Vol. 07, No. 4 (2025) 2519-2528, doi: 10.24874/PES07.04A.014

Proceedings on Engineering Sciences

www.pesjournal.net

A COMPREHENSIVE APPROACH TO SENTIMENT ANALYSIS ON IMDB DATASET USING THE EMBED-FLOW MODEL

Sahitya Satya
Ashok Kumar Shrivastava[1]
Deepak Motwani
Babita Tiwari

Received 11.03.2025
Revised 17.04.2025
Accepted 21.05.2025

Keywords:

Sentiment Analysis, GRU, K-fold cross- validation, Binary Classification, EMBED-FLOW

A B S T R A C T

In solving binary classification problems, deep learning models are very effective, and they are used in a variety of domains such as natural language processing and medicinal research. To achieve better accuracy, this research proposes a deep learning based novel work, a sentiment analysis model that specifically targets the Internet Movie Database (IMDB) dataset. The proposed model leverages embedding, convolutional layers, pooling, multilayer bidirectional gated recurrent units (GRU), dropout, and dense layers and achieves a validation accuracy of 89.34%, surpassing existing models. Experimental analysis on the IMDB dataset, coupled with k-fold validation logistic regression, optimized learning rates, and early stopping mechanisms, ensures robust performance. The proposed model can be described as Embedding Multi-layer Bi-directional GRU with Early Stopping and Dropout using Functional Logistic Regression Optimized for Workflow (EMBED-FLOW).

1. INTRODUCTION

Putting data into one of two groups is known as binary classification, and it's a basic machine learning challenge. Applications for this issue may be found in many domains, such as sentiment analysis (Haque et al., 2019), fraud detection, natural language processing, and healthcare diagnosis. The field of binary classification has changed dramatically over time as a result of the advancement of deep learning techniques, opening up new and more flexible and accurate solutions. Review classification of movies holds significant importance in several aspects. Firstly, it serves as a valuable tool for consumers making informed decisions, helping them gauge the quality and potential enjoyment of a movie before watching. From a business perspective, accurate sentiment analysis enables film studios and distributors to understand audience preferences, tailor marketing strategies, and optimize content creation. Moreover, review classification is integral to building recommendation systems, enhancing user experience by suggesting movies that align with individual's taste. It also helps in the area of natural language processing by using real-life examples to improve machine learning models that analyze feelings. Overall, the classification of movie reviews plays a crucial role in influencing audience choices, shaping industry practices, and advancing the capabilities of computational linguistics.
The increasing adoption of cloud computing addresses the need for higher processing power, while deep learning methods gain prominence in data analysis. One major application area for deep learning methods is sentiment

[1] Corresponding author: Ashok Kumar Shrivastava
Email: akshrivastava1@gwa.amity.edu

analysis, a task involving the classification of text data to estimate sentiment polarity. This is widely utilized in tagging and classifying comments on social media for different purposes. Deep learning, a sophisticated kind about artificial neural network technology that draws inspiration from the composition and operations of the human brain, is characterized by the incorporation of numerous hidden layers. It allows for more complicated and more powerful learning which is very much an analog process rather than a sequential digital process as the human brain functions. There has been a substantial use of deep learning techniques in such domains in the image processing. Text classification, speech recognition, and NLP. GRU (Sen & Chaturvedi, 2023) is a variant of LSTM, which is a Recurrent Neural Network, along with LSTM continued to be the family of the Recurrent Neural Network (RNN). Unlike LSTM, GRU employs a simpler architecture with fewer parameters, which leads to faster training times and reduced computational complexity. Further, GRU simplifies the information flow process by joining an input and forget gates of LSTM into an update gate. This simplification not only enhances computational efficiency but also helps mitigate the vanishing gradient problem more effectively. Despite its simplicity, GRU demonstrates comparable performance to LSTM in many sequence modeling tasks while being more computationally efficient, making it an attractive choice for various applications, including sentiment analysis.

Multi-layer bidirectional GRUs (Cheng et al., 2020) excel over conventional single-layer GRU (Sarika, 2020) and Multi-layer bidirectional LSTM (Başarslan & Kayaalp, 2023) models due to simultaneously detect past and future contextual information. Through the forward and backward integration of numerous layers, these models provide a more profound comprehension of the input text, enabling more nuanced feature extraction and better representation learning. Consequently, they exhibit superior performance in sentiment analysis tasks by effectively encoding the intricate relationships and dependencies present in natural language data.

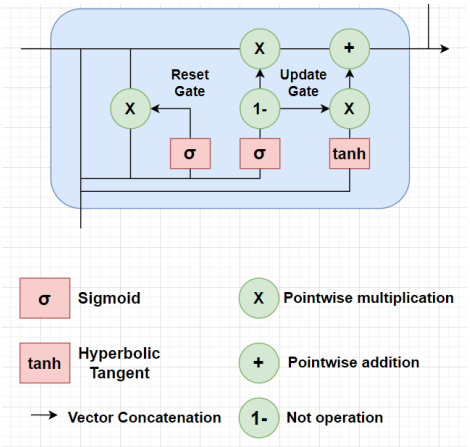

Figure 1. Working of GRU (Sarika, 2020)

In Figure 1 Gated Recurrent Unit (GRU), sequences of input text are processed using two primary components: i) Reset Gate**:** Before integrating new data, it decides how much of the old data should be deleted. It determines a reset gate vector by integrating the current input with the prior hidden state using a sigmoid activation function. Depending on the incoming input, the resultant values, which range from 0 to 1, determine whether certain portions of the previous concealed state should be kept or deleted. ii) Update Gate: This part determines the extent to which the incoming input should change the previous concealed state. But it also produces update gate vector values between 0 and 1 with a sigmoid activation function. We make the computation of the vector with respect to present input and previous hidden state. It creates its updated hidden state by updating the hidden state with the new input from the hidden state before. For Output Calculation**,** GRU creates a candidate hidden state by processing the reset data and the new input after determining the Reset Gate and Update Gate. The final output is then produced by the Update Gate controlling the percentage for this candidate state that is combined with the prior concealed state.

GRU's operations include, i) Sigmoid Function: Compresses values into the range [0, 1], enabling the model to manage information flow by selectively retaining or discarding details. ii) Hyperbolic Tangent (tanh): Maps values into the range [-1, 1], encoding input data into a compact form that represents both positive and negative information effectively. iii) Pointwise Multiplication: Pointwise multiplication involves element-wise multiplication of corresponding elements in two vectors or matrices, enabling the model to selectively emphasize or diminish specific features based on learned weights. iv) Pointwise Addition: Pointwise addition involves element-wise addition of corresponding elements in two vectors or matrices, enabling the model to combine information from different sources or layers while preserving their original dimensions. v) Not Operation: The "not" operation in GRU typically refers to the complement of a gate's activation, indicating the inhibition or suppression of certain information flow, which helps control the update, reset, or output mechanisms of the GRU unit.

The following are the main goals of conducting this study:

- To analyze the impact of activation functions (Farzad et al., 2019) and dropout (Cheng et al., 2017) and other regularization techniques on classification accuracy.
- To propose a deep learning model which work best in real-world binary classification applications such as review classification.
- The aim of this study is to add to the current conversation within the deep learning community and offer useful advice to practitioners and researchers who are tackling binary classification issues.

2. RELATED WORK

2.1 History of Movie Review and its advancements

The evolution of Film Criticism Film critique emerged in the early 1900s with the rising popularity of film media. Newspapers started hiring professional critics to provide in-depth analyses, moving beyond mere entertainment value (Sluga, 2023). New forms of cinema analysis eventually became commonplace in well-known periodicals. With critics like Gene Siskel and Roger Ebert creating programs that not only evaluated movies but also spoke with performers, film criticism has become even more famous in the current day. The main objective of the majority of review media is to explain the idea of a movie as well as its artistic or entertainment value. This is frequently done using grading systems such as grades, numerical scales, or "thumbs up" and "thumbs down" signs. Growth of Internet Film Criticism Online blogs made it possible for people to express their thoughts to a wider audience and signaled the beginning of the usage of digital media for cinema criticism. By adding videos, cut-scenes, animations, and actor viewpoints in film criticism, websites such as YouTube further broadened this audience. Plot flaws, scientific realism, sequel theories, and other facets of critique became the subject of specialized websites that arose. Certain websites customized their analysis to offer material recommendations for parents who are worried. The audience was able to voice their opinions through user-generated material on movie reviews, which was usually accompanied by textual comments and comparative grading systems. However, the modern film criticism industry has shown gender bias, often favored male reviewers and resulted in fewer representations of women. For example in Time magazine or NPR, 70–80 percent of the contributors are male reviewers. The first change brought about by the internet was that fewer women began working as newspaper film reviewers, and on these accounts there were fewer female opinion columnists. Men retained prevalent roles, becoming the dominant voice in film reviews (Onalaja et al., 2021).

A Brief History of Sentiment Analysis Hatzivassiloglou and McKeown's 1997 analysis of adjective phrases in stock market news yielded 90% sentiment accuracy, marking the beginning of sentiment analysis research. In 2004, Pang and Lee determined the polarity of sentences with an accuracy of 86.4% using machine learning, especially a Naive Bayes model (Kolasani & Assaf, 2020). Investigating Machine Learning Instruments for Sentiment Analysis Finding relevant information in reviews and classifying it as either positive or negative attitudes are the main goals of ongoing study. In a study the Amazon product reviews was used to predict ratings ranging from 1 to 5 stars in order to evaluate lexicon-based and machine learning models for text sentiment classification. Their results indicated that machine learning models generally outperformed lexicon-based approaches. The reported 90% accuracy is achieved by a logistic regression model, however, the study talks about a few challenges with processing emojis (Nguyen et al., 2018). In a different research, Yelp review variables were retrieved using logistic regression, a popular technique for predicting whether a review will be approved. The study emphasized the importance of features like complex sentences expressing substantive detail and varied sentiment ranges for achieving higher quality text sentiment analysis (Yao et al., 2018).

2.2 Historical Evolution of Deep Learning Models

Over the past ten years, deep learning has significantly advanced due to major developments in neural network architecture. For binary classification, feedforward neural networks and basic logistic regression models were initially employed. However, the introduction of deep neural networks having multiple hidden layers revolutionized the field. In this section we will explore the simplest of the neural networks and then convolutional neural networks (CNNs), recurrent neural networks (RNNs or LSTM), their respective variants (Bodapati et al., 2021). The investigation into the recurrent neural network (RNN) architectures towards binary classification problems, increasingly complex designs including multi-layer including bidirectional neural networks, gated recurrent units (GRUs), as well as long short-term memory (LSTM) have replaced traditional RNNs. Notably, studies consistently demonstrated how multi-layer and bidirectional GRU architectures outperform conventional RNNs in terms of both learning capacity and model performance. The inherent ability of multi-layer networks to capture hierarchical features and bidirectional architectures to consider contextual information from both past and future time steps contributes to improved sequence modelling. This literature trend underscores the importance of leveraging advanced RNN architectures, specifically multi-layer and bidirectional GRUs, for enhanced accuracy and effectiveness in binary classification tasks. Numerous strategies to improve the precision and resilience in binary classification models have been thoroughly investigated by researchers. This is important to guarantee both the robustness and validity of the proposed model and, for that reason, K fold cross validation, a method of dividing a data set in K proportional groups, was used for training and validation (Moss et al., 2018). This technique aids in mitigating issues related to data partitioning and provides a more reliable assessment of model performance. Additionally, optimizing learning rates, implementing early stopping mechanisms, and incorporating Conv1D layers (Zyout, 2023) along with MaxPool1D layers (Tao & Wu, 2024) have been investigated. Adjusting learning rates allows for more efficient convergence during training, early stopping prevents overfitting, and Conv1D layers with MaxPool1D facilitate feature extraction and

dimensionality reduction. The synergistic application of these techniques has been shown to significantly enhance the accuracy and generalization capabilities of binary classification model, showcasing the importance of a comprehensive approach to model optimization in literature.

2.3 Impact of Regularization Techniques to mitigate Overfitting

A model that remembers the training data excessively well—capturing noise and certain patterns that might not transfer to fresh, unseen data—is said to be overfit. When the model comes across fresh examples, this may result in subpar performance (Yenter, 2017). Regularization strategies like i) Batch Normalization: This method normalizes a neural network layer's input to have a unit variance and zero mean. This promotes improved generalization by stabilizing and speeding up the training process (Bjorck et al., 2018). ii) Dropout: To make sure the model doesn't rely too much on any one neuron during training, dropout randomly removes neurons. It renders the model more robust and further increases its generalization to unseen inputs. iii) Early Stopping: Early stopping means to stop learning given a drop on the performance score on validation set. This method trains the model so that it won't memorize the training data before it starts training on it in an attempt to force the model to not overfit. iv) Embedding: Embedding refers to the mapping of categorical variables (such as words in NLP) into continuous vector spaces (AlSurayyi et al., 2019), help deep learning models become more accurate and better at classification, reduce overfitting, and enhance their capacity for generalization.

3. METHODOLOGY

3.1. Dataset

Among many other datasets, the IMDB dataset (Tripathi, 2020) is a very popular benchmark dataset in the presence of sentiment analysis, especially for the binary classification tasks. Rather than some specific portion of the dataset. In this research, we use all of the IMDB dataset that includes 50,000 movie reviews. There is a set of 25,000 positive and 25,000 negative reviews for training and testing respectively.

It is important to note that no alterations were made to the original dataset, and the entire dataset was used for analysis without skipping any reviews or altering the content in any form. This approach ensures the integrity of the dataset and maintains the original distribution of reviews, thus preserving the dataset's inherent characteristics. This dataset is appropriate for training and assessing sentiment analysis models since it provides a wide variety of reviews from different genres and historical periods. In order to replicate the natural language used by reviewers, each review is usually represented as a text document with varied lengths and writing styles.

3.2. Data Preprocessing

Proper data preprocessing plays an important part in text pre-processing for training deep learning models. In this research, several important preprocessing steps have been performed to guarantee that the initial data is organized, normalized as well as improved for the learning purpose. The preprocessing pipeline starts with text tokenization and effectively, every review is split into individual tokens (usually words). After that there is integer encoding, where each token is assigned to a unique integer based on its frequency or position in the vocabulary. To provide coding uniformity for input size across all samples, we apply the technique of sequence padding. Shorter string inputs are zero padded, and longer string inputs are truncated, keeping the most central piece of information in each review. Furthermore, Label Encoding is performed to transform the categorical Sentimental label (Positive or Negative) to binary numerals format (1 or 0), which can be understood by the classification model. This binary encoding is also aligned with the type of problem, which is binary sentiment classification problem. Additional, preprocessing includes the conversion of all text to lowercase to decrease vocabulary size and additional to improve generalization and removal of punctuation and HTML tags, to eliminate noise. Stop words are let in to semantically know that they are often important sentiment. Concerning training of the model, we specify train time hyperparameters like optimizer (i.e. Adam), batch size (i.e. 64), and number of epochs (i.e. 10). These was hand-picked from the beginning trials - training time vs performance of the model. Another preprocessing issues which include shuffling the data and make reproducible via using random number initialization and split the dataset into a training and a validation corpus.

3.3. Architecture of proposed model

The below depicted model in Figure 2 is a Functional Embedding multi-layer bi-directional GRU Logistic Regression Conv1D MaxPooling1D Model with dropout with K-fold validation with Learning Rates with Early Stopping(EMBED-FLOW).

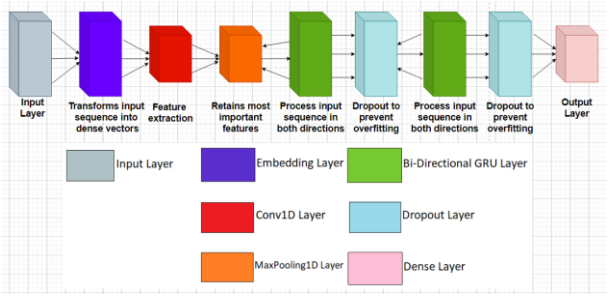

Figure 2. Architecture Diagram of the proposed model (EMBED-FLOW)

Different layers of the proposed model in Figure 3 includes: i) Input Layer: The input layer forms the form of the input data, and this means that the number of input nodes is determined. Since the input dimension in this model is set to 200, it anticipates an input sequence of integers with a length of 200. It converts the input data into a format that the neural network's later layers can understand. In this instance, it transfers word-representative integer sequences to the embedding layer. By forming connections with the neural network's later layers, the input layer permits data to go through the network both forward and backward (i.e., during training and inference). ii) Embedding Layer: It is crucial in natural language processing (NLP) tasks where words are represented as dense vectors. The Embedding layer which is used in this model is the default Keras Embedding layer and no other pre-trained Embeddings are used. It has an input dimension (vocabulary size) of 88,587. The dimension should be 256, and the last element of the dimension or vocabulary implies the number of unique words in words and represents the fact that this model will process 88,587 unique words. It enables the model to acquire continuous word representations, incorporating contextual information and semantic links. It learns to represent words in a dense vector space by modifying its weights during training in response to the input data. Words with comparable representations in this embedding space are those that share semantic similarities or commonly appear together in the dataset. Better performance on tasks like sentiment analysis and natural language processing is made possible by this approach, which enables the model to record meaningful associations between words. iii) Conv1D Layer: The Conv1D layer introduces convolutional operations, typically used for image processing, to NLP models. In this context, it helps capture local patterns and features within the embedded sequences. By applying convolution, the model can identify specific features or patterns within a certain window size, allowing it to recognize local structures in the input data. The choice of a kernel size is 5 which means the convolutional operation considers five consecutive words at a time, which can be effective in capturing n-gram relationships. iv) MaxPooling Layer: MaxPooling is a down sampling operation that retains the most important information while reducing the spatial dimensions of the input. MaxPooling helps in reducing computational complexity, focusing on the most salient features identified by the Conv1D layer. This down-sampling process aids in maintaining relevant information while discarding less crucial details. A pooling size of 4 is taken which indicates that, for every four consecutive values, the maximum value is retained. v) Bi-Directional GRU Layer: It enables the model to acquire continuous word representations, incorporating contextual information and semantic links. Words with comparable representations in this embedding space are those that share semantic similarities or commonly appear together in the dataset. Better performance on tasks like sentiment analysis and natural language processing is made possible by this approach, which enables the model to record meaningful associations between words. vi) Dropout Layer: For regularization and to prevent overfitting, dropout with a rate of 0.5 is used after each bidirectional GRU layer. Randomly deleting nodes makes it better and decreases the model's reliance on specific connections. vi) Dense Layer: If we are working with binary classification problems the Sigmoid activation function gives us a forecast that comes out like a probability because it squashes the output between 0 and 1. Only one neuron in the output layer is suitable for binary classifier task, in which the goal is to decide to which of two classes a sequence belongs.

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 200)]	0
embedding (Embedding)	(None, 200, 256)	22678272
conv1d (Conv1D)	(None, 196, 64)	81984
max_pooling1d (MaxPooling1D)	(None, 49, 64)	0
bidirectional (Bidirectional)	(None, 49, 128)	49920
dropout (Dropout)	(None, 49, 128)	0
bidirectional_1 (Bidirectional)	(None, 49, 128)	74496
dropout_1 (Dropout)	(None, 49, 128)	0
dense (Dense)	(None, 49, 1)	129

Total params: 22884801 (87.30 MB)
Trainable params: 22884801 (87.30 MB)
Non-trainable params: 0 (0.00 Byte)

Figure 3. Layer-wise Architecture and Parameter Details of the proposed model (EMBED-FLOW)

3.4. Evaluation Metrics

To evaluate the performance of the proposed model, the following common binary classification evaluation metrics has been employed:

- Accuracy: Determines the number of correctly identified instances among the total data set.

Accuracy = Number of correct Predictions / Total Number of Predictions (1)

- Precision: The ratio of actual positive predictions divided by the total number of predictions obtains this value.

Precision = True Positives / True Positives + False Positives (2)

- Recall: The ratio of true positive predictions to the actual positive instances.

Recall = True Positives / True Positives + False Negatives (3)

- F1-Score: Unites precision and recall in one measurement while it balances confusion between false positive and false negative results.

F1-Score = 2×Precision×Recall / Precision + Recall (4)

4. RESULT

4.1 Model Performance

Finally, in this subsection we illustrate the detail of how the proposed model works on the IMBD dataset. This model has introduced some performance parameters (accuracy, precision, recall and F1 score). These metrics facilitate the evaluation of the model's generalization capability on different binary classification tasks based on these metrics.

4.2 Comparative Analysis

To facilitate model comparison, a summary of the performance of pre-existing model and new proposed model is given below in a tabular view.

Table 1. Performance of pre-existing model

Model	Validation Accuracy	Validation Loss	Training Accuracy	Training Loss
GRU	88.19%	47.28%	97.20%	70.83%

EMBED-FLOW architecture displayed superior performance compared to the original GRU model by enhancing generalization together with robustness capabilities, as mentioned in Table 1 and Table 2. The original single-layer GRU model achieved a validation accuracy of 88.19% that EMBED-FLOW improved to 89.34% while providing a relative improvement of 1.15 percentage points. EMBED-FLOW exhibited superior performance for unseen data after its validation loss fell from 47.28% in the baseline GRU to 26.50% in EMBED-FLOW.

Table 2. Performance of new proposed model (EMBED-FLOW)

Validation Accuracy	Validation Loss	Training Accuracy	Training Loss	Precision	Recall	F1-Score
89.34%	26.50%	97.00%	38.41%	88.66%	87.52%	87.20%

Training accuracy performance is equivalent between the GRU model at 97.20% and the EMBED-FLOW model at 97.00% although their training loss outcomes prove dissimilar. Training loss from the conventional GRU reaches 70.83% even though its training data contains 70.83% noise yet EMBED-FLOW demonstrates a substantially lower training loss at 38.41% due to its implementation of dropout and early stopping and its Conv1D layers plus multi-layer bidirectional GRUs learning mechanism. EMBED-FLOW demonstrates superior generalization capabilities because it minimizes the difference between training loss and validation loss. The precision rate of EMBED-FLOW on the IMDB dataset reached 88.66% while its recall reached 87.52% and F1-score achieved 87.20% - metrics which outperformed the unpublished results of the baseline GRU. The proposed model reflects its ability to determine positive reviews correctly together with negative reviews while sustaining a proportionate measurement of false positives against false negatives. EMBED-FLOW improves binary sentiment analysis by providing trustworthy and sophisticated sentiment predictions which prove superior to traditional GRU models during practical applications.

4.3 Graph plot for Training vs Validation Accuracy and Training vs Validation Loss of proposed model

The training development of EMBED-FLOW can be observed through Figure 4 using the data presented in Section 4.3. Initial model development starts with 70% training accuracy that progresses to reach more than 90% accuracy while validation accuracy improves from 68% to 85% between epoch 1 and 5. During epochs 5 to 10 both training accuracy and validation accuracy increase upward with training reaching approximately 96 percent and validation reaching 90 percent before both metrics become comparable. The model shows pattern acquisition through its close measurement relationship between training and validation phases. A visual representation of Training vs. Validation Loss on the graph measures the effectiveness of balanced algorithm learning. From start to epoch 10 both training loss falls from 0.45 to 0.37 while validation loss emerges faster from 0.44 to 0.27 until epoch 10 finishes. The performance of EMBED-FLOW increases across epochs because its validation loss drops more quickly than its training loss. The proposed system achieves quick convergence along with a stable balance between bias and variance factors resulting in reliable performance for unknown data sets through its performance analysis charts.

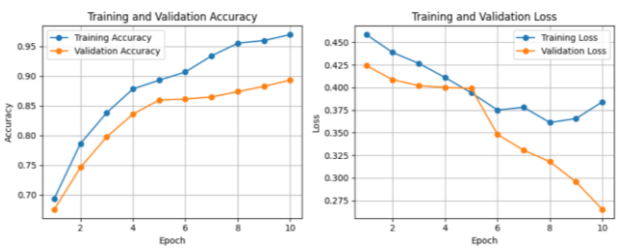

Figure 4. Graph Plot for Training vs Validation Accuracy and Training vs Validation Loss of the proposed Model (EMBED-FLOW)

5. DISCUSSION AND RESULT ANALYSIS

5.1 Model Complexity

By using many layers to process information both forward and backward, the multi-layer bidirectional design improves the model's capacity to identify complex patterns and connections in text data. This method performs better despite its increased complexity, proving its effectiveness in sentiment analysis tasks by successfully collecting both local and global aspects.

5.2 Overfitting and Generalization.

As for overfitting risk and the amount of generalization of model to new, unseen data. An investigation is made of the effects that methods as simple as dropout or K-fold cross validation can have on the model in terms of resilience to a set of datasets, as modeling resilience. We can see that the Validation Accuracy of the proposed model improved by a little for which K-fold cross validation is applied. By embedding one or more layers of Dropout, K-fold cross-validation, Learning Rates or Early Stopping, they have a very significant effect on the Validation accuracy.

5.3 Analysis of Proposed model

After applying some modifications in pre-existing model, the EMBED-FLOW model, as shown in Figure 2 is obtained with higher accuracy and robustness over pre-existing model. Important Changes are listed below: -

- **Conv1D and MaxPooling1D Layers:** The addition of Conv1D and MaxPooling1D layers introduces convolutional operations to capture local patterns and reduce spatial dimensions. The capacity of the model to identify subtle patterns in the input may be enhanced by this combination, which may be useful for extracting pertinent characteristics from sequential data.
- **Learning Rate Scheduling and Early Stopping**: During training, it also adjusts the learning rate schedule as well as the early stopping mechanism through the model EMBED-FLOW. Learning rate scheduling helps optimize the training process, and early stopping prevents overfitting, contributing to more stable and efficient model training.
- **Comprehensive Dropout:** The model EMBED-FLOW includes dropout not only in the bidirectional GRU layers but also in the overall architecture. Comprehensive dropout usage helps mitigate overfitting at different levels of the model, promoting better generalization to unseen data.

The EMBED-FLOW model, which falls under the category of non-pre-trained models and under specific criteria that are discussed above, currently, it is the State of the Art of Sentiment Analysis on the IMDB dataset. A few well-known pre-trained models include XLNet, Word Embeddings (Word2Vec, GloVe, and FastText) and GPT (Generative Pre-trained Transformer).

6. CONCLUSION

This paper proposes a new and robust approach for sentiment analysis based on the developed EMBED-FLOW model, which combines multi-layer bidirectional GRU model with Conv1D, MaxPooling1D, embedding, and enhanced regularization as shown in Figure 2. Using both deep sequential model and efficient optimisation techniques, the model shows better result on the IMDB dataset, a validation accuracy of 89.34%, surpassing non sequential GRU based networks. The strength of the model is its ability to integrate convolutional operations for obtaining localized features and bidirectional recurrent structures to be able to grasp sentence context. In addition, the use of dropout layers, early stopping, k-fold cross-validation, and adaptive learning rate also boosts the model to learn a good generalization over the model to the unseen data by reducing the overfitting of the model. Furthermore, the experimental results also demonstrate that EMBED-FLOW holds a good balance between precision, recall and F1-score, which proves the validation of EMBED-FLOW in binary classification tasks. Both the analysis of training dynamics show that the model converges fast, has small loss gaps and its learning is stable, thus suitable for real-world sentiment analysis scenarios.

7. FUTURE PROSPECTS

7.1 Dataset Bias

The widespread use of the IMDB dataset in sentiment analysis benchmarking has resulted in substantial problems because its built-in structural biases negatively distort the data distribution. The fundamental properties of model performance alongside general practice application change because of present bias that exists within the data. The unbalanced distribution of negative versus positive reviews exists throughout all subgroups that contain release years and genres and different categories of reviewers. Despite having equal positive-negative review proportions at the macro level (25,000 per group), model training could be affected by possible regional imbalances and associated distribution issues in the IMDB dataset. Homogeneous linguistic patterns in review data establish themselves as the main cause of this present research matter. Regular linguistic and cultural patterns used by IMDB reviewers differ from typical social media and e-commerce expressions as well as multilingual textual patterns. Model interpretation of sentiment expressions becomes difficult because IMDB datasets contain limited language styles that mainly include casual and

sarcastic or domain-specific expressions. Multiple evaluation databases need evaluation because they present different linguistic patterns together with cultural elements and structural features. Data augmentation systems need to unite with bias detection technologies to achieve fair analysis results while fulfilling generalized performance standards.

7.2 Hyperparameter Tuning

For the current experiment researchers used set hyperparameters from standard configurations following preliminary trials. The experiment failed to apply a thorough hyperparameter evaluation through either grid search or random selection of parameter value combinations. Within deep learning circles scientists strongly agree that hyperparameters control model outcomes so poor parameter choices restrict the model from reaching its maximum accuracy potential while lowering generality and slowing convergence rates. All crucial parameters like learning rate together with batch size and number of epochs and dropout rate and number of GRU units and the dimensions of Conv1D kernel and MaxPooling1D affect the model's capability to identify patterns present in the dataset. Higher dropout settings enhance generalization, but they might additionally produce underfitting behaviour while inappropriate learning rate values may result in slow convergence together with excessive optimization minimization. Model performance enhancement becomes attainable by using organized hyperparameter optimization techniques which merge grid search and random search with Bayesian optimization along with automated tools Optuna and Keras Tuner. The model performance could potentially be enhanced with adaptive learning rate methods which include ReduceLROnPlateau and AdamW with weight decay. Future researchers should construct effective hyperparameter optimization systems which will let them fully examine their model space to achieve best possible performance results. The addition of this step would enhance both accuracy levels and expose broader configurations capable of achieving consistent success across various datasets while performing different tasks.

7.3 Advanced models can be used to achieve better accuracy

The proposed EMBED-FLOW model shows competitive performance for sentiment classification but the performance can be enhanced by using advanced deep learning models especially pre-trained language models. BERT (Bidirectional Encoder Representations from Transformers) along with XLNet and RoBERTa and other large-scale Transformer-based models have transformed the field of NLP during recent years. The training process of these models occurs through large language databases while they master complex language relationships and can clarify ambiguous sentences together with maintaining syntactic linguistic precision. Standard RNN-based architecture lacks these capabilities. The Transformer-based models establish word relationships through self-attention computation which ignores word sequence positioning for determining their relative value producing advanced dynamic representation structures. Model improvement heavily relies on Word embedding technology which includes FastText, Word2Vec and GloVe because these systems deliver semantic vectors for word representation. Deep models with pre-trained word embeddings are able to achieve even faster training convergence and little better convergence rates, also better generalization at low resource and domain specific though. Using pre-trained models for a particular task such as IMDB and fine-tuning these models gives top-level sentiment analysis results. For the subsequent research, the next step would be applying EMBED- FLOW via transfer learning and developing a mixed model architecture which combines GRU bases with attentions or switch to transformer encoders to improve the performance.

7.4 Limitations

Although the EMBED-FLOW model shows impressive results on the IMDB dataset, this study accepts main constrains which could impact upon the generalizability and robustness of the suggested method. Even though the dataset is very well-known and balanced on a high level, it mainly contains long-form, formally written English reviews and may not appropriately generalize the linguistic variance as well as diversity experienced by other domains. The absence of analysis on other datasets, for example, Amazon product reviews, Yelp reviews, Twitter sentiment data, or Netflix user comments, weaken the switches to the model in generalisation. The difference among these datasets is very large in terms of language patterns, sentence lengths and vocabulary expressions of sentiment, and domain specialized vocabulary. An example of Twitter might be that posts contain informal language, emojis, slang, and abbreviation, which a model must appropriately manage nonstandard grammar and noisy input. On the other hand, Amazon reviews might involve hanging terminology along with product-specific material. Therefore, the model’s effectiveness may decrease when applied to real-world scenarios of short text, domain transfer, or multilingual. One more limitation is the domain adaptation and language transfer. The model is trained only on English language data and its multilingual sentiment classification performance is unexplored so far. Furthermore, it empowers no equation for dealing with code-mixed content, which is basic in the social networking sites and global datasets.

To deal with these challenges, the upcoming work will focus on testing the EMBEDFLOW model on a number of benchmark datasets across domains, including cross-lingual training, testing and on the application combination of adversarial training, adaptation of domains, and real-time inference. Such demonstration can build a qualitative insight of robustness, flexibility, and operationalize model.

References:

AlSurayyi, W. I., Alghamdi, N. S., & Abraham, A. (2019). Deep learning with word embedding modeling for a sentiment analysis of online reviews. *International Journal of Computer Information Systems and Industrial Management Applications*, *11*, 15-15. Retrieved from https://www.ijcisim.org/abstract.php?article=15

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*(1), 5. https://doi.org/10.1186/s13677-022-00386-3

Bjorck, N., Gomes, C. P., Selman, B., & Weinberger, K. Q. (2018). Understanding batch normalization. Advances in neural information processing systems, 31. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2018/file/36072923bfc3cf47745d704feb489480-Paper.pdf

Bodapati, S., Bandarupally, H., Shaw, R. N., & Ghosh, A. (2021). Comparison and analysis of RNN-LSTMs and CNNs for social reviews classification. Advances in Applications of Data-Driven Computing, (pp. 49–59). Springer. https://doi.org/10.1007/978-981-33-6922-3_5.

Cheng, G., Peddinti, V., Povey, D., Manohar, V., Khudanpur, S., & Yan, Y. (2017). An Exploration of Dropout with LSTMs. In Interspeech (pp. 1586-1590). https://doi.org/10.21437/Interspeech.2017-129

Cheng, Y., Yao, L., Xiang, G., Zhang, G., Tang, T., & Zhong, L. (2020). Text sentiment orientation analysis based on multi-channel CNN and bidirectional GRU with attention mechanism. *IEEE Access*, *8*, 134964-134975. https://doi.org/10.1109/ACCESS.2020.3005823

Farzad, A., Mashayekhi, H., & Hassanpour, H. (2019). A comparative performance analysis of different activation functions in LSTM networks for classification. *Neural Computing and Applications*, *31*, 2507-2521. https://doi.org/10.1007/s00521-017-3210-6

Haque, M. R., Lima, S. A., & Mishu, S. Z. (2019). Performance analysis of different neural networks for sentiment analysis on IMDb movie reviews. *3rd International conference on electrical, computer & telecommunication engineering (ICECTE)* (pp. 161-164). IEEE. https://doi.org/10.1109/ICECTE48615.2019.9303573

Kolasani, S. V., & Assaf, R. (2020). Predicting stock movement using sentiment analysis of Twitter feed with neural networks. *Journal of Data Analysis and Information Processing*, *8*(4), 309-319. https://doi.org/10.4236/jdaip.2020.84017

Moss, H. B., Leslie, D. S., & Rayson, P. (2018). Using JK fold cross validation to reduce variance when tuning NLP models. arXiv preprint arXiv:1806.07139. Retrieved from https://arxiv.org/abs/1806.07139

Nguyen, H., Veluchamy, A., Diop, M., & Iqbal, R. (2018). Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, *1*(4), 7. Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss4/7

Onalaja, S., Romero, E., & Yun, B. (2021). Aspect-based sentiment analysis of movie reviews. *SMU Data Science Review*, 5(3), Article 10. Retrieved from https://scholar.smu.edu/datasciencereview/vol5/iss3/10

Sarika, P. K. (2020). Comparing LSTM and GRU for Multiclass Sentiment Analysis of Movie Reviews. DiVA Portal. Retrieved from https://www.diva-portal.org/smash/get/diva2:1454870/FULLTEXT02.pdf

Sen, M., & Chaturvedi, K. (2023). Sentiment Analysis of IMDB Movies Dataset Using Deep Learning Based GRU Model. *International Conference on Sustainable Communication Networks and Application (ICSCNA)* (pp. 176-181). IEEE. https://doi.org/10.1109/ICSCNA58489.2023.10370081

Sluga, M. (2023). Art Film Writing in American Modernist Periodicals, 1910s–1930s. *The Journal of Modern Periodical Studies*, *14*(2), 159-184. https://doi.org/10.5325/jmodeperistud.14.2.0159

Tao, Z., & Wu, Z. (2024). Sentiment Analysis of Product Reviews Based on Bi-LSTM and Max Pooling. *Journal of Intelligent & Fuzzy Systems.* https://doi.org/10.3233/JIFS-223456

Tripathi, S., Mehrotra, R., Bansal, V., & Upadhyay, S. (2020, September). Analyzing sentiment using IMDb dataset. In 2020 *12th International Conference on Computational Intelligence and Communication Networks (CICN)* (pp. 30-33). IEEE. https://doi.org/10.1109/CICN49253.2020.9242524

Yao, Y., Angelov, I., Rasmus-Vorrath, J., Lee, M., & Engels, D. W. (2018). Yelp's Review Filtering Algorithm. *SMU Data Science Review*, *1*(3), Article 3. Retrieved from https://scholar.smu.edu/datasciencereview/vol1/iss3/3

Yenter, A., & Verma, A. (2017, October). Deep CNN-LSTM with combined kernels from multiple branches for IMDb review sentiment analysis. In 2017 *IEEE 8th annual ubiquitous computing, electronics and mobile communication conference (UEMCON)* (pp. 540-546). IEEE. https://doi.org/10.1109/UEMCON.2017.8249049

Zyout, M. A., Shatnawi, R., & Najadat, H. (2023). Malware classification approaches utilizing binary and text encoding of permissions. *International Journal of Information Security*, *22*(6), 1687-1712. https://doi.org/10.1007/s10207-023-00670-5

Sahitya Satya
Amity University Madhya Pradesh,
Gwalior,
India
sahityasaty16@gmail.com
ORCID 0009-0004-0090-9983

Ashok Kumar Shrivastava
Amity University Madhya Pradesh,
Gwalior,
India
akshrivastava1@gwa.amity.edu
ORCID 0000-0002-7245-9916

Deepak Motwani
Amity University Madhya Pradesh,
Gwalior,
India
dmotwani@gwa.amity.edu
ORCID 0000-0002-0217-7155

Babita Tiwari
Manipal University Jaipur,
Jaipur, Rajasthan
India
babita.tiwari@jaipur.manipal.edu
ORCID 0009-0007-2839-5349