CSE556 NLP Assignment-4

Arnav Goel	*Medha Hira*	*Siddharth Rajput*	*Amil Bhagat*
2021519	*2021265*	*2021102*	*2021309*

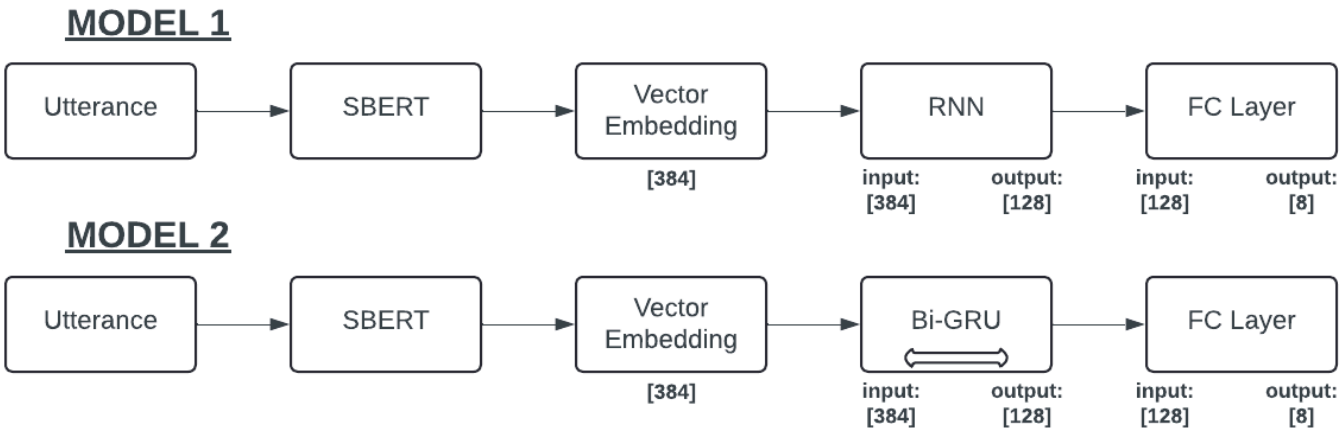
Note: We choose Point 2 for both the tasks

TASK 1 - ERC (Emotion Recognition in Conversation)

1. Two model checkpoints M1 and M2 in proper format.[5*2=10]

We have submitted two checkpoints for this tas, which are models trained on the ERC task:
'M1_Task1.pth'
'M2_Task1.pth'

2. Well-labeled model architectures used for both M1 and M2.[5*2=10]



We made architectures for this task:

4. Properly mention which of the two architectures was better and why.[5]
Out of the 2 architectures, M2 was better as:

- We employed the use of Bi-Directional GRU.

- The bi-directional nature of the architecture helped us look at the future as well as the past context to predict the emotion of the current utterance. This improved the prediction of emotion in a conversation setting.
- Moreover, GRUs pose to be a solution to the vanishing gradients problem present in the RNNs.

5. Proper report explaining the intuition behind the models, splits, and everything relevant.[5]

- **As 1st step,** we converted the sentences in each data sample into Sentence embeddings using the Sentence-Transformers library. We employed the model: **"`all-MiniLM-L6-v2`".**
- It leverages the attention mechanism as a part of its underlying architecture and converts the sentences into embeddings (vector representations) that capture their semantic content, including emotional nuances. Following this, we treat it like a normal sequence labeling task.

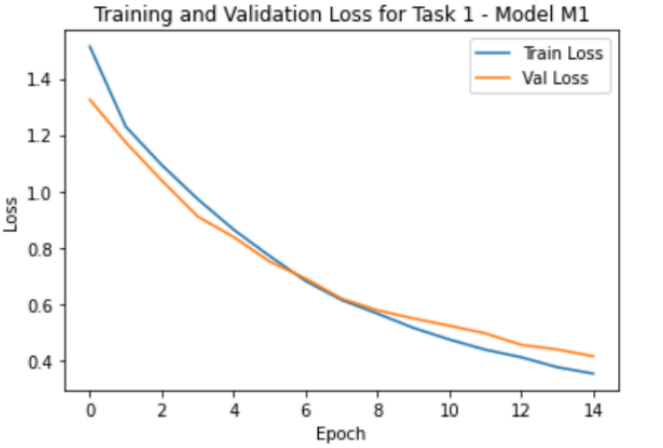
- **Model M1 for ERC task:**
  - Following this, we have implemented a simple (single-layer) **Recurrent Neural Network (RNN) as model M1 for this task**. The use of RNNs is justified due to the sequential nature of the task at hand.
  - The RNN accepts an input with a dimension of 384 and produces an output with a dimension of 128. Subsequent to the RNN, there is a fully connected layer that has an input size of 128 and an output size of 8.
  - These 8 output neurons are mapped to 7 distinct emotion classes (and one padding class -> -1) categories present in the dataset.
  - The intuition behind using an RNN stems from the recurrent nature of the fact, i.e., emotion prediction in the conversation would depend on the emotional states of sentences said before the current sentence.

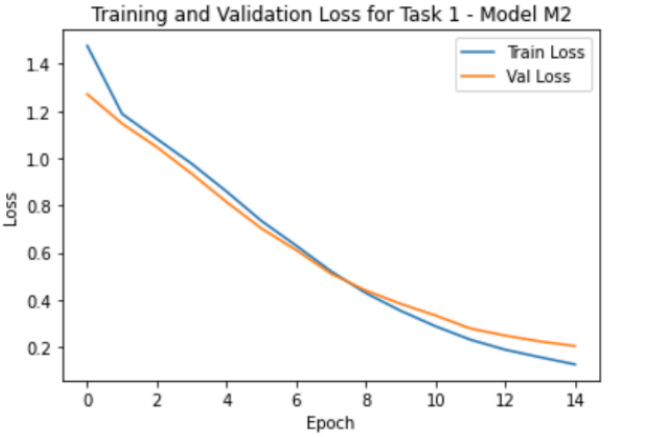
- **Model M2 for ERC task:**
  - Following this, we have implemented a **Bi-direction Gated Recurrent Unit (Bi-GRU) based model as M2 for this task**.
  - The intuition behind using a Bi-GRU as an improvement to the RNN architecture (in M1) is the bi-directional nature of the GRU, as emotional cues for a given utterance can lie in the present or future. Further, GRUs are better as compared to RNNs in terms of vanishing gradients.
  - We use a GRU unit instead of LSTM because GRUs are computationally cheaper to train and more stable and robust in their training.
  - The Bi-GRU layer accepts an input with a dimension of 384 and produces a hidden layer output with a dimension of 256. Subsequent to the Bi-GRU layer, there is a fully connected layer that has an input size of 256 and an output size of 8.

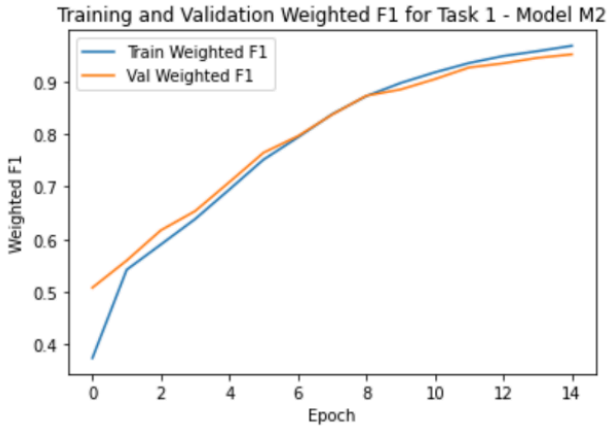
- These 8 output neurons are mapped to 7 distinct emotion classes (and one class, i.e., -1 for the masked/padded entries in the batch), which are present in the dataset.

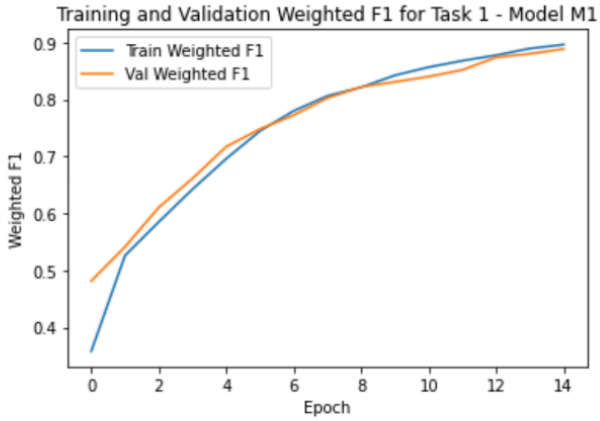
6. Add train loss and val loss vs. epochs plots for each model in the report. [2.5*2*2=10]

- We train both models for 15 epochs to ensure a fair comparison between the two.
- We plot the loss curves on the train and validation datasets for the 15 epochs.
- We additionally plot the curves for Weighted F1 scores on the training and validation dataset over the 15 epochs.









7. Final Results Table (on the validation set):

Evaluation Metric	Model M1	Model M2
Weighted-F1	0.888	**0.952**
Macro-F1	0.861	**0.933**

Subtask under Task 1 - Emotion Flip Detection

- As specified in the assignment, we additionally performed an emotion flip detection task using the same models, M1 and M2, that were trained on the ERC task.
- The methodology is described as follows:
    - We used the models M1 and M2 described above to predict emotions on the validation data set.
    - Then, for each speaker, we made a dictionary with all the episodes. It had more than one utterance, and subsequently, we got the predicted and true emotion labels.
    - We checked for flips in this and marked the valid and invalid flips using our algorithm.
    - We report the accuracy of flip detection in the table below for the two models **on the VALIDATION DATASET**:

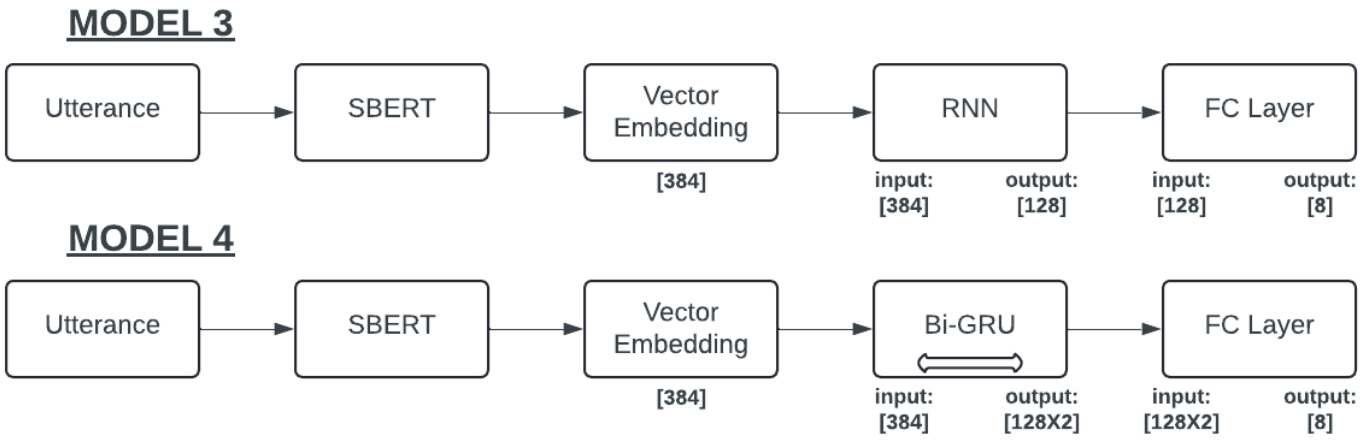
Evaluation Metric	Model M1	Model M2
Accuracy (Flip Detection)	0.825	**0.921**

TASK 2 - EFR (Emotion Flip Reasoning)

1. Two model checkpoints M3 and M4 in proper format.[5*2=10]

‘M1_Task2.pth’
‘M2_Task2.pth’

2. Well-labeled model architectures used for both M3 and M4.[5*2=10]



4. Properly mention which of the two architectures was better and why.[5]
Out of the 2 architectures, **M4(Bi-GRU) was better as**:

- We employed a bi-directional gated recurrent unit layer (Bi-GRU).
- The bi-directional nature of the architecture helped us look at the future as well as the past context to predict the triggers for emotion flips. Since the target sentence was the last utterance in the episode (as specified in the comments), making use of a model that takes context bi-directionally helped to improve performance drastically on our task without even giving emotion as input to the model.
- Moreover, GRUs pose to be a solution to the vanishing gradients problem present in the RNNs.

5. Proper report explaining the intuition behind the models, splits, and everything relevant.[5]

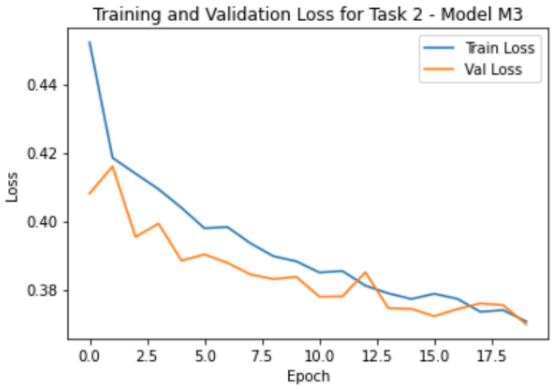
- **As 1st step,** we converted the sentences in each data sample into Sentence embeddings using the Sentence-Transformers library. We employed the model: **"`all-MiniLM-L6-v2`".**
- It leverages the attention mechanism as a part of its underlying architecture and converts the sentences into embeddings (vector representations) that capture their semantic content, including emotional nuances. Following this, we treat it like a normal sequence labeling task.

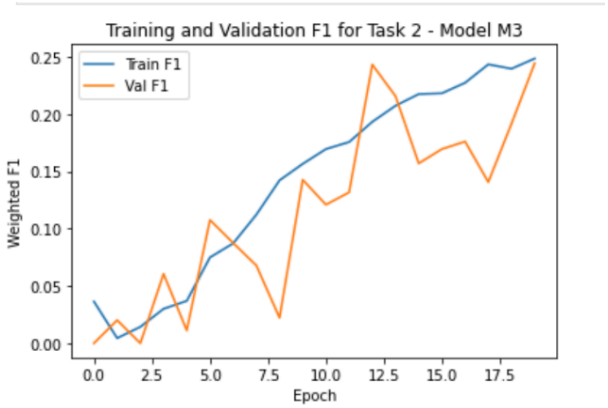
- **Model M1 for EFR task:**
    - Following this, we have implemented a simple (single-layer) Recurrent Neural Network. The use of RNNs is justified due to the sequential nature of the task at hand. We assume that knowing the previous context of triggers can help the model learn which kind of a statement is a trigger and which is not.
    - The RNN accepts an input with a dimension of 384 and produces an output with a dimension of 128. Subsequent to the RNN, there is a fully connected layer that has an input size of 128 and an output size of 8.
    - These 8 output neurons are mapped to 7 distinct emotion classes (and one padding class -> -1) categories present in the dataset.
    - This, however, performs very poorly, as seen in the results, because it fails at the basic notion of the task.

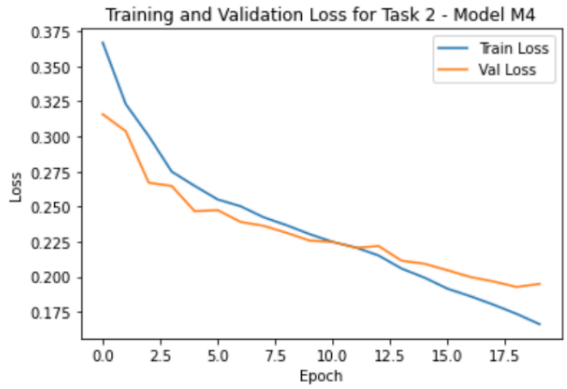
- **Model M2 for EFR task:**
    - Following this, we have implemented a Bi-direction Gated Recurrent Unit.
    - The intuition behind using a Bi-GRU as an improvement to the RNN architecture (in M1) is the bi-directional nature of the GRU, as cues for a given utterance being in a trigger can lie in the present and the future. This is because the target utterance is the last in the episode. Knowing cues for future utterances helps the model better understand which utterances are triggers and which are not.

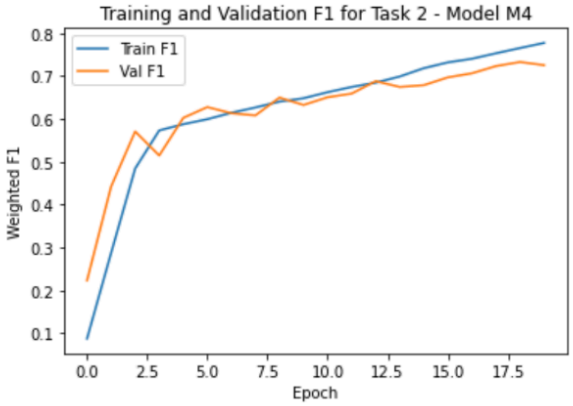
- The Bi-GRU accepts an input with a dimension of 384 and produces an output with a dimension of 128. Subsequent to the Bi-GRU, there is a fully connected layer that has an input size of 128 and an output size of 8.
- These 8 output neurons are mapped to 7 distinct emotion classes (and one padding class -> -1) categories present in the dataset.

6. Add train loss and val loss vs. epochs plots for each model in the report. [2.5*2*2=10]









7. Final Results Table (on the validation set):

Evaluation Metric	Model M3	Model M4
Binary-F1	0.244	**0.725**
Weighted-F1	0.218	**0.631**
Macro-F1	0.122	**0.363**

Credit Statement:

While each member contributed equally to all tasks, we are mentioning a division of tasks just for the sake of the question here.

- **Amil Bhagat (Roll No: 2021309)** - Model 3, Task 2
- **Arnav Goel (Roll No: 2021519)** - Model 4, Task 2, Report and Analysis
- **Medha Hira (Roll No: 2021265)** - Model 2, Task 1, Report and Analysis
- **Siddharth Rajput (Roll No: 2021102)** - Model 1, Task 1