Assignment 6

**1. Aim**

To implement a Sentiment Analysis model using Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) networks to classify text data as positive or negative, and evaluate the performance of the model.

**2. Objectives**

* To understand the fundamentals of sentiment analysis and its applications.
* To explore the use of LSTM and GRU networks for text classification tasks.
* To preprocess textual data for sentiment analysis.
* To train an LSTM/GRU model to classify text into positive or negative sentiments.
* To evaluate the model's accuracy and performance using appropriate metrics.

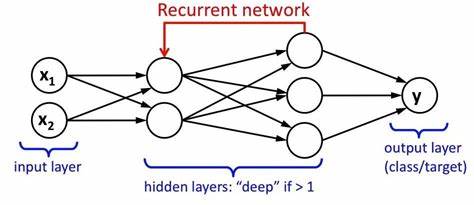
**3. Theory**

**3.1 Sentiment Analysis**

Sentiment analysis is the process of identifying and categorizing opinions expressed in a piece of text, especially in terms of whether the writer’s attitude towards a particular topic is positive, negative, or neutral. It is widely used in applications such as customer feedback, reviews, and social media analysis.

**3.2 Recurrent Neural Networks (RNNs)**

Traditional RNNs are used for processing sequences of data. However, they face challenges when learning long-term dependencies due to the vanishing gradient problem. This led to the development of more advanced models like LSTMs and GRUs, which can capture longer dependencies in the sequence data.



RNN architecture

**3.3 Long Short-Term Memory (LSTM)**

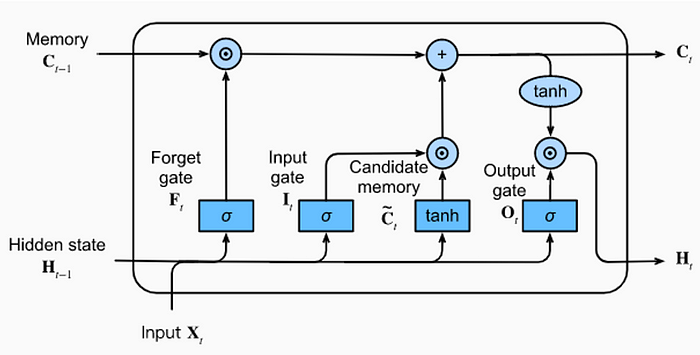
LSTMs are a type of RNN capable of learning long-term dependencies by using a special gating mechanism. An LSTM unit consists of:

* **Forget Gate:** Decides what information should be discarded from the cell state.
* **Input Gate:** Updates the cell state with new information.
* **Output Gate:** Controls what part of the cell state should be output.

LSTMs are effective for sentiment analysis as they can retain information over longer text sequences, improving context understanding.

**3.4 Gated Recurrent Unit (GRU)**

GRU is a simplified version of LSTM that uses fewer gates (reset and update gates), making it computationally more efficient while still retaining the ability to capture dependencies in the sequence data. GRUs are also commonly used for sentiment analysis tasks.



LSTM architecture

**4. Working/Algorithm Used**

**4.1 Data Collection and Preprocessing:**

1. **Data Collection:**
   * Sentiment analysis datasets like IMDb movie reviews, Amazon reviews, or custom datasets are collected for training and testing. These datasets usually contain labeled text data with positive and negative sentiments.
2. **Data Preprocessing:**
   * **Tokenization:** Split the text into individual words or tokens.
   * **Stopword Removal:** Remove common words (e.g., "the", "is") that don’t contribute to sentiment analysis.
   * **Padding:** Pad the sequences to a uniform length for batch processing.
   * **Word Embeddings:** Use pre-trained embeddings such as Word2Vec, GloVe, or train embeddings from scratch using Keras' Embedding layer to convert text into numeric form.

**4.2 Model Architecture:**

The sentiment analysis model is built using either LSTM or GRU layers. The architecture includes embedding layers for text representation, followed by LSTM/GRU layers for capturing the sequence patterns, and fully connected layers for classification.

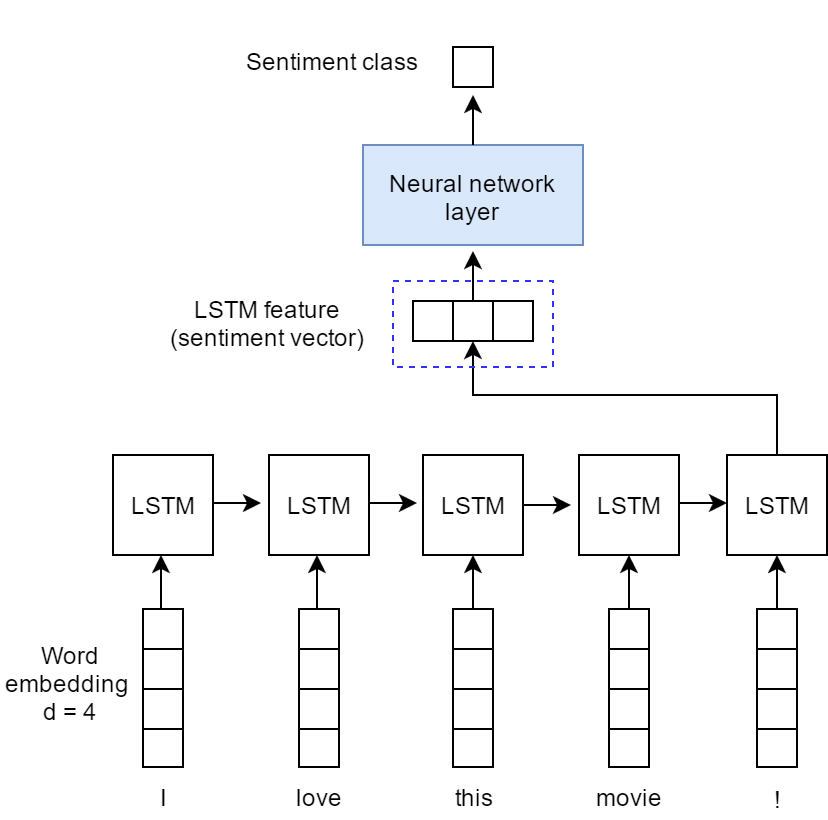
1. **Embedding Layer:**
   * The embedding layer converts the input text into dense vectors of fixed size, where words with similar meanings have similar vector representations.
2. **LSTM/GRU Layers:**
   * These layers capture sequential dependencies between words in a text. For sentiment analysis, they help capture context, such as the polarity of the sentence.
3. **Fully Connected Layers:**
   * After the LSTM/GRU layers, a Dense layer with a sigmoid activation function is used to classify the sentiment as either positive or negative.

**4.3 Model Compilation and Training:**

1. **Loss Function:**
   * Use binary cross-entropy loss for binary classification tasks (positive/negative sentiment).
2. **Optimizer:**
   * Adam optimizer is typically used for optimizing the model during training.
3. **Training:**
   * The model is trained using the training dataset with a validation split to monitor performance during training. The training is done over several epochs with a fixed batch size.

**4.5 Evaluation Metrics:**

* **Accuracy:** The percentage of correct predictions over the total number of predictions.
* **Confusion Matrix:** Used to evaluate the performance of the binary classifier by showing true positives, false positives, true negatives, and false negatives.
* **Precision, Recall, and F1-Score:** These metrics provide more detailed performance evaluation by focusing on different aspects of classification.



**5. Conclusion**

In this project, we successfully implemented a sentiment analysis system using LSTM or GRU networks. The model was trained on a dataset of labeled reviews and was able to classify new text into positive or negative sentiments with high accuracy. The use of LSTM/GRU networks allowed the model to capture long-term dependencies between words in the text, improving the sentiment classification performance. Further improvements could include the use of more complex architectures, such as bidirectional LSTMs or attention mechanisms, and experimenting with larger datasets.