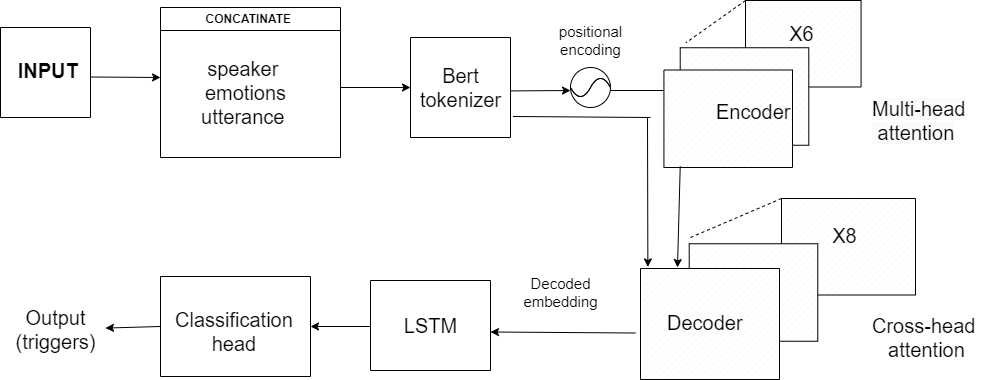
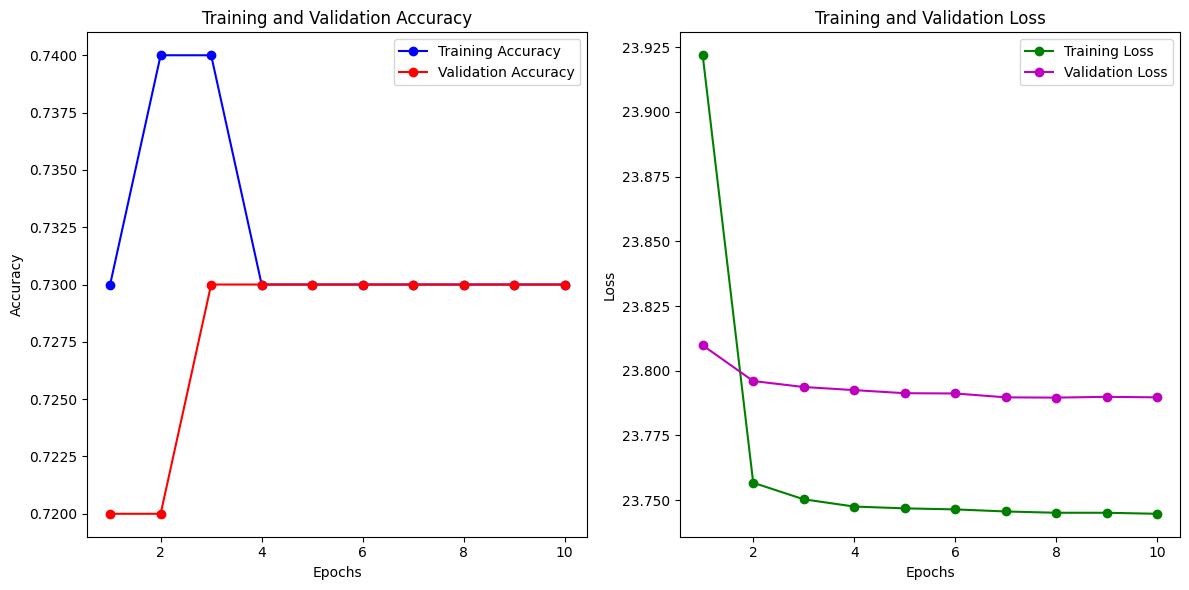
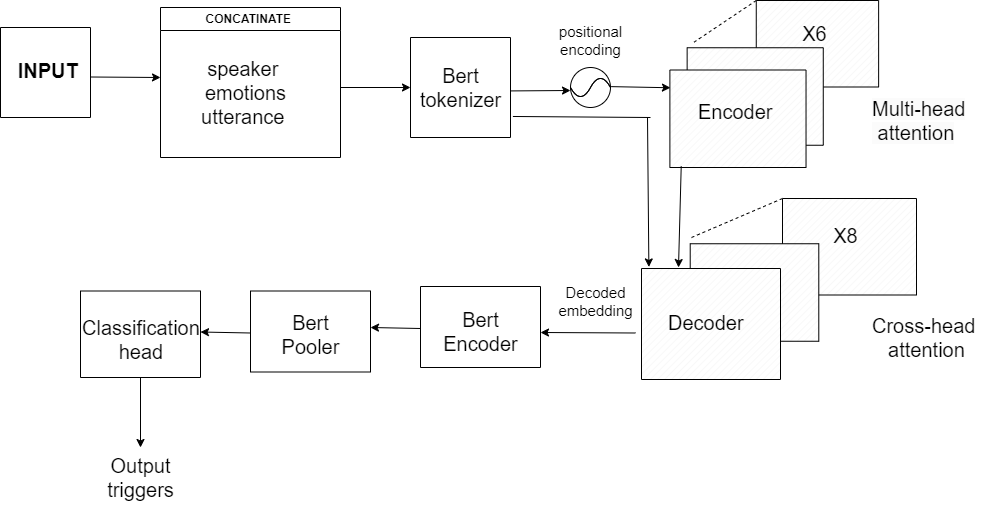
**M3**



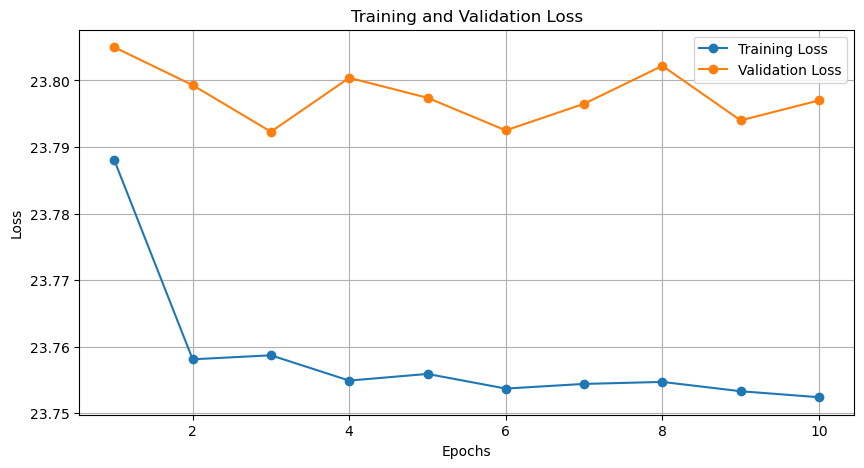
The input, comprising concatenated speaker, emotion, and utterance data, is tokenized using BERT to convert it into numerical values while maintaining positional encoding. These values are then processed by a transformer encoder, enabling the capture of relationships via self-attention, creating a memory unit. The decoder receives the last sentence's embedding from this memory unit, facilitating cross-attention between the last utterance and memory to integrate contextual cues. Additionally, an LSTM is utilized for sequence identification, followed by a classification head for prediction.

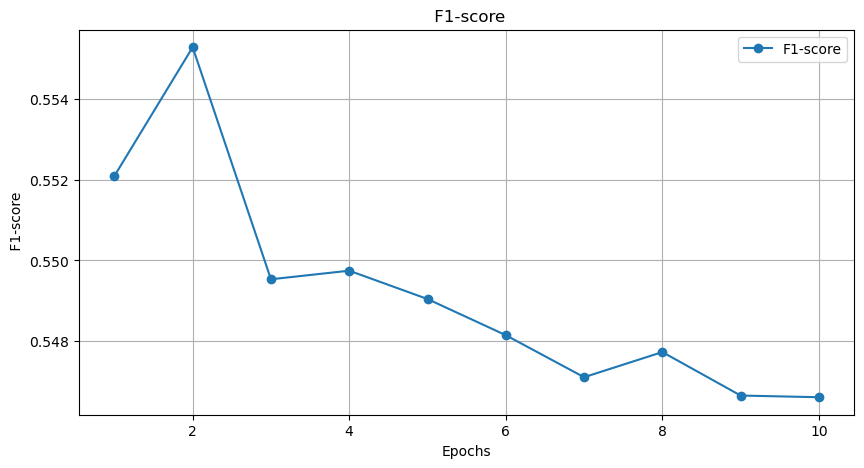


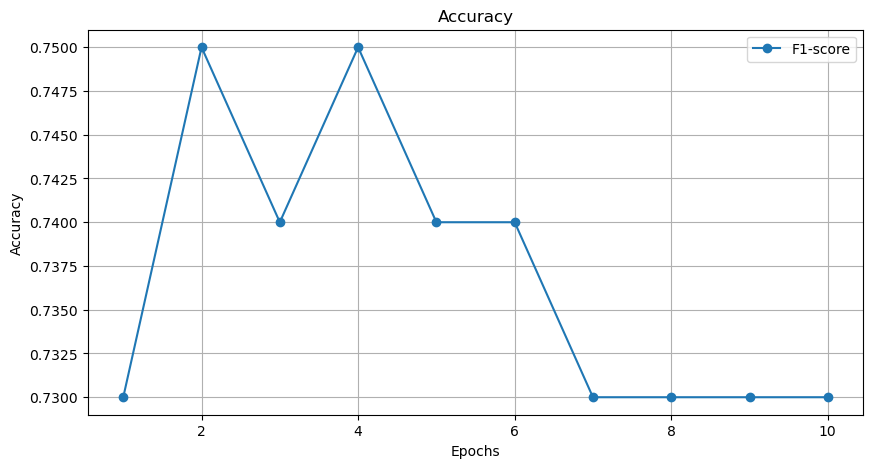
**M4**



The input consists of concatenated speaker, emotion, and utterance information, which undergoes tokenization using the BERT tokenizer to convert it into numerical values while preserving the positional encoding of each word and sentence. These numerical representations are then fed into a transformer encoder to capture relationships through self-attention mechanisms, forming a memory unit. The last sentence's embedding from this memory unit is transferred to the decoder. Here, cross-attention is applied between the last utterance and the memory unit to incorporate contextual information. Finally, leveraging a pre-trained BERT encoder and pooler, along with a classification head, facilitates accurate predictions.







**Comparison:**

Both models are working almost identical because both are transformer based encoder-attension-decoder architecture. Before classification layer, both the model have different implementation. Former one has LSTM which captures temporal features better than the later one which has BERT encoder and pooler.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Loss (train)** | **Loss (val)** | **F1-macro** |
| M3 | 23.7449 | 23.789 | 0.5474 |
| M4 | 23.7524 | 23.797 | 0.5466 |