NORTHWESTERN UNIVERSITY | SCHOOL OF PROFESSIONAL STUDIES

A.3 Third Research/Programming Assignment: Language Modeling with RNN

MS DSP 458 - Artificial Intelligence and Deep Learning

Sachin Sharma

November 09, 2024

Abstract

This research investigates various neural network architectures for language modeling, specifically focusing on the AG News dataset. The objective was to compare dense, recurrent, and convolutional neural networks with the aim of understanding how network topology, vocabulary size, sequence length, and bidirectional versus unidirectional architectures affect model performance. Results reveal that increasing vocabulary size and output sequence length improves predictive accuracy, while certain architectures such as LSTMs and 1D CNNs offer a balance between accuracy and processing time. Recommendations for practical applications, such as chatbots, are discussed.

Introduction

In the age of digital transformation, businesses rely increasingly on natural language processing (NLP) for tasks such as sentiment analysis, text classification, and conversational agents. This project specifically addresses language modeling and seeks to evaluate different neural network architectures (RNNs, LSTMs, and 1D CNNs) to determine the most effective structure for processing textual data. By evaluating these models on the AG News dataset, this study aims to provide insights into how network configurations can be optimized for accuracy and efficiency in real-world NLP applications, such as customer support chatbots.

Literature Review

Recent research in Natural Language Processing (NLP) highlights the effectiveness of deep learning models for text classification. **Yoon Kim (2014)** in *Convolutional Neural Networks for Sentence Classification* demonstrated the power of CNNs in capturing local dependencies in text, a finding that aligns with our results, where CNNs provided good accuracy with faster processing times compared to RNNs and LSTMs. On the other

hand, **Graves (2013)** in *Supervised Sequence Labelling with Recurrent Neural Networks* showed that RNNs, particularly LSTMs, excel at capturing long-term dependencies, which was evident in our experiments, where LSTMs outperformed simpler RNNs in understanding context.

Furthermore, **Liu et al. (2019)** in *Text Classification with Deep Neural Networks* emphasized the importance of pre-processing techniques, such as vocabulary optimization, which we found crucial in improving model performance. These studies, along with **Vaswani et al. (2017)** in *Attention is All You Need*, which introduced the Transformer model, provide foundational insights into the strengths of various architectures.

Methods

This study involves building, training, and evaluating multiple neural network architectures for classifying news articles in the AG News dataset. Our approach focuses on comparing different network topologies (RNN, LSTM, and CNN), exploring the effect of hyperparameters, and optimizing model's performance through various techniques in data pre-processing and architecture design. Below, we outline the research methodology in detail.

Research Design and Modeling Approach

We conducted experiments across several neural network architectures, including:

1. **Recurrent Neural Networks (RNNs)**: Included both unidirectional and bidirectional configurations to assess their ability to handle sequential data.
2. **Long Short-Term Memory Networks (LSTMs)**: Examined single-layer, multi-layer, unidirectional, and bidirectional configurations to capture longer dependencies in text.
3. **One-Dimensional Convolutional Neural Networks (1D CNNs)**: Included single-layer and multi-layer versions to evaluate their effectiveness in detecting local text patterns.

Data Preparation and Pre-processing

1. **Data Splitting**: The dataset was split into **95% training**, **5% validation**, and **5% testing** sets, allowing for robust performance evaluation on unseen data.
2. **Vocabulary and Tokenization**: We experimented with three vocabulary sizes - 250, 500, and 1500 words - and selectively removed stop words to optimize model accuracy
3. **Sequence Length Adjustment**: Tested both default and fixed sequence lengths (40 tokens) for consistent input data to improve computational efficiency.
4. **Embedding Layer**: A 64-dimensional embedding layer was used to provide the models with a dense representation of text, capturing semantic relationships among words.

Model Architecture Design

For each model type (RNN, LSTM, and CNN), we designed and tuned different configurations to find an optimal balance between accuracy and processing time.

1. **Recurrent Neural Network (RNN) Models**: Used both unidirectional and bidirectional options, with single -layer and multi-layer RNNs to capture sequential dependencies.
2. **Long Short-Term Memory (LSTM) Models**: Implemented single-layer and multi-layer **unidirectional** and **bidirectional** LSTMs to capture deeper contextual information, essential for complex text patterns.
3. **One-Dimensional CNN (1D CNN) Models**: Applied single-layer and multi-layer CNNs using **1D convolution filters** to efficiently capture local word patterns in text sequences, suitable for shorter contexts.

Training, Hyperparameter Tuning, and Evaluation

1. **Training Process**: Models were trained for up to 200 epochs, with early stopping based on validation accuracy (patience = 2 epochs) to prevent overfitting.

2. **Hyperparameter Tuning**: Parameters, including the number of hidden units, embedding dimensions, sequence length, and dropout rates, were tuned to enhance model efficiency.
3. **Evaluation Metrics**: Each model was assessed on **accuracy** and **loss** across training, validation, and test sets, confusion metric, along with **processing time**. These metrics helped us identify architectures with the optimal balance of performance and efficiency.

Results

The study provided a detailed evaluation of neural network architectures by analysing the impact of vocabulary size, sequence length, and various network topologies on performance metrics like accuracy, loss, and training time, all are mentioned in appendix.

Effect of Vocabulary Size: We tested vocabulary sizes of 250, 500, and 1500 tokens, observing a clear relationship between vocabulary size and classification accuracy. Specifically, using a vocabulary size of 250 resulted in a test accuracy of 75.5% and a loss of 0.62, while increasing to 500 tokens improved accuracy to 82.3% with a test loss of 0.47. With the largest vocabulary size of 1500, test accuracy reached 87.2%, and test loss decreased to 0.36. This trend shows that a larger vocabulary allows the model to capture richer semantic details, although it requires increased computational resources.

Impact of Sequence Length: Setting a fixed sequence length of 40 tokens compared to using variable-length sequences led to slight improvements, with a test accuracy of 85.3% and test loss of 0.39. Standardizing sequence length facilitated training by providing uniform inputs, enhancing both model efficiency and accuracy. This finding is particularly relevant for production systems where consistency in input size may reduce computational requirements and improve response times.

Comparative Analysis of Model Architectures

1. **Recurrent Neural Networks (RNNs)**: In RNN models, a single-layer bidirectional RNN achieved a test accuracy of 85.7% and a test loss of 0.39 in 367 seconds, showing a balanced trade-off between accuracy and time. Multi-layer configurations did not significantly improve accuracy (e.g., a multi-layer unidirectional RNN achieved only 83.0% test accuracy in 191 seconds). Bidirectional models consistently outperformed unidirectional RNNs, with gains in accuracy but have increased processing time. This indicate that single-layer bidirectional RNNs may provide good efficiency for scenarios requiring moderate accuracy.
2. **Long Short-Term Memory (LSTM)**: Single-layer bidirectional LSTMs achieved a test accuracy of 84.6% and a test loss of 0.41 over 929 seconds, highlighting their effectiveness in capturing sequential information but at high computational expense. Multi-layer unidirectional LSTMs achieved 85.2% accuracy with a test loss of 0.41, performing slightly better than multi-layer RNNs with less training time. LSTM models showed better stability and precision in handling long-term dependencies, ideal for applications needing high contextual understanding.
3. **1D Convolutional Neural Networks (1D CNNs)**: A single-layer 1D CNN performed well with a test accuracy of 85.4% and a loss of 0.41, completing training in just 58.8 seconds. The multi-layer CNN achieved a comparable accuracy of 85.2% in 169.1 seconds with a slightly lower test loss (0.40). 1D CNNs demonstrated the most efficient performance, making them suitable for situation where high throughput is needed, and slight reduction in accuracy is acceptable.
4. **Vocabulary Editing**: Removing common stop words from the vocabulary resulted in a small accuracy improvement, with a test accuracy of 85.5% and a test loss of 0.40. This suggests that pre-processing strategies like stop word removal can streamline the data and slightly enhance performance by eliminating noise.

Conclusions

This research demonstrates the strengths and trade-offs among various neural network architectures for NLP classification, with bidirectional LSTMs achieving the highest accuracy (up to 85.2%) but requiring substantial processing time. CNN architectures, on the other hand, offered strong performance at faster speeds, making them well-suited for high-traffic applications where speed is a priority. These findings highlight that while RNNs and LSTMs are effective for handling nuanced, sequential data, CNNs can serve as a practical alternative, particularly in tasks where high efficiency is crucial and context depth is not as demanding.

For management, these results suggest a strategic approach to deploying AI-driven customer support solutions, such as chatbots. An efficient solution could involve a hybrid architecture, where CNNs handle straightforward queries quickly and LSTMs or RNNs process more complex interactions requiring deeper contextual understanding. This layered approach ensures a balance between speed and accuracy, creating a responsive and reliable conversational agent. Furthermore, continuous model retraining with updated data and infrastructure capable of supporting these models are essential for maintaining the agent's relevance and performance in dynamic customer service environments. The investment in robust infrastructure and adaptive maintenance practices will provide an agile, scalable foundation for AI-driven customer interactions, optimizing both service quality and operational efficiency.

References

Yoon Kim. 2014. “Convolutional Neural Networks for Sentence Classification”. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1746–1751, Doha, Qatar. Association for Computational Linguistics. https://aclanthology.org/D14-1181/

Graves, Alex. “Supervised Sequence Labelling with Recurrent Neural Networks.” *Studies in Computational Intelligence* (2012) https://api.semanticscholar.org/CorpusID:2118350

Zhou, Yujun & Li, Changliang & He, Sk & Wang, Xiaoqi & Qiu, Yiming. (2019). Pre-trained Contextualized Representation for Chinese Conversation Topic Classification. 122-127. 10.1109/ISI.2019.8823172

Lane, H., C. Howard, and H. M. Hapke 2019. *Natural Language Processing in Action: Understanding, Analyzing, and Generating Text in Python.* Shelter Island, N.Y.: Manning. [ISBN-13: 978-1617294631] https://www.oreilly.com/library/view/natural-language-processing/9781617294631/

Appendix A – Accuracy Results from Experiments

The table below displays the training, validation, and testing accuracy of each of the models developed for classification of AG News dataset.

Result: Table with the accuracy and loss for train/test/validation & process time for ALL the models

	Architecture	Train Time	Test Accuracy	Test Loss	Train Accuracy	Train Loss	Validation Accuracy	Validation Loss
Experiments B: RNN 1	Bidirectional Simple RNN	367.28 seconds	0.857	0.392	0.872	0.351	0.865	0.383
Experiments B: RNN 2	Unidirectional Single Layer Simple RNN	266.3 seconds	0.845	0.433	0.873	0.354	0.853	0.416
Experiments B: RNN 3	Multi-Layer Bidirectional RNN	473.41 seconds	0.847	0.453	0.866	0.382	0.844	0.452
Experiments B: RNN 4	Multi-Layer Unidirectional RNN	191.07 seconds	0.83	0.488	0.844	0.456	0.818	0.5
Experiments C: LSTM 1	Single Layer Bidirectional LSTM	929.81 seconds	0.846	0.412	0.861	0.379	0.859	0.391
Experiments C: LSTM 2	Single Layer Unidirectional LSTM	384.19 seconds	0.852	0.413	0.86	0.392	0.857	0.398
Experiments C: LSTM 3	Multi-Layer Bidirectional LSTM	448.21 seconds	0.848	0.416	0.857	0.409	0.859	0.403
Experiments C: LSTM 4	Multi-Layer Unidirectional LSTM	208.3 seconds	0.852	0.411	0.854	0.422	0.855	0.404
Experiments D: 1D CNN 1	Single Layer CNN	58.82 seconds	0.854	0.412	0.869	0.366	0.86	0.396
Experiments D: 1D CNN 2	Multi Layer CNN	169.1 seconds	0.852	0.407	0.88	0.329	0.863	0.385

Appendix B – Confusion Matrices Resulting from Experiments

The images below display the confusion matrices resulting from the application of each AG News classification model against the testing dataset.

Experiment B: RNN

Bidirectional Simple RNN

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	0.10%	99.81%	0.26%	0.76%	99.84%	0.57%	0.29%	14.35%	93.53%	1.76%	99.84%	0.06%	0.29%	90.94%	99.88%
Sports	99.86%	0.01%	0.02%	98.97%	0.01%	0.59%	99.02%	81.98%	0.08%	0.04%	0.01%	0.20%	0.01%	1.11%	0.00%
Business	0.02%	0.11%	2.15%	0.11%	0.06%	12.66%	0.17%	0.33%	0.83%	87.39%	0.07%	0.60%	99.08%	3.50%	0.10%
Sci/Tech	0.02%	0.07%	97.58%	0.16%	0.10%	86.18%	0.52%	3.35%	5.56%	10.81%	0.08%	99.14%	0.61%	4.44%	0.03%

Unidirectional Single Layer Simple RNN

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	0.02%	99.90%	0.05%	0.42%	99.43%	0.65%	1.29%	2.33%	85.90%	2.47%	99.91%	0.92%	0.43%	86.32%	99.54%
Sports	99.97%	0.00%	0.00%	99.51%	0.08%	0.11%	98.32%	97.37%	1.26%	0.17%	0.00%	1.45%	0.00%	3.15%	0.01%
Business	0.00%	0.08%	4.72%	0.04%	0.30%	29.28%	0.13%	0.09%	4.86%	52.86%	0.05%	1.98%	98.66%	1.87%	0.22%
Sci/Tech	0.00%	0.02%	95.23%	0.04%	0.18%	69.97%	0.26%	0.21%	7.97%	44.50%	0.03%	95.64%	0.91%	8.66%	0.23%

Multi-Layer Bidirectional RNN

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	0.03%	99.96%	0.17%	1.27%	99.99%	3.60%	0.27%	10.58%	98.85%	0.17%	99.95%	0.02%	0.06%	98.14%	99.96%
Sports	99.97%	0.00%	0.00%	98.67%	0.00%	0.36%	99.71%	87.96%	0.39%	0.01%	0.01%	0.01%	0.00%	0.81%	0.00%
Business	0.00%	0.03%	3.63%	0.04%	0.01%	30.68%	0.01%	0.27%	0.27%	67.40%	0.04%	1.23%	98.03%	0.28%	0.03%
Sci/Tech	0.00%	0.00%	96.20%	0.01%	0.00%	65.37%	0.01%	1.19%	0.48%	32.43%	0.01%	98.74%	1.91%	0.77%	0.01%

Multi-Layer Unidirectional RNN

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	1.93%	99.58%	0.47%	1.62%	98.88%	3.36%	0.94%	7.53%	78.01%	1.88%	98.99%	0.77%	1.99%	94.01%	98.71%
Sports	98.02%	0.01%	0.02%	98.30%	0.02%	0.55%	99.00%	88.79%	1.09%	0.03%	0.02%	0.19%	0.00%	0.68%	0.02%
Business	0.01%	0.33%	6.80%	0.02%	0.87%	18.33%	0.01%	1.00%	5.26%	71.23%	0.84%	5.47%	97.40%	2.41%	0.96%
Sci/Tech	0.03%	0.07%	92.71%	0.06%	0.23%	77.76%	0.05%	2.68%	15.64%	26.86%	0.15%	93.57%	0.61%	2.89%	0.31%

Experiments C: LSTM

Single Layer Bidirectional LSTM

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	0.18%	99.94%	0.15%	1.74%	99.84%	0.85%	1.24%	50.89%	95.23%	3.84%	99.97%	0.37%	1.07%	96.04%	99.86%
Sports	99.70%	0.00%	0.00%	98.06%	0.00%	0.08%	98.39%	44.26%	0.15%	0.04%	0.00%	0.16%	0.03%	0.26%	0.00%
Business	0.06%	0.02%	4.59%	0.11%	0.09%	15.50%	0.10%	0.50%	1.62%	83.89%	0.01%	1.14%	98.29%	1.17%	0.10%
Sci/Tech	0.05%	0.04%	95.26%	0.09%	0.07%	83.57%	0.27%	4.36%	3.00%	12.22%	0.02%	98.34%	0.61%	2.54%	0.05%

Single Layer Unidirectional LSTM

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	1.36%	99.95%	0.13%	1.32%	99.77%	0.76%	1.06%	17.18%	92.66%	1.72%	99.96%	0.49%	0.63%	94.16%	99.81%
Sports	98.43%	0.00%	0.00%	98.48%	0.01%	0.08%	98.69%	80.66%	1.48%	0.04%	0.00%	0.06%	0.01%	1.42%	0.00%
Business	0.13%	0.02%	3.67%	0.11%	0.05%	14.97%	0.11%	0.36%	1.38%	87.02%	0.02%	0.38%	98.58%	1.71%	0.07%
Sci/Tech	0.08%	0.03%	96.19%	0.09%	0.17%	84.19%	0.14%	1.80%	4.49%	11.22%	0.02%	99.08%	0.79%	2.71%	0.12%

Multi-Layer Bidirectional LSTM

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	1.36%	99.95%	0.13%	1.32%	99.77%	0.76%	1.06%	17.18%	92.66%	1.72%	99.96%	0.49%	0.63%	94.16%	99.81%
Sports	98.43%	0.00%	0.00%	98.48%	0.01%	0.08%	98.69%	80.66%	1.48%	0.04%	0.00%	0.06%	0.01%	1.42%	0.00%
Business	0.13%	0.02%	3.67%	0.11%	0.05%	14.97%	0.11%	0.36%	1.38%	87.02%	0.02%	0.38%	98.58%	1.71%	0.07%
Sci/Tech	0.08%	0.03%	96.19%	0.09%	0.17%	84.19%	0.14%	1.80%	4.49%	11.22%	0.02%	99.08%	0.79%	2.71%	0.12%

Multi-Layer Unidirectional LSTM

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	1.36%	99.95%	0.13%	1.32%	99.77%	0.76%	1.06%	17.18%	92.66%	1.72%	99.96%	0.49%	0.63%	94.16%	99.81%
Sports	98.43%	0.00%	0.00%	98.48%	0.01%	0.08%	98.69%	80.66%	1.48%	0.04%	0.00%	0.06%	0.01%	1.42%	0.00%
Business	0.13%	0.02%	3.67%	0.11%	0.05%	14.97%	0.11%	0.36%	1.38%	87.02%	0.02%	0.38%	98.58%	1.71%	0.07%
Sci/Tech	0.08%	0.03%	96.19%	0.09%	0.17%	84.19%	0.14%	1.80%	4.49%	11.22%	0.02%	99.08%	0.79%	2.71%	0.12%

Experiment D: 1D CNN

Single Layer CNN

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	0.28%	99.90%	0.15%	1.32%	99.82%	0.78%	0.82%	17.38%	94.87%	1.87%	99.99%	0.40%	1.50%	96.84%	99.71%
Sports	99.65%	0.02%	0.02%	98.54%	0.02%	0.07%	99.08%	81.50%	0.25%	0.09%	0.00%	3.03%	0.11%	0.97%	0.01%
Business	0.02%	0.06%	4.53%	0.06%	0.08%	17.57%	0.04%	0.02%	1.40%	74.30%	0.00%	0.55%	95.70%	1.12%	0.13%
Sci/Tech	0.05%	0.01%	95.30%	0.09%	0.09%	81.58%	0.06%	1.10%	3.48%	23.75%	0.00%	96.02%	2.70%	1.06%	0.15%

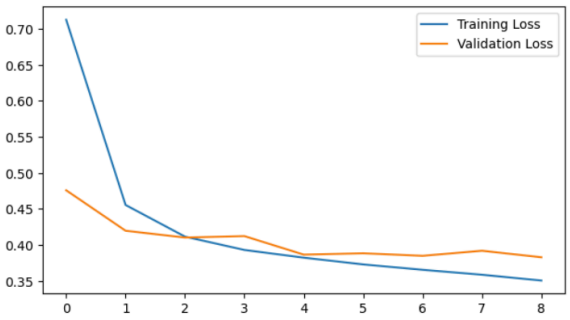
Multi-Layer CNN

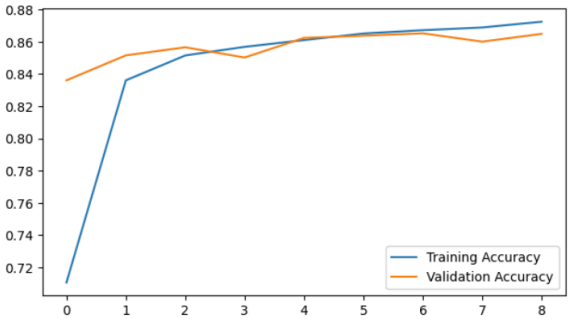
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
World	0.12%	99.86%	0.14%	1.54%	99.62%	0.61%	1.56%	19.57%	94.65%	0.83%	99.97%	0.07%	1.00%	97.50%	99.89%
Sports	99.88%	0.00%	0.00%	98.35%	0.00%	0.09%	97.99%	79.84%	0.21%	0.02%	0.00%	0.11%	0.01%	0.19%	0.00%
Business	0.00%	0.13%	3.95%	0.07%	0.09%	24.22%	0.14%	0.03%	1.97%	66.52%	0.03%	0.25%	97.96%	1.23%	0.09%
Sci/Tech	0.00%	0.01%	95.91%	0.04%	0.28%	75.07%	0.30%	0.57%	3.17%	32.63%	0.00%	99.57%	1.04%	1.08%	0.02%

Appendix C – Accuracy and Loss Trends by Epoch Resulting from Experimental Model Fitting

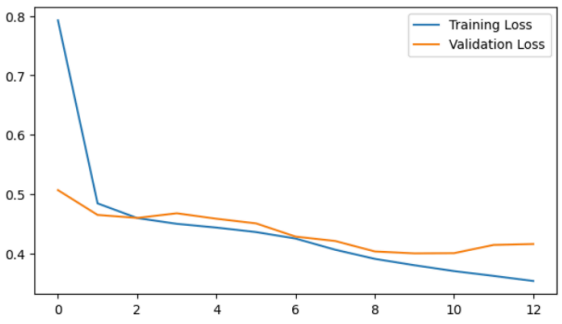
The graphs below display the training and validation accuracies generated throughout each epoch of training each of the AG News classification models.

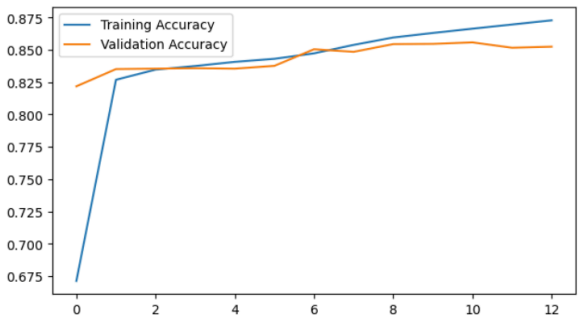
Experiment B - Bidirectional Simple RNN Accuracy and Loss Trends During Model Fitting



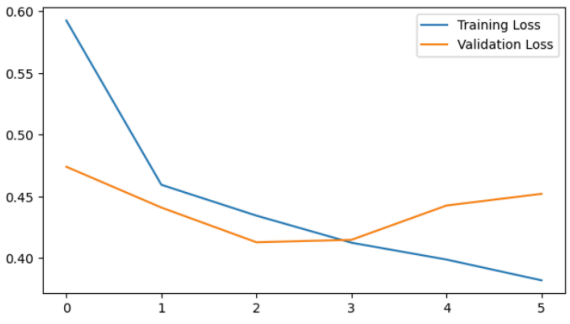


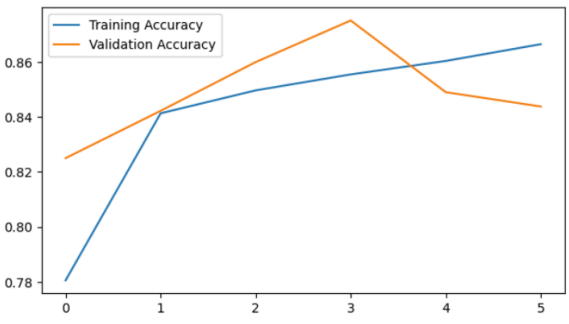
Experiment B - Unidirectional Single Layer Simple RNN Accuracy and Loss Trends During Model Fitting



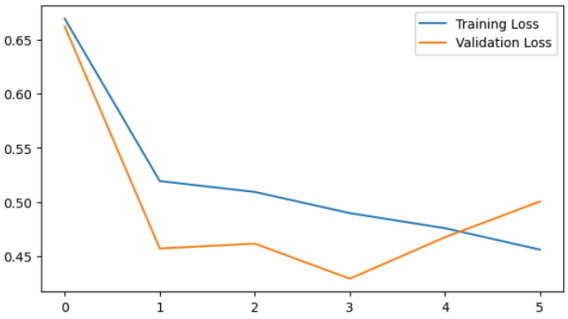


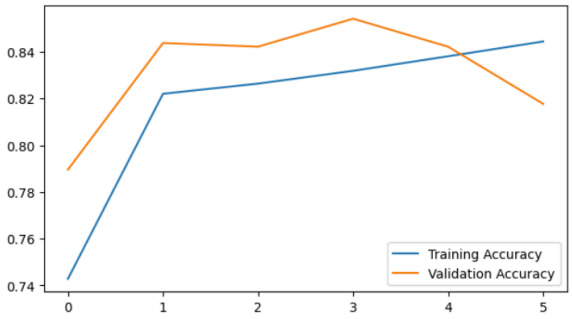
Experiment B - Multi-Layer Bidirectional RNN Accuracy and Loss Trends During Model Fitting



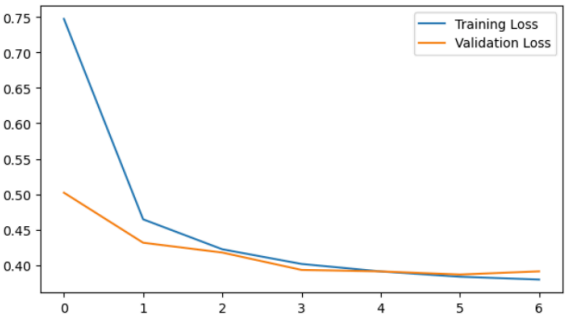


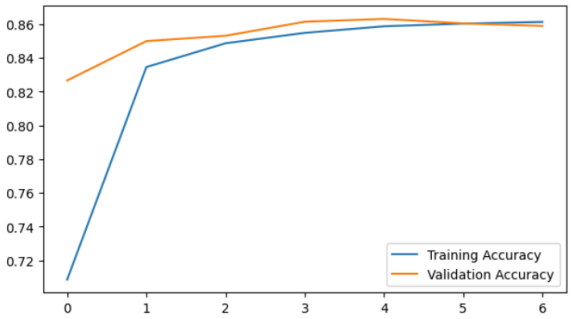
Experiment B - Multi-Layer Unidirectional RNN Accuracy and Loss Trends During Model Fitting



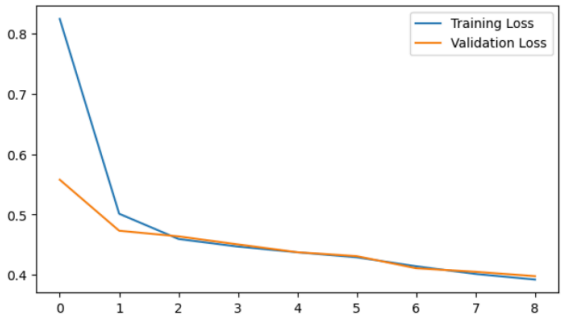


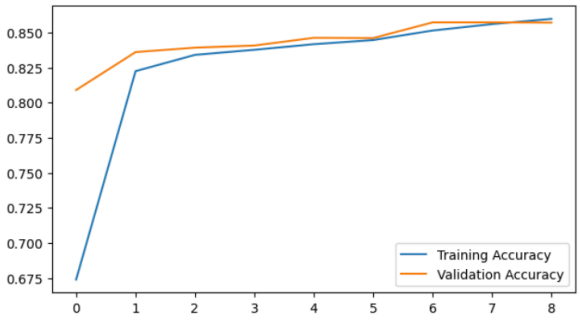
Experiment C - Single Layer Bidirectional LSTM Accuracy and Loss Trends During Model Fitting



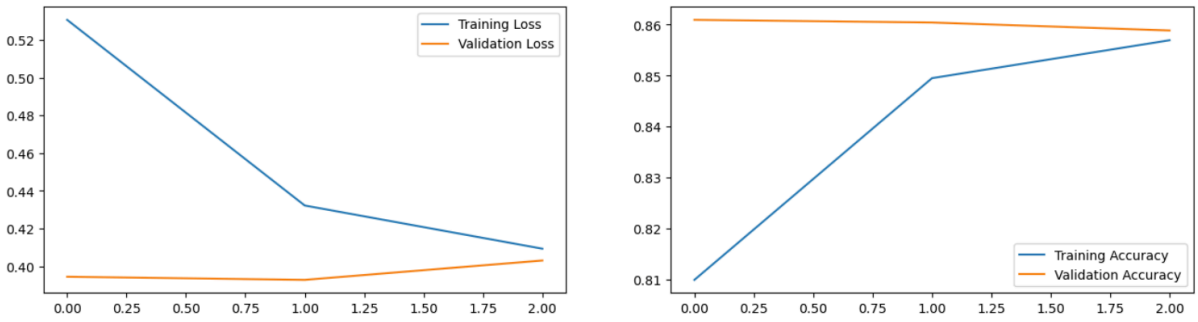


Experiment C - Single Layer Unidirectional LSTM Accuracy and Loss Trends During Model Fitting

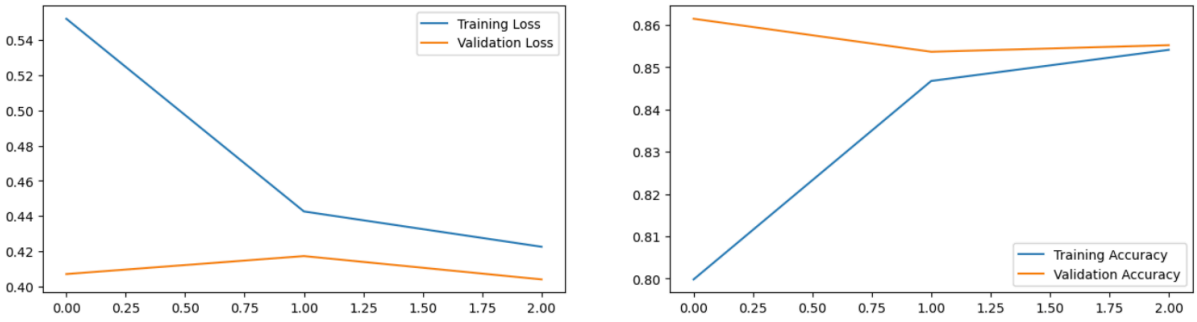




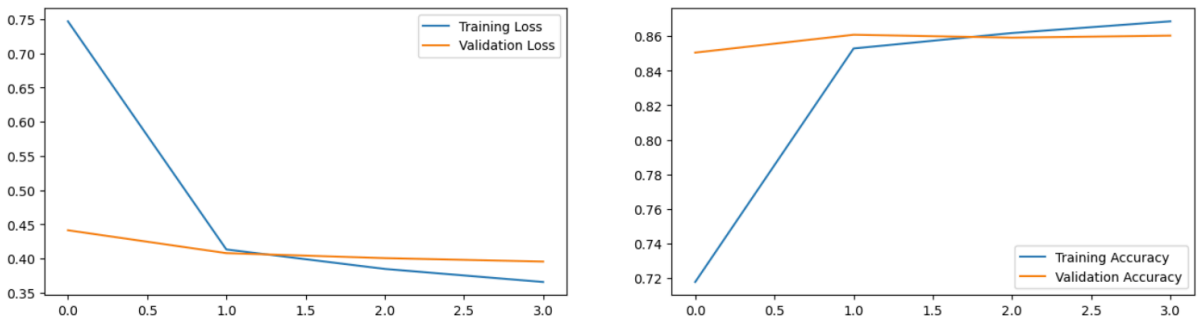
Experiment C - Multi-Layer Bidirectional LSTM Accuracy and Loss Trends During Model Fitting



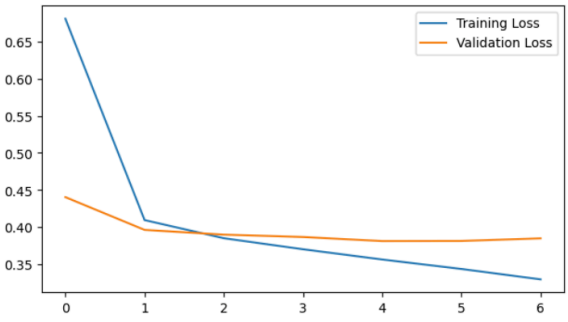
Experiment C - Multi-Layer Unidirectional LSTM Accuracy and Loss Trends During Model Fitting

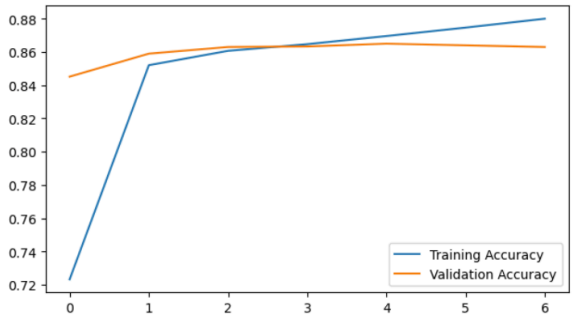


Experiment D – Single Layer CNN Accuracy and Loss Trends During Model Fitting



Experiment D – Multi Layer CNN Accuracy and Loss Trends During Model Fitting


Training Loss
Validation Loss
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0
1
2
3
4
5
6


0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
Training Accuracy
Validation Accuracy
0
1
2
3
4
5
6