
GRU	0.9987	0.8828	0.8682
LSTM	0.9991	0.8816	0.8744
Bidirectional LSTM	0.9977	0.8692	0.8646
Bidirectional GRU	0.9976	0.8754	0.8754
Convolutional Neural Networks	0.9973	0.8622	0.8556

Table 3 shows the training performance, validation performance, and testing performance of different models for a sentiment analysis task. Training accuracy indicates the effectiveness of a model on the dataset it was trained with. Validation accuracy demonstrates how well the model performs on a separate set of data (validation set) employed to evaluate how effectively the model generalizes to new, unseen content. Testing accuracy shows how well the model performs on a completely different set of data (testing set) used to provide an unbiased estimate of model performance. LSTM shows the highest accuracy of 99.9% on the training data followed by GRU(99.8%), bidirectional LSTM(99.77%), bidirectional GRU (99.76%), convolutional 1D Model (99.73%) and least accuracy on training data is shown by Simple Recurrent Neural Networks(99.7%).

GRU shows the highest accuracy of 88.28% on validation data followed by LSTM(88.16%), bidirectional GRU(87.54%), Simple Recurrent Neural Networks(87.3%), bidirectional LSTM(86.9%) and convolutional 1D Model(86.2%). Bidirectional GRU shows the highest accuracy of 87.5% on testing data, followed by LSTM(87.44%), GRU(86.82%), bidirectional LSTM(86.46%), Simple Recurrent Neural Networks(86.4%) and convolutional 1D Model(85.56%). Figure 4,5,6,7,8 and 9 explain the Training vs Validation Accuracy and Loss. [20]

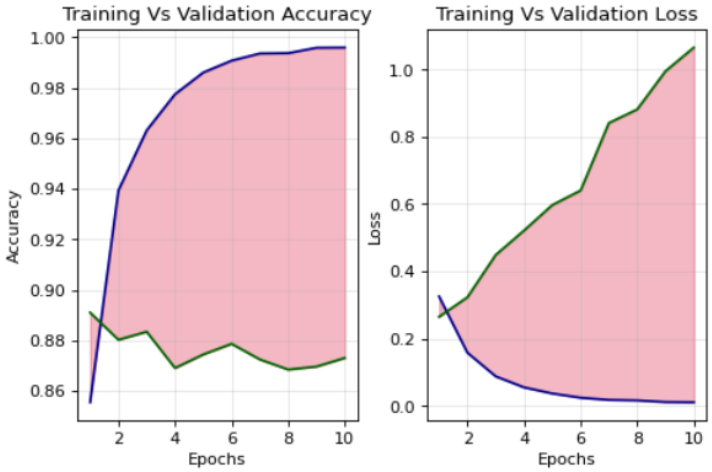

Fig. 4Training vs validation accuracy and loss of Simple Recurrent Neural Networks

Figure 4 shows the two graphs, one showing comparison between training accuracy and validation accuracy, and the other showing contrast between training loss and validation loss for Simple Recurrent Neural Networks. Both graphs depict performance over a set number of epochs, which are iterations of the training process. The training accuracy graph shows that the training accuracy for all models increases with increasing epochs. This suggests that the models are learning from the training data. The training loss graph shows a decrease in training loss as the number of epochs increases for all models.

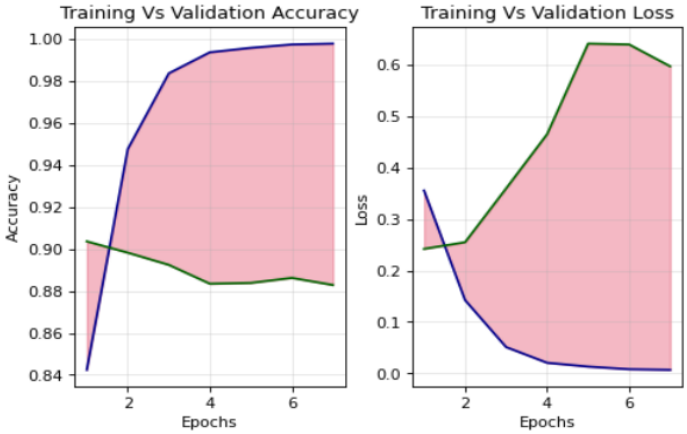

Fig. 5 Training vs validation accuracy and loss of a GRU neural network model

Figure 5 shows the accuracy reaches to approximately 100% after 6th epoch for training accuracy while the loss reaches to almost 0 after 6th epoch.

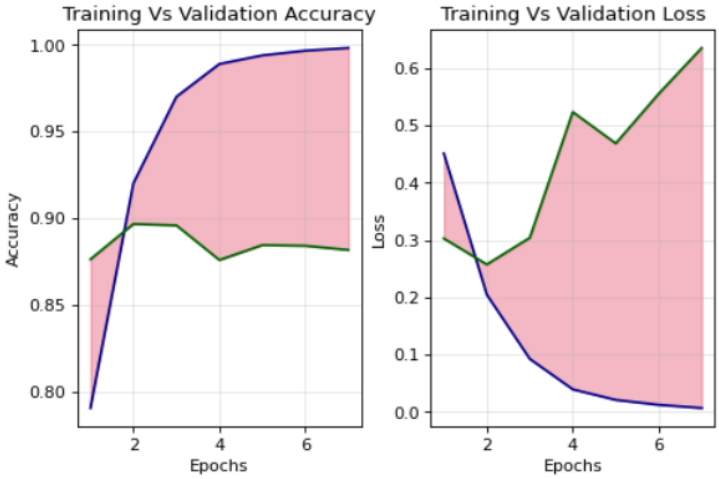

Fig. 6 Training vs validation accuracy and loss of a LSTM neural network model

Figure 6 shows a wavy behaviour in increase in Validation Loss which suggests that validation loss fluctuates over the epochs. The reasons might be Stochastic Gradient Descent or Hyperparameter Tuning.

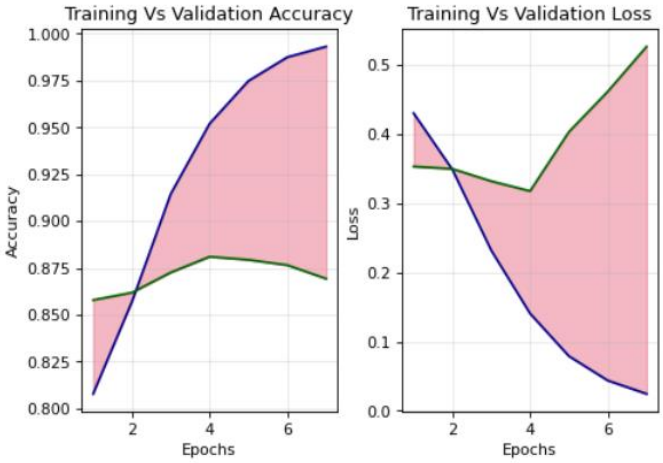

Fig. 7 Training vs validation accuracy and loss of a Bidirectional LSTM neural network model

Figure 7 shows that training accuracy takes some time to reach an accuracy of approximately 100% after 6th epoch as compared to that of LSTM and GRU.

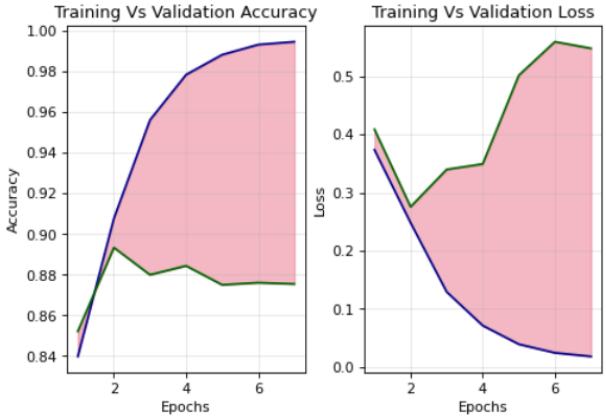

Fig. 8 Training vs validation accuracy and loss of a Bidirectional GRU neural network model

Figure 8 suggests that the bidirectional GRU model is learning from the data, but it may be worth employing techniques like early stopping or data augmentation to address overfitting and smoothen the validation loss trajectory.

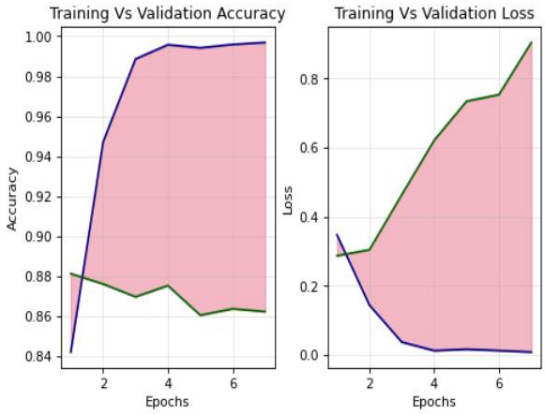

Fig. 9 Training vs validation accuracy and loss of a Convolutional 1D neural network model

Figure 9 shows training vs validation accuracy and training vs validation loss which is almost similar to that of other models.

IV. CONCLUSION

In conclusion, the comprehensive evaluation of various deep learning architectures utilized for opinion evaluation on IMDb reviews has provided valuable insights into their performance across different datasets. While LSTM achieved the highest accuracy on the training data, GRU demonstrated superior performance on the validation dataset. Interestingly, Bidirectional GRU emerged as the top-performing model on the testing data, showcasing its robustness in generalizing to unseen instances. These results underscore the importance of assessing model performance on multiple datasets to gauge their effectiveness in real-world scenarios. Additionally, the comparative analysis highlights the strengths and weaknesses of each model, providing valuable guidance for selecting the most suitable architecture based on specific task requirements. Overall, this study contributes to the understanding of deep learning approaches in sentiment analysis and offers practical insights for optimizing model performance in similar tasks. Additionally, sentiment analysis finds applications in finance, online commerce, and Blogger Centric Contextual Advertising Analyzer, among other domains[21].

REFERENCES

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

[2] Zhao, Z., Liu, W., & Wang, K. (2023). Research on sentiment analysis method of opinion mining based on multi-model fusion transfer learning. Journal of Big Data, 10, 155. [DOI: 10.1186/s40537-023-00837-x]

[3] Choudhary, K., DeCost, B., Chen, C., et al. (2022). Recent advances and applications of deep learning methods in materials science. npj Computational Materials, 8, 59. [DOI: 10.1038/s41524-022-00734-6]

[4] Başarslan, M. S., Kayaalp, F. (2023). MBi-GRUMCONV: A novel Multi Bi-GRU and Multi CNN-Based deep learning model for social media sentiment analysis. Journal of Cloud Computing, 12, 5. [DOI: 10.1186/s13677-022-00386-3]

[5] Qaisar, S. M. (2020). Sentiment Analysis of IMDb Movie Reviews Using Long Short-Term Memory. In 2020 2nd International Conference on Computer and Information Sciences (ICCIS) (pp. 1-4). Sakaka, Saudi Arabia. [DOI: 10.1109/ICCIS49240.2020.9257657]

[6] Shaukat, Z., Zulfiqar, A. A., Xiao, C., et al. (2020). Sentiment analysis on IMDB using lexicon and neural networks. SN Applied Sciences, 2, 148. [DOI: 10.1007/s42452-019-1926-x]

[7] Islam, M. S., Ghani, N. A. (2022). A Novel BiGRUBiLSTM Model for Multilevel Sentiment Analysis Using Deep Neural Network with BiGRU-BiLSTM. Lecture Notes in Electrical Engineering, 730(July), 403–414. [DOI: 10.1007/978-981-33-4597-3_37]

[8] Pimpalkar, A., & Raj R, J. R. (2022). MBiLSTMGloVe: Embedding GloVe knowledge into the corpus using multi-layer BiLSTM deep learning model for social media sentiment analysis. Expert Systems with Applications, 203, 117581. [DOI: 10.1016/j.eswa.2022.117581]

[9] Shaukat, Z., Zulfiqar, A. A., Xiao, C., Azeem, M., & Mahmood, T. (2020). Sentiment analysis on IMDB using lexicon and neural networks. SN Applied Sciences, 2(2), 148. [DOI: 10.1007/s42452-019-1926-x]

[10] Asudani, D. S., Nagwani, N. K., & Singh, P. (2023). Impact of word embedding models on text analytics in deep learning environment: a review. Artificial Intelligence Review, 56, 10345–10425. [DOI: 10.1007/s10462-023-10419-1]

[11] Li, D., & Du, L. (2022). Recent advances of deep learning algorithms for aquacultural machine vision systems with emphasis on fish. Artificial Intelligence Review, 55(5), 4077–4116. [DOI: 10.1007/s10462-021-10102-3]

[12] Rana, R. (2016). Gated Recurrent Unit (GRU) for Emotion Classification from Noisy Speech (Version 1). arXiv. [DOI: 10.48550/ARXIV.1612.07778]

[13] Jordan, I. D., Sokół, P. A., & Park, I. M. (2021). Gated Recurrent Units Viewed Through the Lens of Continuous Time Dynamical Systems. Frontiers in Computational Neuroscience, 15, 678158. [DOI: 10.3389/fncom.2021.678158]

[14] Bahad, P., Saxena, P., & Kamal, R. (2019). Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. Procedia Computer Science, 165, 74-82. [DOI: 10.1016/j.procs.2020.01.072]

[15] Moghar, A., & Hamiche, M. (2020). Stock Market Prediction Using LSTM Recurrent Neural Network. Procedia Computer Science, 170, 1168-1173. [DOI: 10.1016/j.procs.2020.03.049]

[16] Abduljabbar, R. L., Dia, H., & Tsai, P. W. (2021). Development and evaluation of bidirectional LSTM freeway traffic forecasting models using simulation data. Scientific Reports, 11, 23899. [DOI: 10.1038/s41598-021-03282-z]

[17] Zeyer, A., Doetsch, P., Voigtlaender, P., Schlüter, R., & Ney, H. (2017). A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition. In 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2462-2466). New Orleans, LA, USA. [DOI: 10.1109/ICASSP.2017.7952599]

[18] Wang, S., Shao, C., Zhang, J., et al. (2022). Traffic flow prediction using bi-directional gated recurrent unit method. Urban Information, 1, 16. [DOI: 10.1007/s44212-022-00015-z]

[19] Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing, 151, 107398. [DOI: 10.1016/j.ymssp.2020.107398]

[20] Al-qaydeh, N. (2021). IMDB Sentiment with Deep Neural Networks. Kaggle. [https://www.kaggle.com/code/naseralqaydeh/imdb-sentiment-with-deep-neural-networks]. Accessed 24 March 2024.

[21] Patil, H. P., & Atique, M. (2017). Applications, issues and challenges in sentiment analysis and opinion mining–a user's perspective. International Journal of Control Theory and Application, 10(19), 33-43.