Sentiment Analysis of Movie Reviews :

A Comparative Study of Deep Learning Architectures and Embedding Strategies

Table of Content

1. Introduction 2
2. Model Comparison and Best Configuration 2
   2.1 Evaluation 2
   2.2 Best Embedding Configuration and Model Architecture 4

Table of Figure

Figure 1 Evaluation metrics 2
Figure 2 ROC curves of all models 3
Figure 3 Architecture Details: CNN + GloVe 4

1. Introduction

With the rise of online platforms and social media, individuals frequently share their thoughts and opinions in the form of reviews, comments, and posts. This results in vast volumes of unstructured text data, which businesses and analysts can leverage to understand public sentiment and inform decision-making. **Sentiment analysis** plays a crucial role in this process, using natural language processing (NLP) and machine learning techniques to automatically determine whether expressed opinions are positive or negative.

This project applies Python and deep learning techniques to perform sentiment classification on a dataset of movie reviews. The primary objective is to accurately classify each review as either positive (1) or negative (0). Beyond just building classifiers, the study emphasizes the importance of data preprocessing—cleaning and preparing the raw text—to ensure high model performance.

To assess model effectiveness, two deep learning architectures are explored: **Convolutional Neural Networks (CNN)** and **Bidirectional Long Short-Term Memory (BiLSTM)** networks. Each architecture is evaluated with two different word embedding strategies:

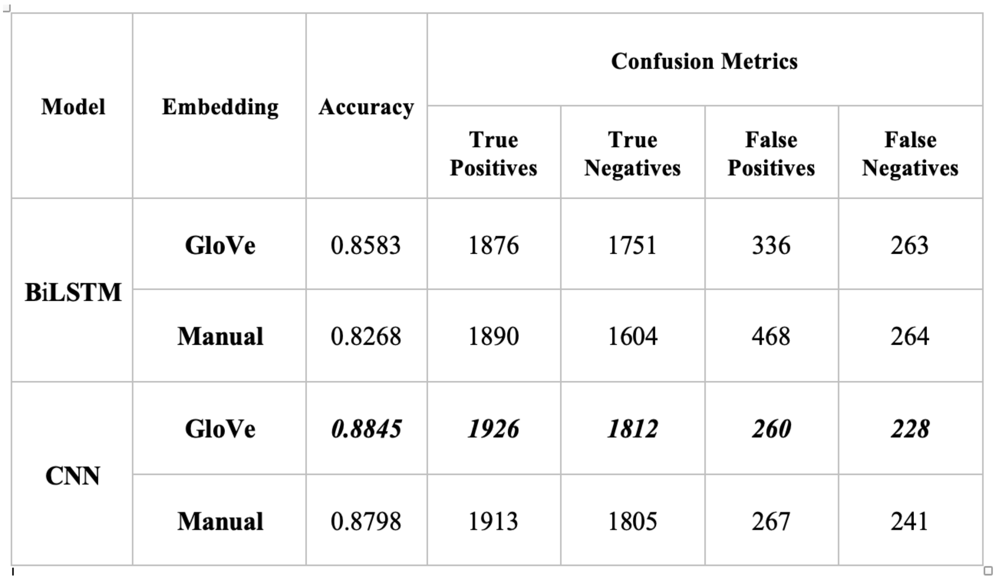
1. **Pre-trained GloVe embeddings**, which capture rich semantic relationships from large corpora, and
2. **Manually trained embeddings**, which are learned directly from the movie review dataset.

This comparison aims to evaluate the impact of both model architecture and embedding strategy on classification accuracy. Although the project was developed on a local environment without large-scale tools, it also reflects on **big data technologies**—such as distributed computing frameworks—that would be suitable for scaling sentiment analysis in real-world applications.

2. Model Comparison and Best Configuration

2.1 Evaluation

Four model configurations were compared based on metrics below (Figure 1).



Model	**Embedding**	**Accuracy**	**Confusion Metrics**			
			True Positives	**True Negatives**	**False Positives**	**False Negatives**
BiLSTM	**GloVe**	0.8583	1876	1751	336	263
	Manual	0.8268	1890	1604	468	264
CNN	**GloVe**	***0.8845***	***1926***	***1812***	***260***	***228***
	Manual	0.8798	1913	1805	267	241

Figure 1 Evaluation metrics

Among all configurations, the CNN model with GloVe embeddings achieved the highest accuracy (88.45%), establishing itself as the top-performing setup for sentiment classification. It also produced the most favourable confusion matrix results, with 1926 true positives and 1812 true negatives, while maintaining low false positives (260) and false negatives (228). The CNN with manual embeddings also performed well, achieving 87.98% accuracy, though it fell slightly short of the GloVe-enhanced counterpart.

In contrast, BiLSTM models showed overall weaker performance, especially when paired with manually initialized embeddings. The LSTM-GloVe model reached an accuracy of 85.83%, while the BiLSTM-manual variant recorded the lowest accuracy (82.68%). Its confusion matrix reflected this decline, with a higher count of false positives (468) and fewer true negatives (1604)—indicating weaker classification reliability.

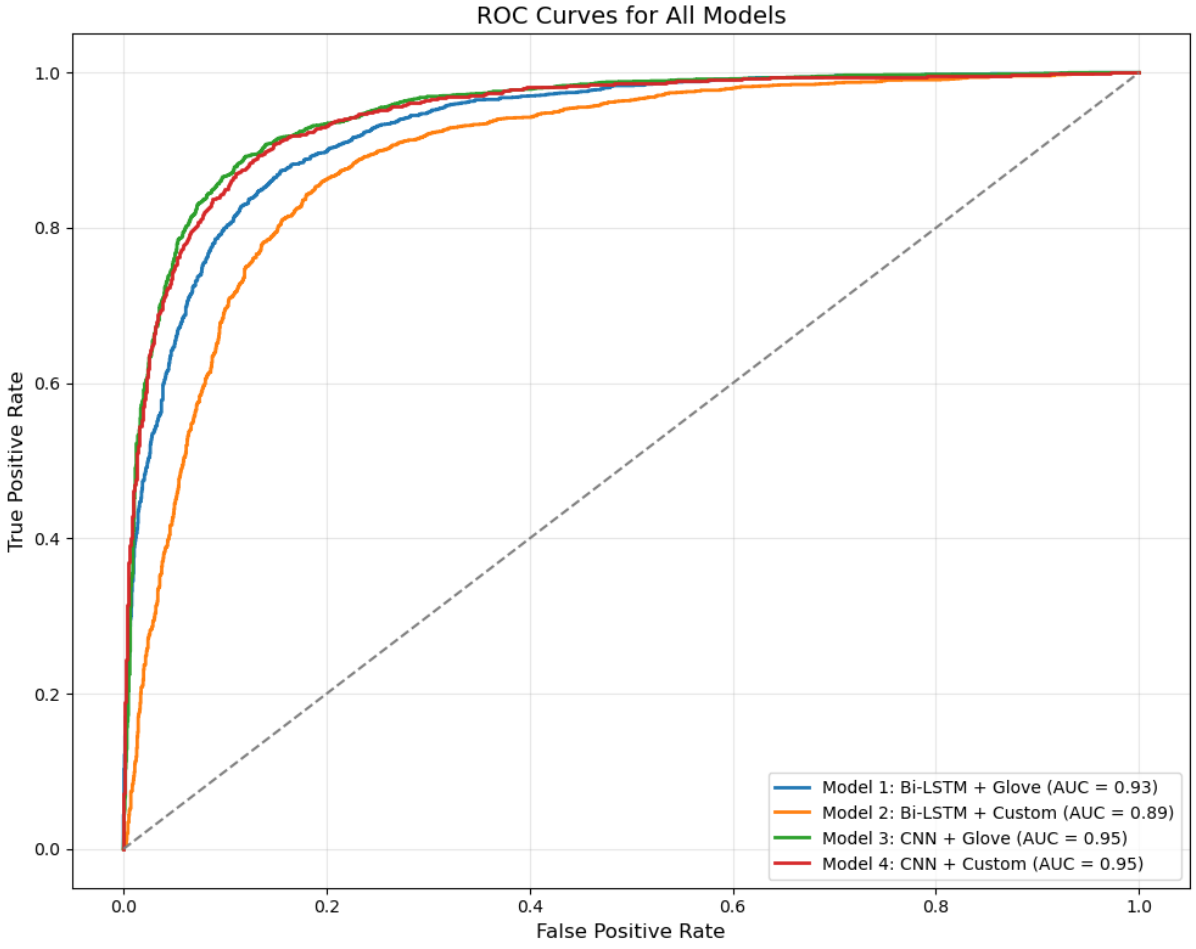
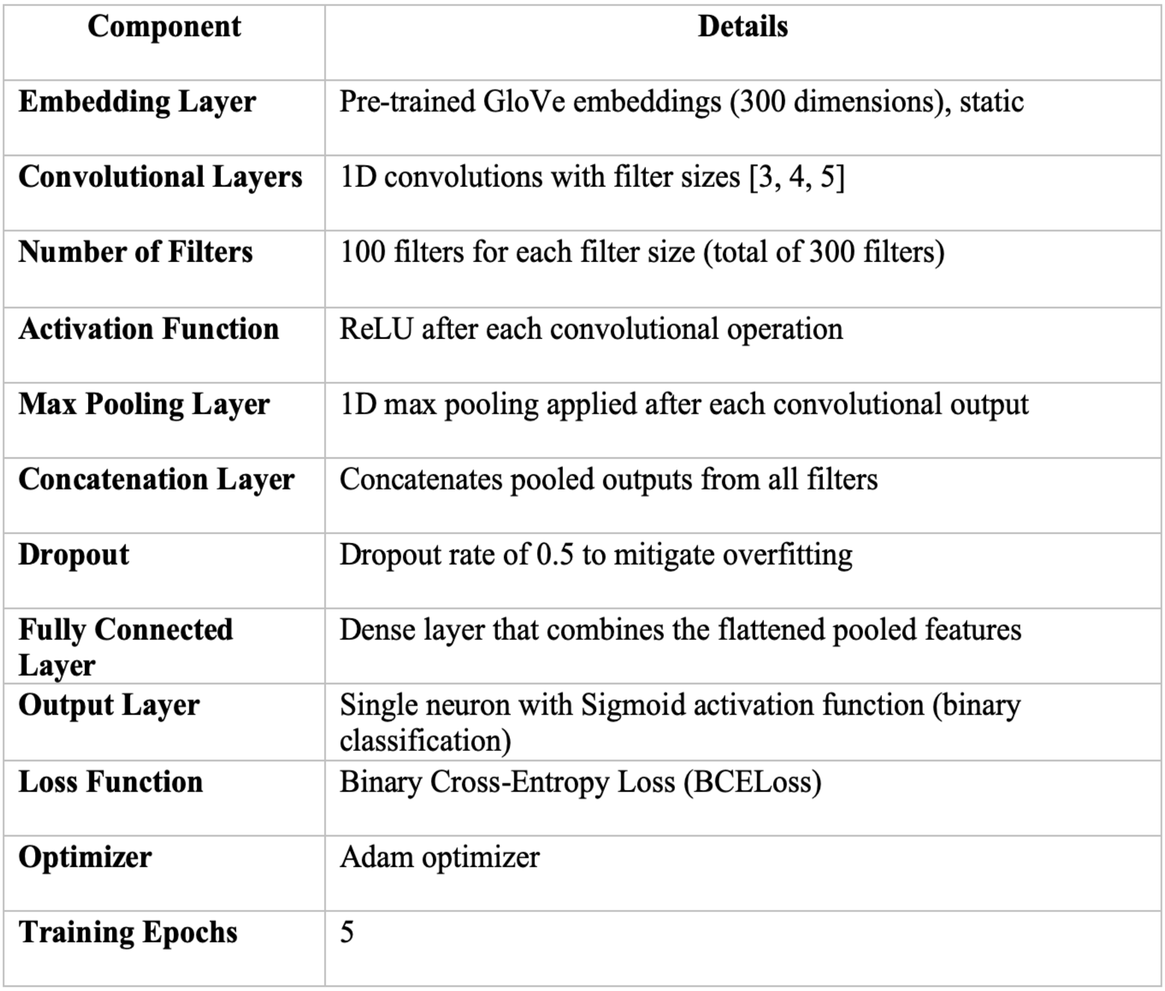


Figure 2 ROC curves of all models

The ROC curves further illustrated these differences in model capability (Figure *2*). The CNN-GloVe configuration achieved the largest area under the curve, with an AUROC of 0.95, confirming its strong discriminatory power. The BiLSTM-GloVe followed with an AUROC of 0.93, while both manual embedding variants (CNN and LSTM) exhibited reduced AUROC scores, with BiLSTM-manual scoring the lowest at 0.89.

These results collectively highlight that model architecture and embedding type are both critical to performance. While CNN models consistently outperformed BiLSTM models, the use of pre-trained GloVe embeddings significantly enhanced accuracy and discrimination in both architectures. The contrast with manual embeddings, particularly in the BiLSTM models, underscores the importance of rich semantic representations. Overall, pairing CNNs with GloVe embeddings emerges as the most effective strategy for accurate and reliable sentiment prediction.

2.2 Best Embedding Configuration and Model Architecture



Component	Details
Embedding Layer	Pre-trained GloVe embeddings (300 dimensions), static
Convolutional Layers	1D convolutions with filter sizes [3, 4, 5]
Number of Filters	100 filters for each filter size (total of 300 filters)
Activation Function	ReLU after each convolutional operation
Max Pooling Layer	1D max pooling applied after each convolutional output
Concatenation Layer	Concatenates pooled outputs from all filters
Dropout	Dropout rate of 0.5 to mitigate overfitting
Fully Connected Layer	Dense layer that combines the flattened pooled features
Output Layer	Single neuron with Sigmoid activation function (binary classification)
Loss Function	Binary Cross-Entropy Loss (BCELoss)
Optimizer	Adam optimizer
Training Epochs	5

Figure 3 Architecture Details: CNN + GloVe

Based on a comprehensive comparison, the CNN model with pre-trained GloVe embeddings emerges as the most effective configuration for this sentiment classification task.

The CNN combined with 300-dimensional GloVe embeddings outperforms other models due to its well-suited architecture and parameter choices tailored to the dataset and task. The pre-trained GloVe embeddings provide rich, semantically meaningful word representations learned from a large corpus, enabling the model to better capture the diverse vocabulary and expressions found in customer reviews, including context-dependent phrases (Wagh and Sinha Anupa, 2024).

The model's multiple convolutional layers, with filter sizes of 3, 4, and 5, effectively capture important local patterns and phrases of varying lengths—such as tri-grams and five-grams—that frequently occur in sentiment expressions within customer feedback. With 100 filters per size, the model covers a broad range of linguistic features and subtle sentiment cues present in the large, diverse dataset, ensuring robust feature extraction.

Global max-pooling layers highlight the most salient features by down sampling the convolution outputs, helping the model focus on key sentiment indicators while reducing noise from less relevant words typical of natural language reviews. To prevent overfitting, a dropout rate of 0.5 randomly deactivates neurons during training—an essential strategy given the variability in review length, style, and vocabulary. This helps the model avoid memorizing dataset-specific patterns and instead generalize effectively.

The fully connected layer with sigmoid activation consolidates the extracted features into a calibrated probability score for binary classification, aligning perfectly with the binary sentiment labels in the dataset (Vishwakarma, 2024). Finally, the Adam optimizer paired with binary cross-entropy loss ensures efficient and stable training, which is critical for learning from the complex and often imbalanced customer sentiment data.

Together, these design choices tailor the CNN-GloVe model to effectively handle the complexity, diversity, and nuances of customer reviews, resulting in its superior performance.