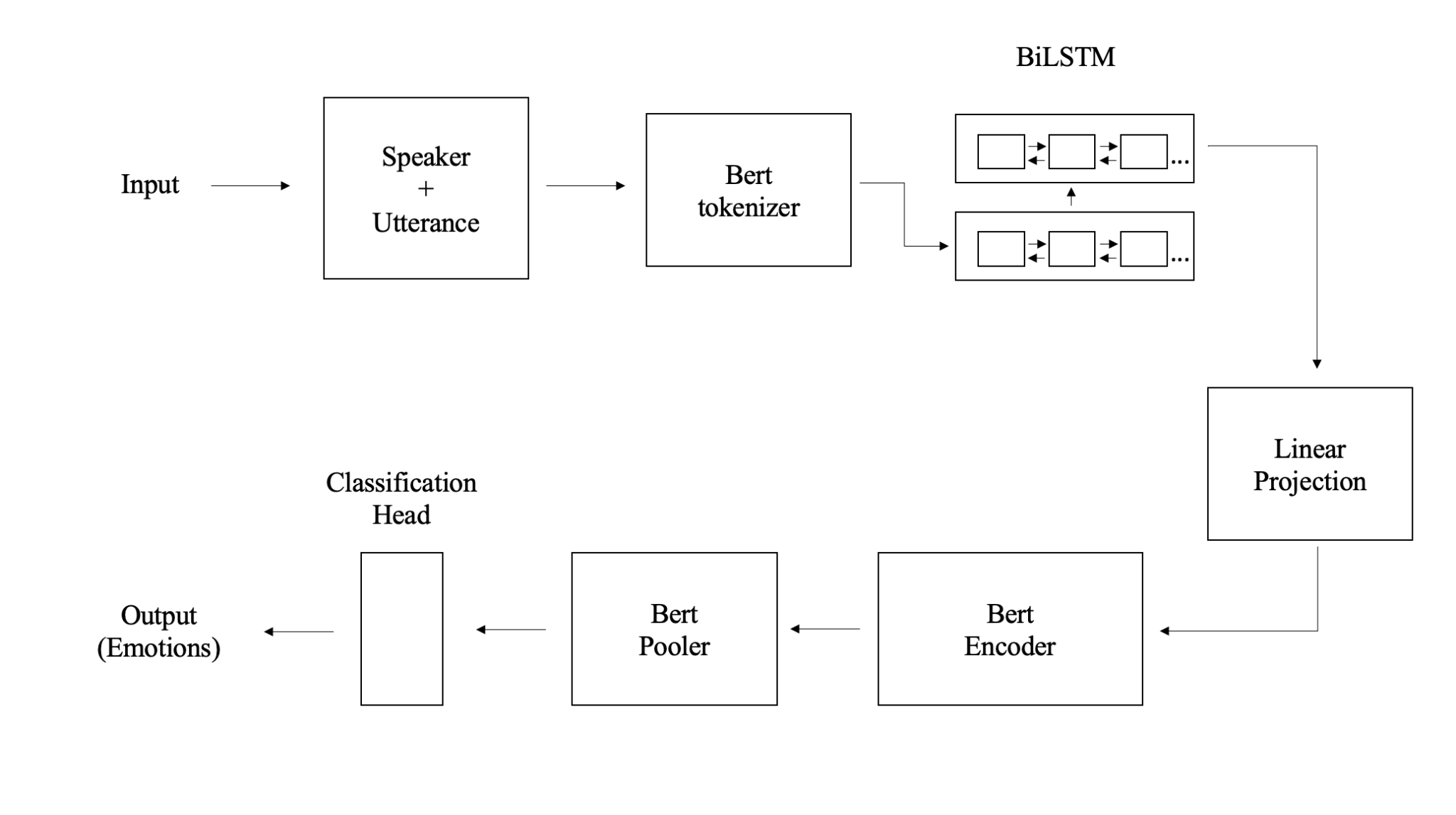
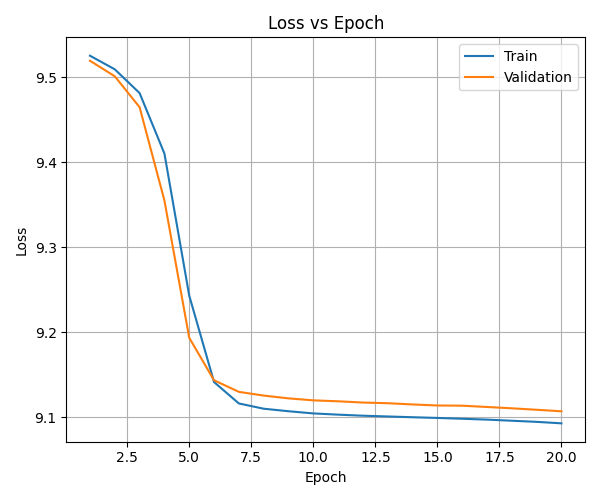
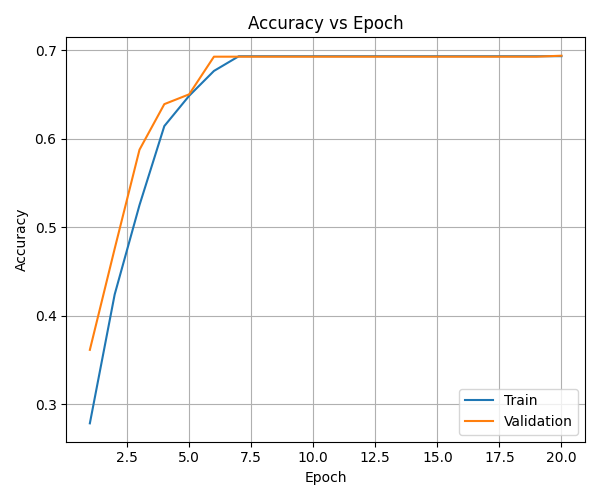
**M1**

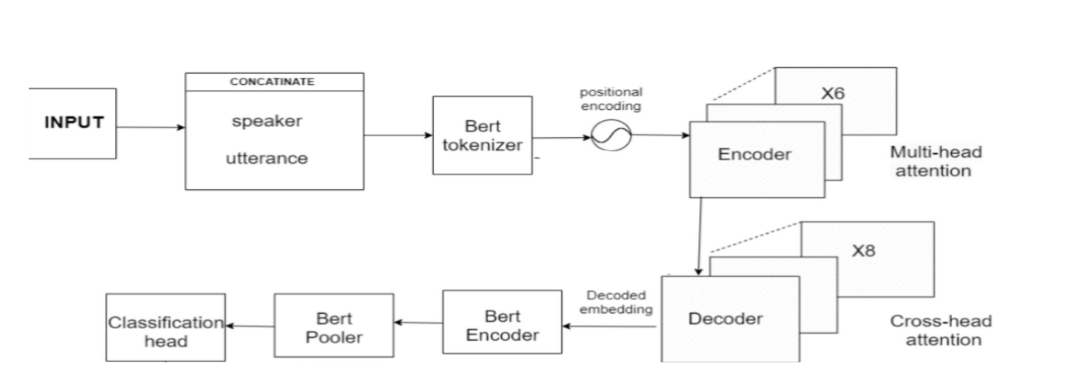


The process begins with transforming the input into numerical values using the BERT tokenizer. These numerical representations are then fed into a Bidirectional Long Short-Term Memory (BiLSTM) model to capture sequential relationships within the data. Next, a linear projection is applied for dimension correction. Following this, the pre-trained BERT encoder and Pooler layers are utilized to further encode the contextual information. Finally, a classification head is employed for prediction

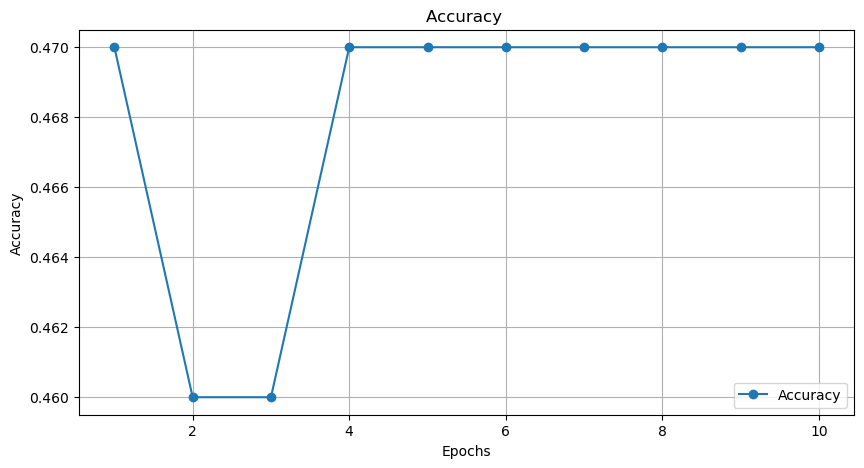


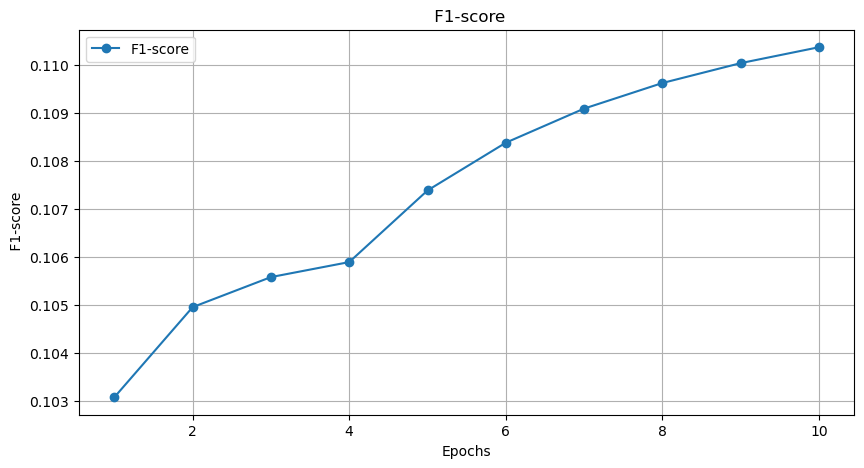


**M2**



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:

Bi-LSTM models have demonstrated superior capability in understanding and capturing sequential patterns in data, due to their temporal nature. In contrast, Model 1, being less complex, is more adept at handling smaller datasets, where it can generalize effectively without requiring vast amounts of data. On the flip side, the more complex model, characterized by a larger parameter set, demands a substantial dataset to reach its full potential and avoid overfitting.