**Project Report**

**Overview**

The goal of this project was to classify text data from the AG News dataset into four categories which are World, Sports, Business, and Science/Technology, using three different neural network models: a Feedforward Neural Network (FFNN), a Recurrent Neural Network (RNN) and a Bidirectional Long Short-Term Memory (BiLSTM) network. Each model was trained using three optimizers: SGD, Adam and Adagrad. The project involved text preprocessing, model development, training, evaluation and visualization of results.

**Dataset Definition**

The AG News dataset was loaded and analyzed, consisting of 120,000 training samples and 7,600 testing samples. Each sample was categorized into one of the four news categories. The dataset was preprocessed and cleaned for model training. The textual data included both article titles and descriptions, providing comprehensive information for classification.

Dataset Highlights :

* Training Samples: 120,000
* Test Samples: 7,600
* Classes: World, Sports, Business, Science/Technology
* Source: AG's corpus of news articles
* Language: English

**Data Preprocessing**

Before feeding the data into the neural networks, extensive preprocessing was conducted. This involved:

* Lowercasing all text.
* Removing URLs, punctuation, extra whitespaces, and stopwords using both Spacy and NLTK.
* Lemmatization of words.

The cleaned data was then tokenized and converted into numerical sequences using TensorFlow’s Tokenizer. The sequences were padded to ensure uniform input size for the models.

**Visualizations**

Several visualizations were performed to understand the dataset distribution and the effect of preprocessing:

* Word Clouds: Displayed the most frequent words in both the raw and cleaned datasets.
* Category Distribution: Visualized the number of samples per category using a bar chart.
* Text Length Distribution: Plotted the distribution of text lengths in the cleaned dataset.

These visualizations provided insights into the data structure and the impact of cleaning operations.

**Model Implementation**

Three neural network architectures were implemented and trained:

1. Feedforward Neural Network (FFNN)
2. Recurrent Neural Network (RNN)
3. Bidirectional Long Short-Term Memory (BiLSTM)

For each model, three optimizers were used: SGD, Adam, and Adagrad. The models were trained for 10 epochs using a batch size of 64, and early stopping was applied to avoid overfitting.

**Model Details**:

* FFNN: A simple architecture with embedding, flattening, and dense layers.
* RNN: Included an embedding layer followed by a SimpleRNN layer for capturing sequential information.
* BiLSTM: Used bidirectional LSTM units to handle long-range dependencies in text data.

**Results Comparison**

After training, the models were evaluated on the test dataset. The following results were obtained:

**FFNN**:

* SGD Accuracy: 86.10%
* Adam Accuracy: 85.45%
* Adagrad Accuracy: 85.75%

**RNN**:

* SGD Accuracy: 26.07%
* Adam Accuracy: 25.31%
* Adagrad Accuracy: 25.31%

**BiLSTM**:

* SGD Accuracy: 36.99%
* Adam Accuracy: 87.48%
* Adagrad Accuracy: 87.58%

Insights:

* The BiLSTM model with Adagrad provided the highest accuracy (87.58%) on the test dataset.
* The RNN model performed poorly across all optimizers due to the simple architecture and inability to capture long-term dependencies in the text.
* FFNN with SGD had comparable performance to more complex models showing the effectiveness of simpler architectures when combined with the right optimizer.

**Conclusion**

The project successfully demonstrated the implementation of multiple neural network architectures and optimizers for text classification. The BiLSTM model with the Adagrad optimizer emerged as the best-performing combination. The project also highlighted the importance of preprocessing, model selection, and optimizer choice in achieving high classification accuracy.