Summary Report

Assignment 4

Advanced Machine Learning_64061

MSBA

Submitted by: Dileep Kumar Pasala

❖ **Introduction**

This assignment uses bidirectional LSTM models on the IMDB dataset to examine the effects of embedding technique and training sample size on sentiment categorization. The 150-word cutoff, top 10,000-word vocabulary, fixed 10,000-sample validation split, and training subsets of 100–10,000 samples were all met by all models.

❖ **Models Overview**

Models 1–9 were constructed using two embedding types:
- Trainable Embedding (random initialization, learned during training)
- Pretrained GloVe Embedding (100-dimensional, frozen)

❖ **Model Results**

Model	Model Type	Training Sample Size	Train Loss	Train Accuracy	Val Loss	Val Accuracy	Test Loss	Test Accuracy
Model 1	Trainable Embedding	100	0.2388	0.9800	0.8149	0.5702	0.8050	0.5811
Model 2	Pretrained Embedding	100	0.4732	0.7800	0.6946	0.5802	0.6996	0.5794
Model 3	Trainable Embedding	500	0.2306	0.9220	0.7863	0.6336	0.7912	0.6314
Model 4	Pretrained Embedding	500	0.5109	0.7440	0.6772	0.6064	0.6836	0.6014
Model 5	Trainable Embedding	1000	0.1447	0.9500	0.6602	0.7164	0.6280	0.7342
Model 6	Pretrained Embedding	1000	0.4772	0.7710	0.6286	0.6768	0.6224	0.6758
Model 7	Trainable Embedding	5000	0.1401	0.9550	0.6462	0.8060	0.6450	0.8050
Model 8	Pretrained Embedding	5000	0.4206	0.8028	0.8905	0.6088	0.8879	0.6057
Model 9	Trainable Embedding	10000	0.1517	0.9458	0.4984	0.8182	0.5406	0.8057

❖ **Key Findings**
1. For sample size of 100 samples Model 2 (Pretrained Embedding) slightly outperformed Model 1 in validation accuracy: Model 1 Val Acc is 0.5702 where as Model 2 Val Acc is 0.5802. *Pretrained embeddings have a slight advantage with very small datasets.*
2. Trainable Embedding surpass the Pretrained Embeddings when we increase the training sample sizes which we can see it by Comparing paired models: *The Trainable Embedding becomes consistently better starting at 500 samples.*
   - 500 samples: Trainable wins (M3 Val Acc 0.6336 > M4 0.6064)
   - 1,000 samples: Trainable wins (M5 0.7164 > M6 0.6768)
   - 5,000 samples: Trainable strongly wins (M7 0.8060 > M8 0.6088)

❖ **Best Performing Model**

Model 9 (Trainable Embedding, 10,000 samples) achieved the highest Validation Accuracy of 0.8182 and Test Accuracy of 0.8057

❖ **Final Recommendations**
- Use Pretrained Embeddings only when training samples ≤ 100.
- For ≥ 500 samples, Trainable Embeddings outperform consistently.
- For large datasets ($\geq 5,000$ samples), Trainable Embeddings are clearly superior.