**Introduction to Text and Sequencing**

**What is Text and Sequencing?**

In our project, "text and sequencing" refers to the process of handling and analyzing text data in a way that respects the order or sequence in which words appear. This is crucial because the meaning of a piece of text often depends not just on the words used, but also on how they are arranged.

**Importance in Sentiment Analysis:**

For example, in sentiment analysis—the task of determining whether the sentiment behind a text is positive or negative—the sequence of words can change the overall sentiment. Phrases like "not good" versus "good" clearly have opposite meanings, despite both containing the word "good."

**Our Approach:**

In this project, we use advanced machine learning models that can understand and process these sequences. This involves:

* Preparing the Text: Converting raw text into a format that our machine learning models can work with, such as turning sentences into lists of numbers where each number represents a different word.
* Using Models That Recognize Patterns: Employing models like LSTM (Long Short-Term Memory), which is designed to remember the order of words in a sentence, helping it capture the intended meaning more accurately.

**Summary and Application of Deep Learning Techniques:**

Our project focused on applying deep learning methods to perform sentiment analysis on the IMDB dataset. We systematically employed a range of techniques to optimize model performance:

* Data Preparation: We carefully segmented the dataset into training, validation, and testing groups to support model training and evaluation.
* Text Vectorization: Using TensorFlow's Text Vectorization, we transformed textual data into a numerical format that our neural networks could process, setting parameters to handle vocabulary size and sequence length.
* Embedding: We explored embedding words using both custom and pre-trained GloVe embeddings, which helped the model to capture semantic meanings of words effectively.
* Recurrent Neural Networks: Bidirectional LSTMs were utilized to leverage both past and future context within text sequences, essential for understanding the nuanced language in reviews.
* Model Training and Evaluation: Our models underwent multiple iterations of training with adjustments in architecture and parameters to enhance accuracy and reduce overfitting.

**Key Observations on Improving Model Accuracy:**

* Increased Training Data: One of the most significant observations was the improvement in model accuracy as we increased the training dataset size. Initially, with smaller datasets, the models struggled to generalize well, leading to modest accuracies. However, as we expanded the dataset from 100 samples to 7100, and then to 14100, there was a notable increase in validation and test accuracies. This improvement underscores the importance of large datasets in training deep learning models, as more data provide a better representation of the variability in human language, helping the model learn and generalize more effectively.
* Model Adjustments: Refining the model architecture and tuning parameters like the number of LSTM units, dropout rates, and the use of bidirectional layers also contributed to better performance. These adjustments helped the models to capture more complex patterns in the data without overfitting.

**Final Conclusions:**

* Data Volume and Model Complexity: Our findings highlight that both the quantity of training data and the sophistication of the model architecture are pivotal in enhancing performance. Larger datasets enable the model to train on a broader range of text expressions and sentiments, while advanced model features like bidirectional LSTMs and embeddings can more effectively process and understand this information.
* Continuous Refinement: The iterative process of adjusting model configurations and increasing the dataset proved essential in achieving the best outcomes. This iterative refinement is crucial for adapting the model to the nuances of natural language processing tasks.

**Table of Results:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model Configuration | Training Samples | Validation Accuracy | Test Accuracy | Loss |
| Sequence model with embedding layer | 100 | 51.46% | 51.42% | 0.7372 |
| Sequence model with pre-trained embedded layer (GLoVe) | 100 | 53.64% | 53.69% | 0.6908 |
| LSTM + Increased Training Samples | 7100 | 81.18% | 82.05% | 0.5086 |
| LSTM + Further Increased Training Samples | 14100 | 81.80% | 82.20% | 0.6679 |

Above presented is a refined breakdown of the performance across different model configurations and training sizes.

**Visualization of our results:**

A screenshot of a graph

Description automatically generated

**A screenshot of a graph

Description automatically generated**

**Storyline/Recommendation**:

* Improving Results with More Data: This project shows that using more data and updating our models helps us better understand and predict people's opinions from movie reviews. As we added more reviews to train our models, they got better at guessing the right sentiments.
* Learning by Doing: The improvements we saw remind us that experimenting and making changes as we go can lead to better results. It's important to keep trying new things and refining our methods to get the best performance from our models.
* Next Steps and Real-World Use: Moving forward, we should try using even more data and try out new types of models to see if we can do even better. These models could be very useful for businesses to understand what people think about their products or services and help them make smarter decisions.