**Advanced Machine Learning Assignment 4**

**Text Data**

(Arcot Balraj Tanmaiyee [|tarcotba@kent.edu](mailto:|tarcotba@kent.edu) | 811321962)

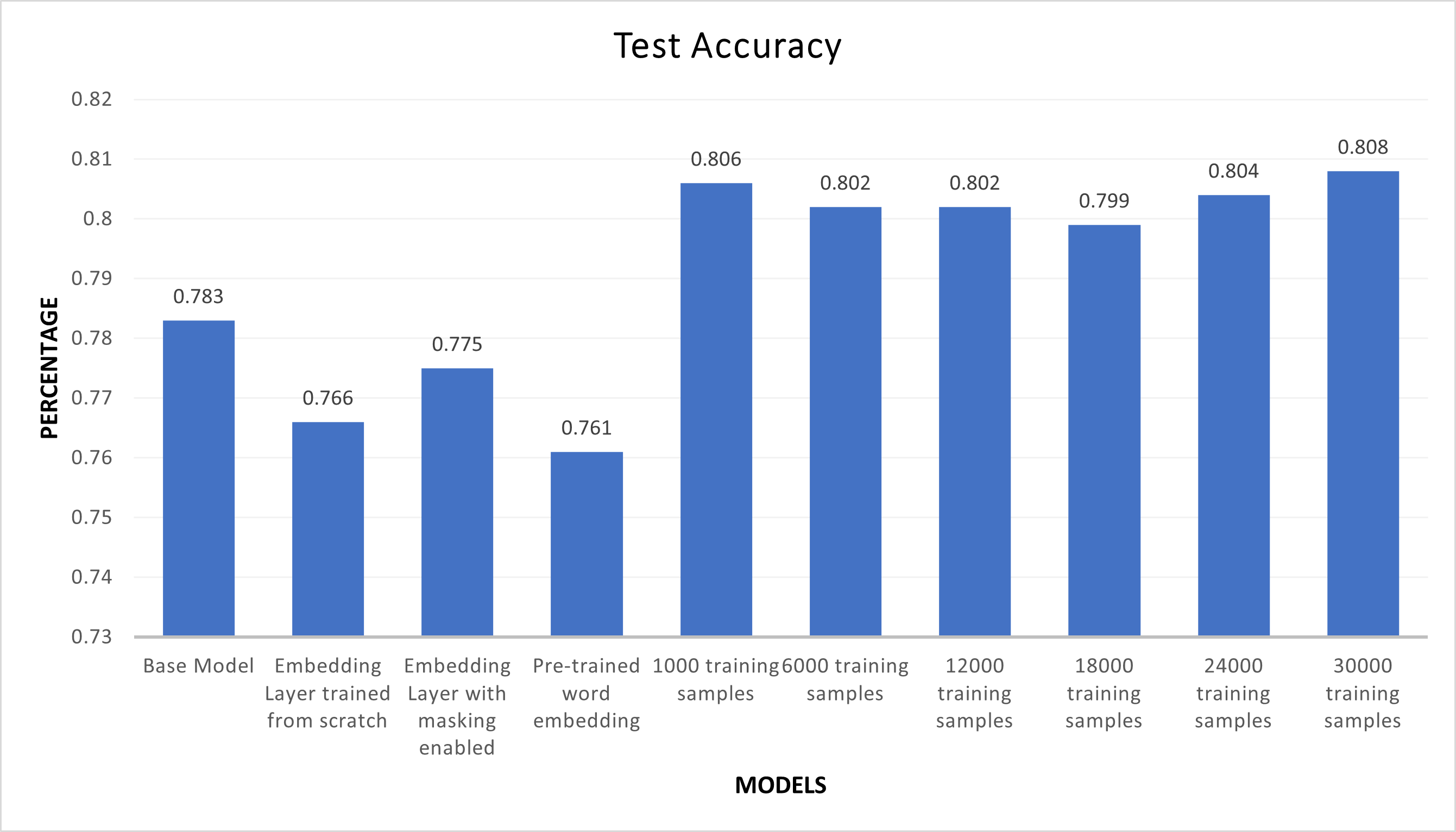
**Overall Summary**

This Assignment examines how different approaches affect a recurrent neural networks (RNN) ability to classify movie reviews according to their emotion. Incorporating embedding layers learned from scratch and using pre-trained word embeddings, training the model on different quantities of data, manipulating datasets to regulate training and validation set sizes, and text preparation to minimize review length and vocabulary size are some of the strategies investigated. The project intends to determine the best practices for creating precise and effective sentiment analysis models by methodically contrasting various approaches using a bidirectional LSTM-based model.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Train Accuracy** | **Validation Accuracy** | **Test Accuracy** |
| Base Model | 0.8232 | 0.7896 | 0.783 |
| Embedding Layer trained from scratch | 0.8238 | 0.7463 | 0.766 |
| Embedding Layer with masking enabled | 0.8593 | 0.7674 | 0.775 |
| Pre-trained word embedding | 0.7140 | 0.7323 | 0.761 |
| 1000 training samples | 0.8667 | 0.7867 | 0.806 |
| 6000 training samples | 0.8651 | 0.7497 | 0.802 |
| 12000 training samples | 0.8686 | 0.7776 | 0.802 |
| 18000 training samples | 0.8676 | 0.7859 | 0.799 |
| 24000 training samples | 0.8605 | 0.7813 | 0.804 |
| 30000 training samples | 0.8691 | 0.7821 | 0.808 |

According to the results, the size of the training data is the most important factor in increasing accuracy for this text categorization task. It may be advantageous to enable masking on the embedding layer when taking the architecture and training configuration into account. Smaller datasets may not benefit as much from pre-trained embeddings. Instead, the model that performed the best (around 80% accuracy) was the one that was trained with more samples. On the training data, the Embedding Layer with masking enabled also produced the best accuracy (around 85%).

Furthermore, it can be seen that the model's accuracy rises with the amount of data used until it reaches roughly 24000. With 30,000 training samples, there was a little decrease in accuracy. The model performed noticeably better as the number of training samples was increased. This implies that a larger training dataset greatly enhances the model's capacity to generalize and produce precise predictions for text categorization tasks. In addition to using the model architecture, training an embedding layer from scratch with additional data yields encouraging outcomes.



**Analysis**

**Ans 1: Comparing and contrasting the results for the best model:**

**Base Model:** Showed a good baseline performance with a reasonable test accuracy of 0.783.

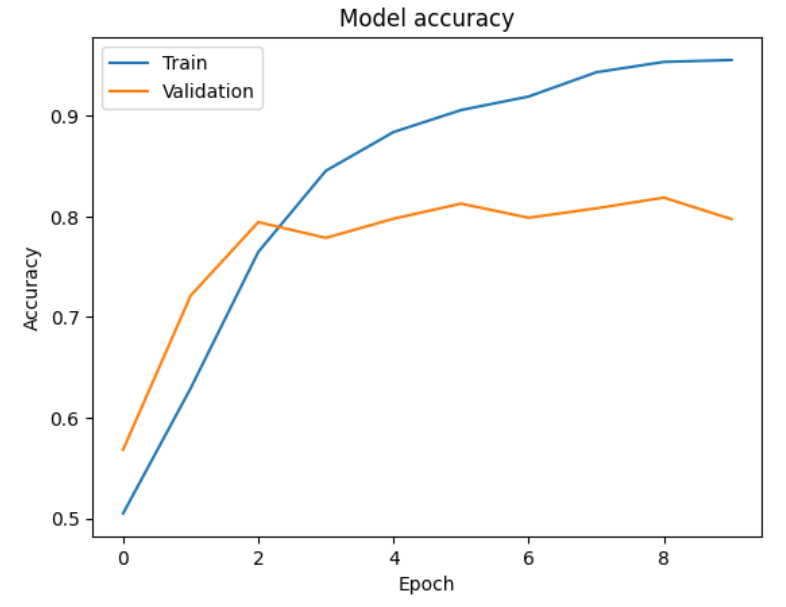
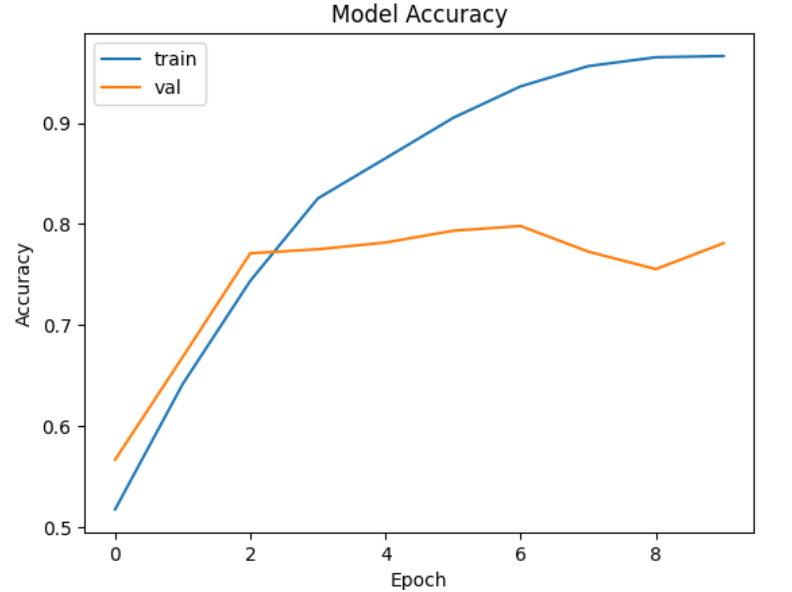
**Embedding Layer:** Was trained from scratch, demonstrated a marginal increase in training accuracy over the base model, but a decline in test and validation accuracy. This suggests that the embedding layer may have overfitted as a result of training on insufficient data.

**Embedding layer with masking:** The maximum training accuracy (0.8593) was attained by this model, indicating efficient use of masked input sequences. Test and validation accuracy, however, continued to lag behind the base model, which would suggest overfitting.

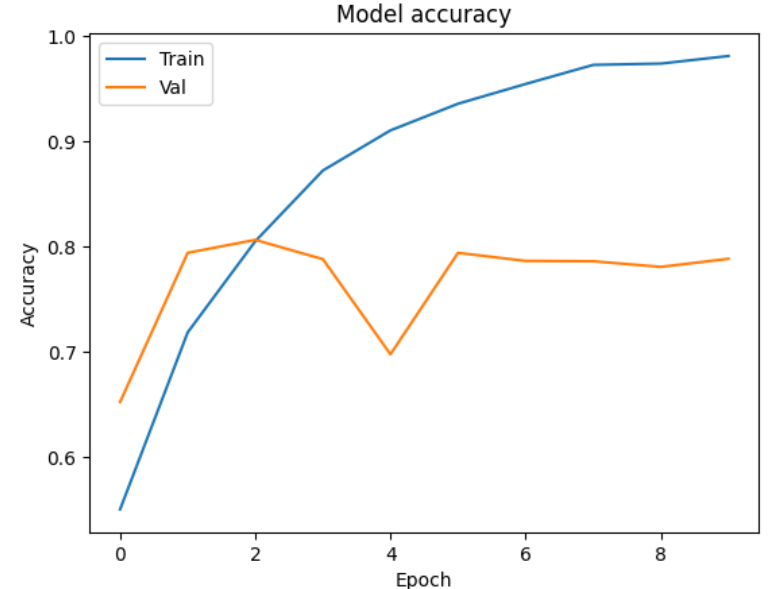
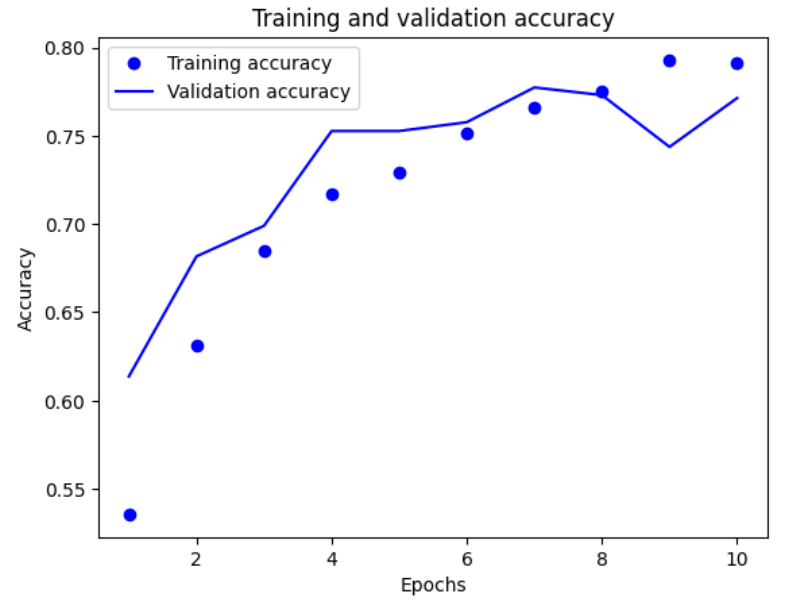
**Pre-trained word embedded (GloVe):** These embeddings may not be the best fit for this particular job or dataset, as the pre-trained word embedding produced the lowest accuracy across all measures.

The Base Model outperforms the other four models in terms of test accuracy. This implies that a more straightforward model with an embedding layer that has already been trained or without one would be a better option for this particular job and dataset. In terms of test accuracy, the Embedding Layer with masking enabled outperforms the other two models, with the exception of the Base Model. It is superior to the other models in this category, despite having a lower validation and test accuracy and a higher training accuracy.

Base Model Embedded Layer Model

Embedded Layer with Masking Pre-Trained Model

**Ans 2: By changing training samples:**

Performance was significantly altered by changing the quantity of the training data. Test accuracy was higher for models trained with more samples than for those trained with fewer. More specifically, test accuracy was 0.806 for 1000 training samples and 0.802 for 6000 training samples. Due to their high training accuracy and low validation and test accuracy, these models also overfit the data. When the data amount is 24000, test accuracy somewhat increases; however, 30000 training samples yielded the greatest results on test data, with an accuracy of 0.808. The pattern indicates that a big amount of training data can increase the model's accuracy. Although there are occasional modest decreases and increases, performance generally improves as the number of training samples grows. With 30,000 samples, we observe the best results.