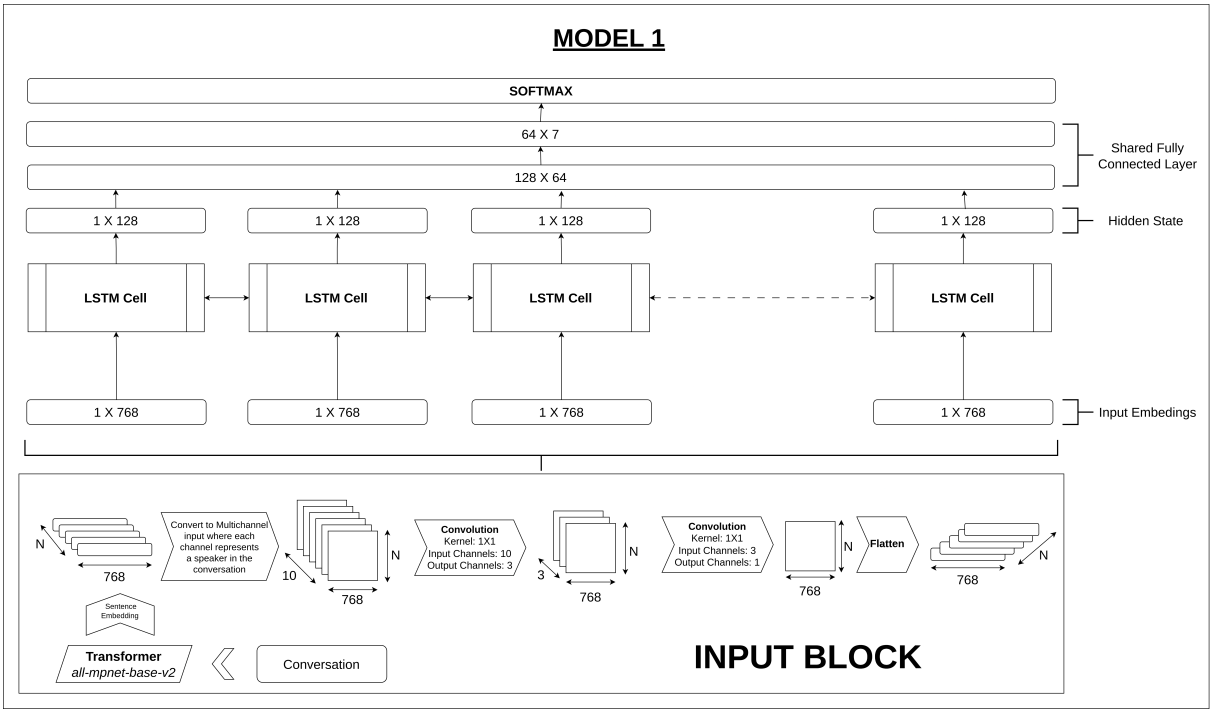
Natural Language Processing Assignment - 4

Task 1 - ERC

Model 1

Architecture



Intuition

Input Block: We want to get good results on the conversational data. Predicting the emotion of an utterance in a conversation can be more difficult than predicting the emotion of a statement or just an English sentence since emotion in the current utterance can be very dependent upon the previous utterances. Moreover, speaker information is also important for the same reasons.

So to preserve these aspects of the conversation we have converted our input into a 3-dimensional input matrix. Where 1 axis preserves the speaker information, 1 axis preserves the information of the timestamps in the conversation and 1 axis contains the input embeddings.

Input Embeddings: We have used sentence-transformer library to be specific we have used the '**all-mpnet-base-v2**' transformer to get the utterance embeddings. We have selected this transformer based on the statistics listed on the website of the sentence-transformer library.

Convolution: To capture the speaker's essence and the timestamp essence in the conversation we have used convolution. We have treated different speakers as channels. We have considered the maximum number of speakers as 10 (8 in the training set and 7 in the validation set). Since LSTM takes single channel input we have reduced the channels from 10 to 1. We want CNN to only capture speaker and timestamp data so we have preserved the utterance embedding dimension i.e. 768. Language understanding task we have left to LSTM.

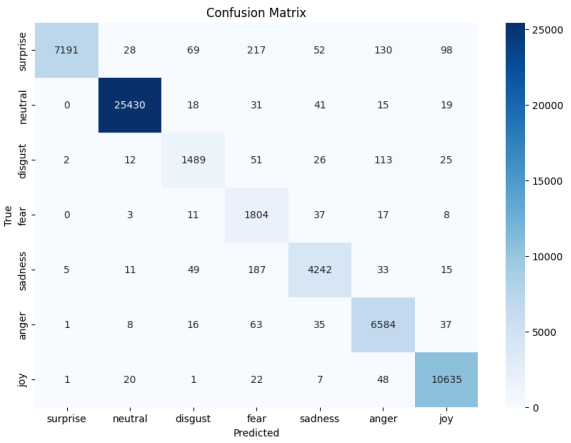
LSTM Cell: It takes 1X768 embedding as input and n such inputs. And gives n outputs for each utterance.

FC Layer: It is a shared layer which is shared among all the LSTM cells. It then predicts one of the 7 emotions for the input.

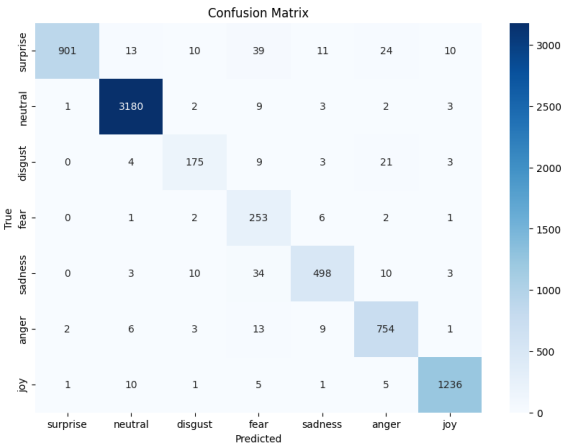
F1 Scores

Emotion	Training - F1 Score	Validation - F1 Score
surprise	0.9597597597597598	0.9419759539989545
neutral	0.9959660047781302	0.9911173445535297
disgust	0.8834173835657075	0.8373205741626795
fear	0.8479435957696826	0.8070175438596492
sadness	0.944555778223113	0.9146005509641872
anger	0.9622917275650396	0.9389788293897884
joy	0.9860460803857031	0.9825119236883944
Macro F1	**0.9399971900067338**	**0.9162175315167403**

Confusion Matrix

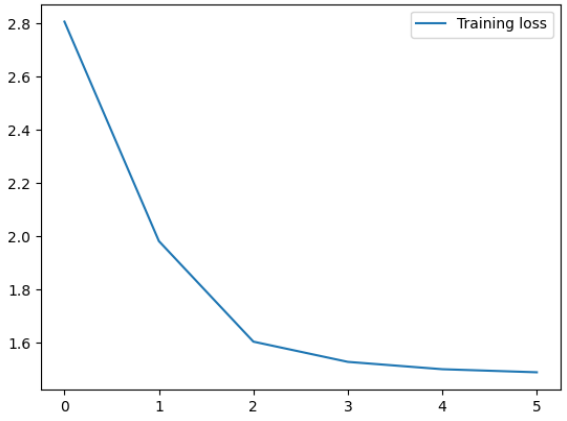


Training

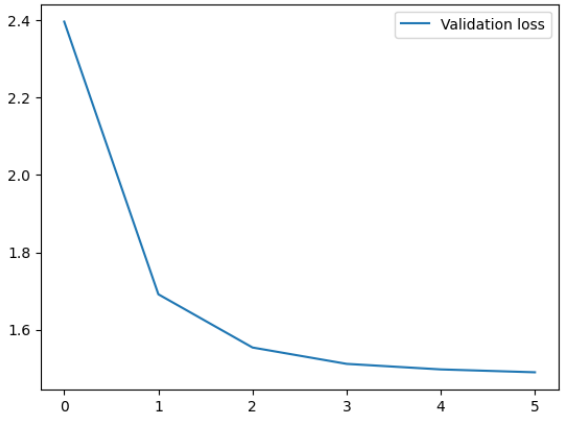


Validation

Loss v/s Epoch Plots



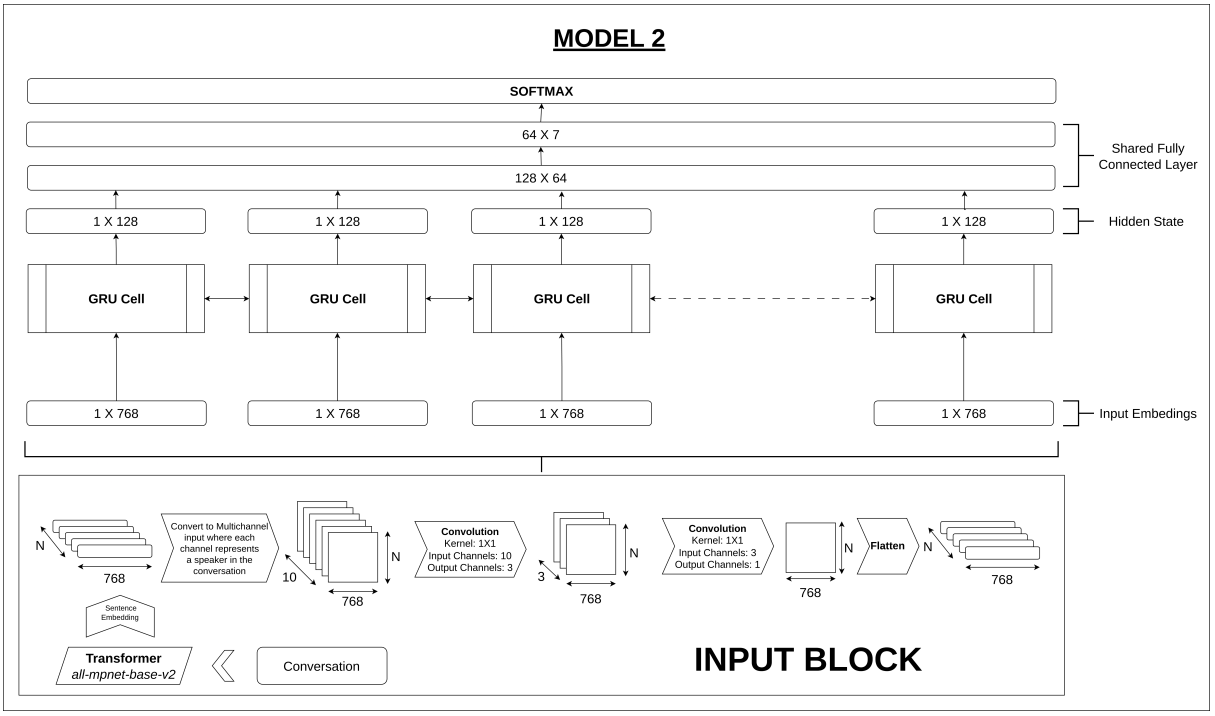
Training loss v/s Epoch



Validation loss v/s Epoch

Model 2

Architecture



Intuition

Input Block: We want to get good results on the conversational data. Predicting the emotion of an utterance in a conversation can be more difficult than predicting the emotion of a statement or just an English sentence since emotion in the current utterance can be very dependent upon the previous utterances. Moreover, speaker information is also important for the same reasons.

So to preserve these aspects of the conversation we have converted our input into a 3-dimensional input matrix. Where 1 axis preserves the speaker information, 1 axis preserves the information of the timestamps in the conversation and 1 axis contains the input embeddings.

Input Embeddings: We have used sentence-transformer library to be specific we have used the '**all-mpnet-base-v2**' transformer to get the utterance embeddings. We have selected this transformer based on the statistics listed on the website of the sentence-transformer library.

Convolution: To capture the speaker's essence and the timestamp essence in the conversation we have used convolution. We have treated different speakers as channels. We have considered the maximum number of speakers as 10 (8 in the training set and 7 in the validation set). Since GRU takes single channel input we have reduced the channels from 10 to 1. We want CNN to only capture speaker and timestamp data so we have preserved the utterance embedding dimension i.e. 768. Language understanding task we have left to GRU.

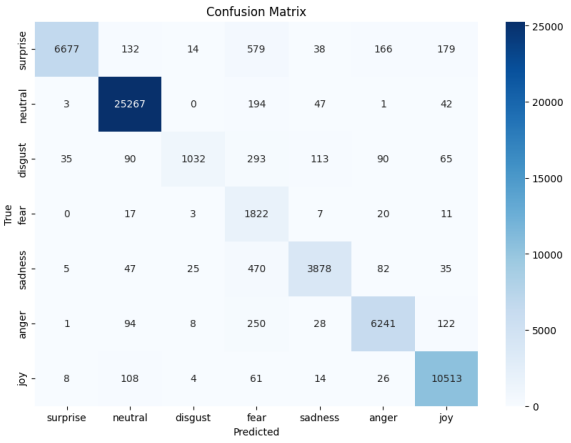
GRU Cell: It takes 1X768 embedding as input and n such inputs. And gives n outputs for each utterance.

FC Layer: It is a shared layer which is shared among all the GRU cells. It then predicts one of the 7 emotions for the input.

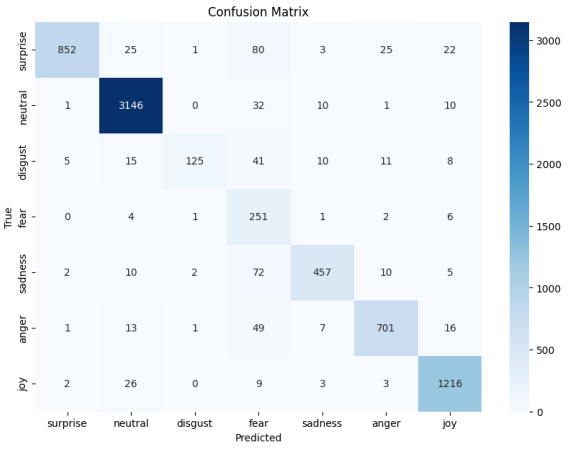
F1 Scores

Emotion	Training - F1 Score	Validation - F1 Score
surprise	0.9200771668733637	0.9107429182255479
neutral	0.9848954374476213	0.9771703680695761
disgust	0.7360912981455064	0.7246376811594203
fear	0.6566948999819788	0.6282853566958698
sadness	0.8948886581285336	0.8713060057197332
anger	0.9335826477187733	0.9097988319273199
joy	0.9688954426063314	0.956726986624705
Macro F1	**0.8707322215574441**	**0.8540954497745961**

Confusion Matrix

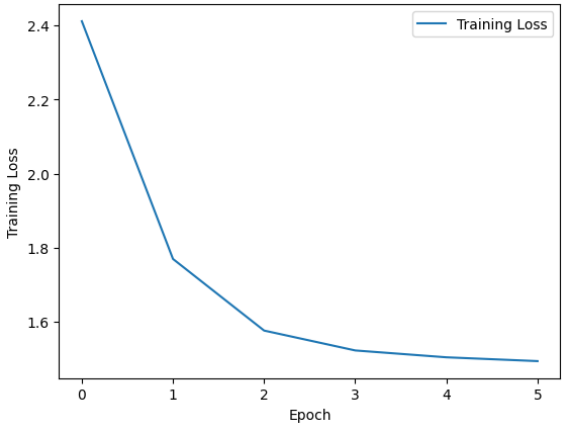


Training

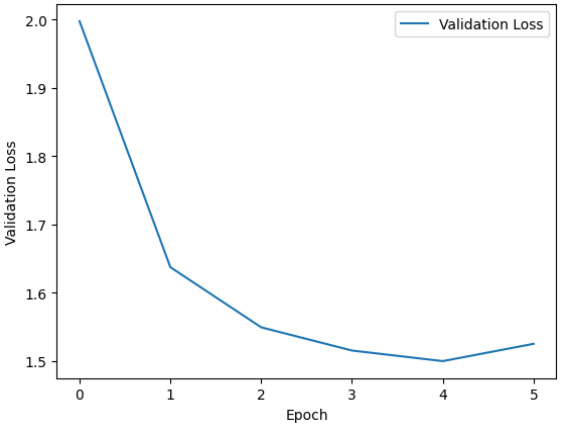


Validation

Loss v/s Epoch Plots:

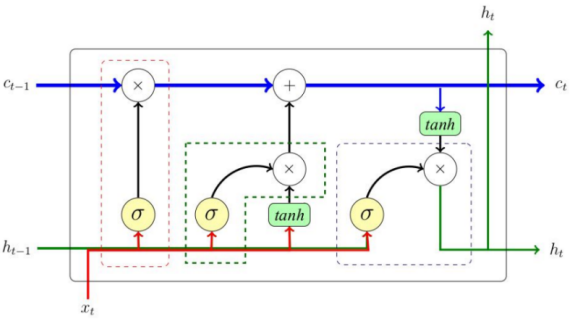


Training Loss v/s Epoch

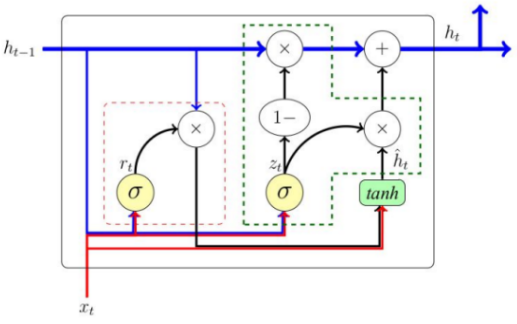


Validation Loss v/s Epoch

Model 1 v/s Model 2



LSTM Cell



GRU Cell

Note: The GRU unit controls the flow of information like the LSTM unit, but without having to use a ***memory unit***. It just exposes the full hidden content without any control.

In ERC we are predicting the emotion of an utterance and **emotion as an entity depends more upon the context of the sentence and less upon the context of the conversion**. Since LSTM controls the injection of current input in the hidden content so it can inject sufficient information for predicting the emotion of the current utterance. While in the case of GRU whole hidden information is exposed to FC predicting the emotion of the current utterance is difficult.

The above hypothesis can be verified from the results on the validation dataset as shown below.

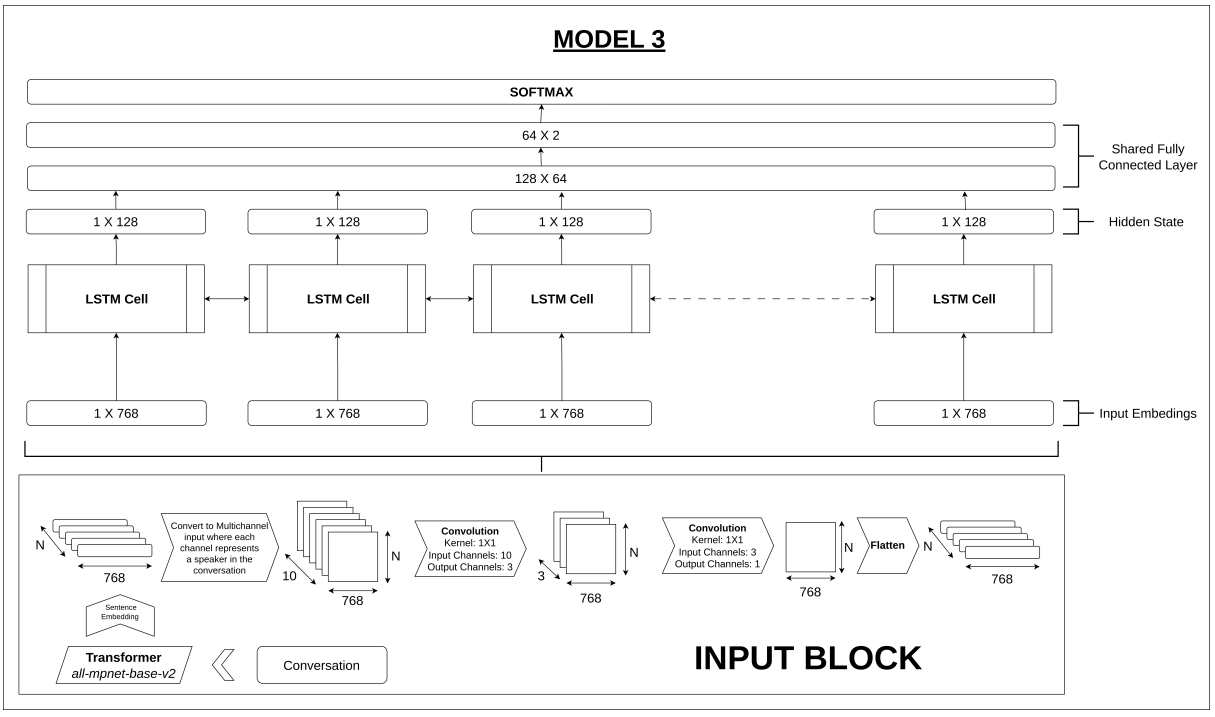
Emotions	Model 1 (Validation F1 scores)	Model 2 (Validation F1 scores)
surprise	0.9419759539989545	0.9107429182255479

Emotions	Model 1 (Validation F1 scores)	Model 2 (Validation F1 scores)
neutral	0.9911173445535297	0.9771703680695761
disgust	0.8373205741626795	0.7246376811594203
fear	0.8070175438596492	0.6282853566958698
sadness	0.9146005509641872	0.8713060057197332
anger	0.9389788293897884	0.9097988319273199
joy	0.9825119236883944	0.956726986624705
Macro F1	**0.9162175315167403**	**0.8540954497745961**

Task 2 - EFR

Model 3

Architecture



Intuition

Input Block: We want to get good results on the conversational data. To predict the emotion flip trigger for a particular utterance in the conversation speaker information and the timestamp of each utterance is important.

So to preserve these aspects of the conversation we have converted our input into a 3-dimensional input matrix. Where 1 axis preserves the speaker information, 1 axis preserves the information of the timestamps in the conversation and 1 axis contains the input embeddings.

Input Embeddings: We have used sentence-transformer library to be specific we have used the '**all-mpnet-base-v2**' transformer to get the utterance embeddings. We have selected this transformer based on the statistics listed on the website of the sentence-transformer library.

Convolution: To capture the speaker's essence and the timestamp essence in the conversation we have used convolution. We have treated different speakers as channels. We have considered the maximum number of speakers as 10 (8 in the training set and 7 in the validation set). Since LSTM takes single channel input we have reduced the channels from 10 to 1. We want CNN to only capture speaker and timestamp data so we have preserved the utterance embedding dimension i.e. 768. Language understanding task we have left to LSTM.

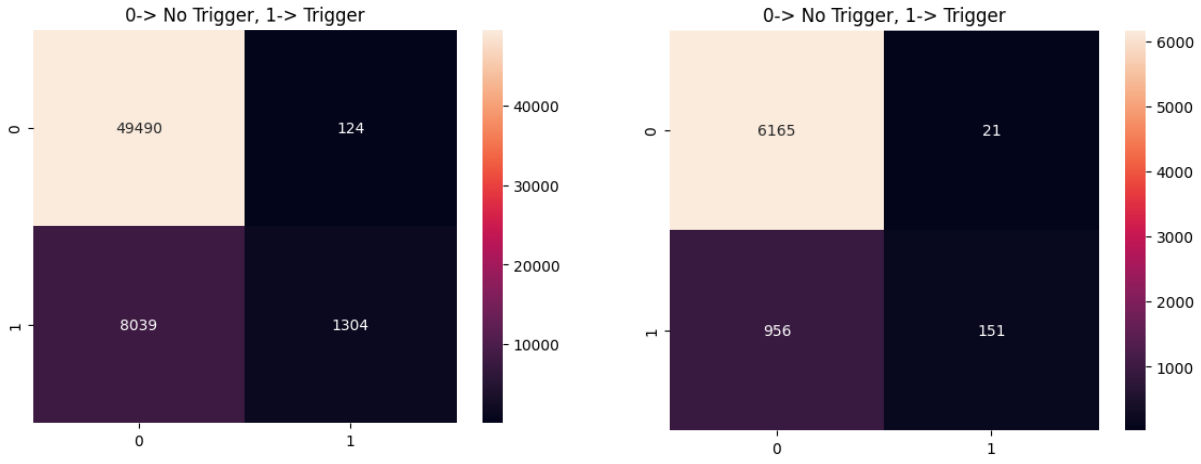
LSTM Cell: It takes 1X768 embedding as input and n such inputs. And gives n outputs for each utterance.

FC Layer: It is a shared layer which is shared among all the LSTM cells. It then predicts one of the 2 classes i.e. trigger or non-trigger for each input, being trigger for the last utterance in a conversation.

F1 Scores

	Training - F1 Score	Validation - F1 Score
F1 Score	0.5829718756921273	0.5813509828901917

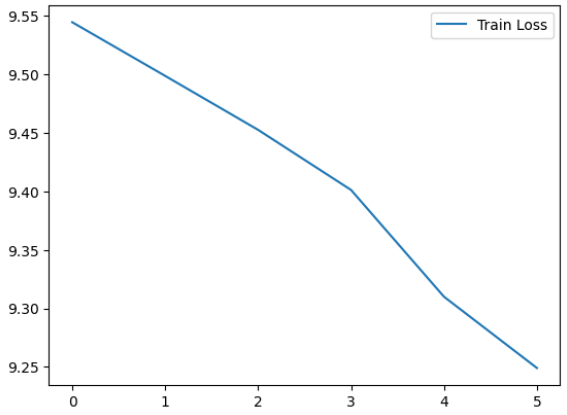
Confusion Matrix

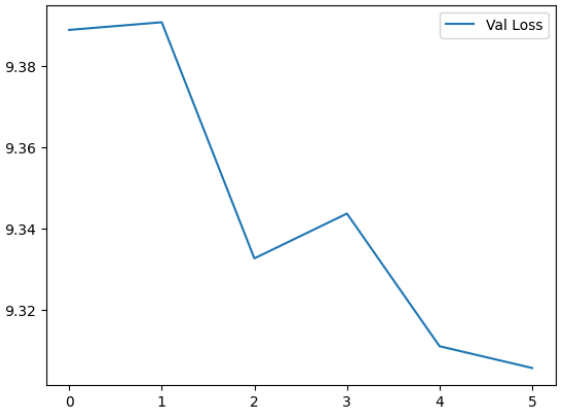


Training

Validation

Loss v/s Epoch Plots



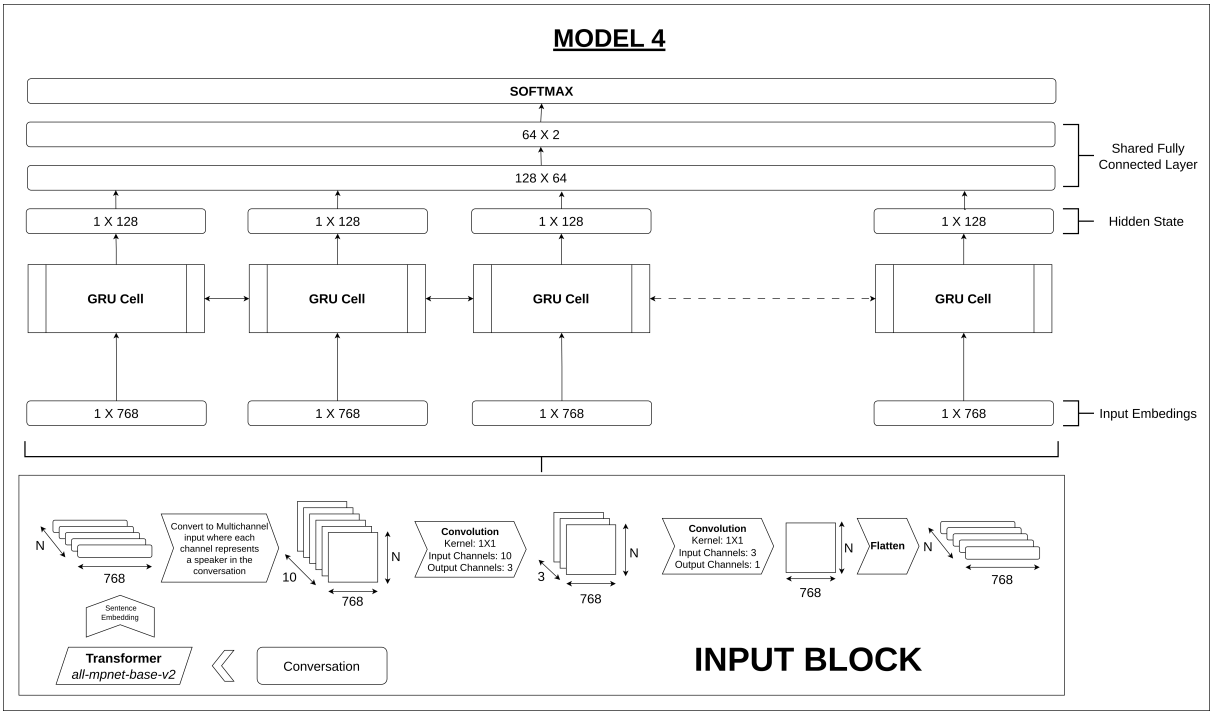


Training Loss v/s Epoch

Validation Loss v/s Epoch

Model 4

Architecture



Intuition

Input Block: We want to get good results on the conversational data. To predict the emotion flip trigger for a particular utterance in the conversation speaker information and the timestamp of each utterance is important.

So to preserve these aspects of the conversation we have converted our input into a 3-dimensional input matrix. Where 1 axis preserves the speaker information, 1 axis preserves the information of the timestamps in the conversation and 1 axis contains the input embeddings.

Input Embeddings: We have used sentence-transformer library to be specific we have used ‘

all-mpnet-base-v2’ transformer to get the utterance embeddings. We have selected this transformer based on the statistics listed on the website of the sentence-transformer library.

Convolution: To capture the speaker's essence and the timestamp essence in the conversation we have used convolution. We have treated different speakers as channels. We have considered the maximum number of speakers as 10 (8 in the training set and 7 in the validation set). Since GRU takes single channel input we have reduced the channels from 10 to 1. We want CNN to only capture speaker and timestamp data so we have preserved the utterance embedding dimension i.e. 768. Language understanding task we have left to GRU.

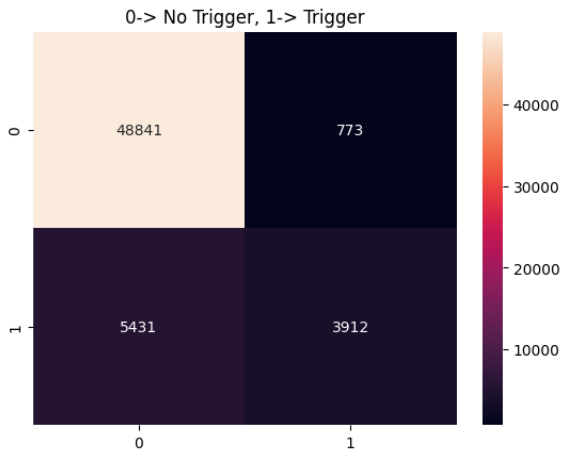
GRU Cell: It takes 1X768 embedding as input and n such inputs. And gives n outputs for each utterance.

FC Layer: It is a shared layer which is shared among all the GRU cells. It then predicts one of the 2 classes i.e. trigger or non-trigger for each input, being trigger for the last utterance in a conversation.

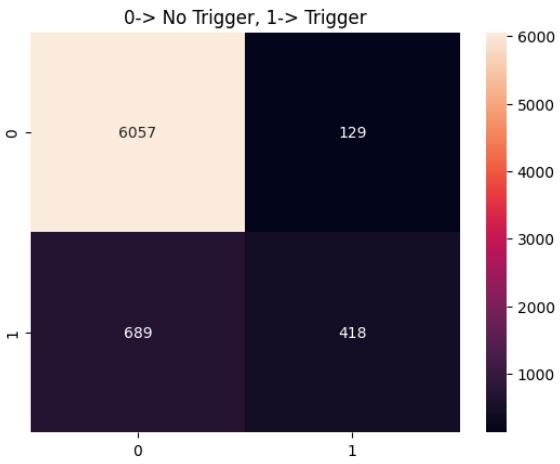
F1 Scores

	Training - F1 Score	Validation - F1 Score
F1 Score	0.7490111759176963	0.7210937052935437

Confusion Matrix

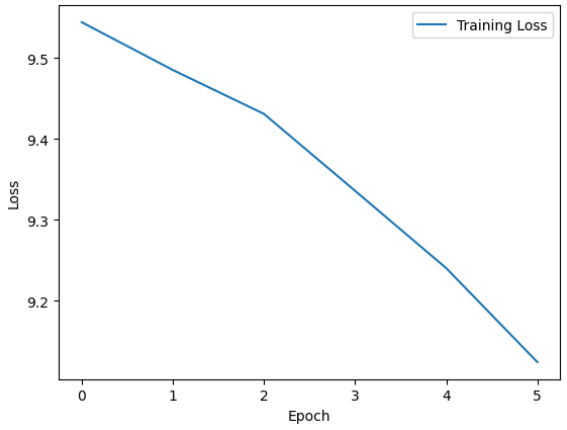


Training

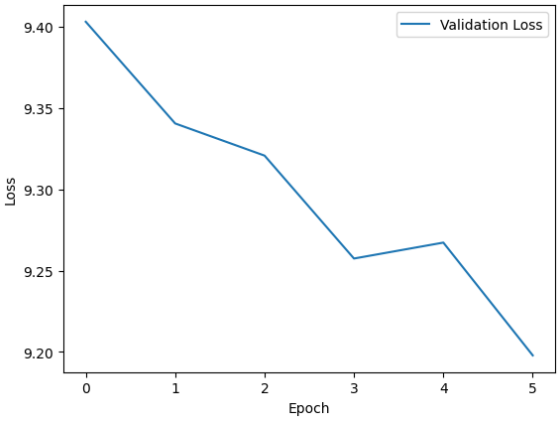


Validation

Loss v/s Epoch Plots

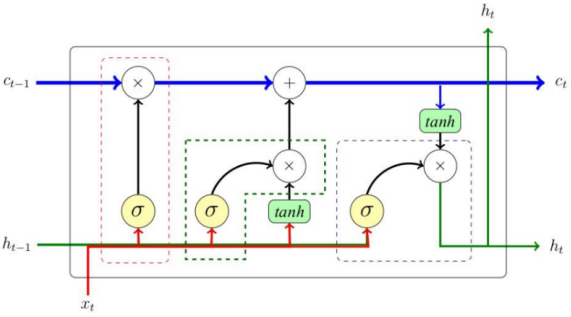


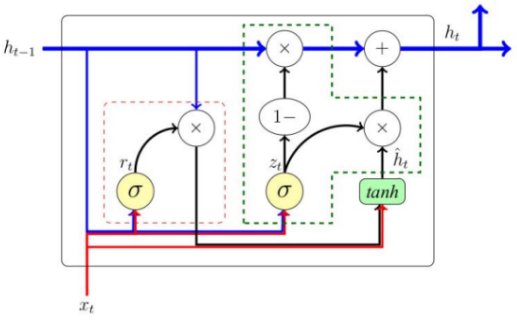
Training Loss v/s Epoch



Validation Loss v/s Epoch

Model 3 v/s Model 4





LSTM Cell **GRU Cell**

The GRU unit controls the flow of information like the LSTM unit, but without having to use a ***memory unit***. It just exposes the full hidden content without any control.

In EFR we are predicting the emotion flip triggers for an utterance in the conversation. For predicting triggers **we need context of the whole conversation**. Since LSTM controls the injection of current input in the hidden content, it can inject information about the current utterance in the hidden content which might be unnecessary and can cause more loss of previous context information. While GRU exposes whole hidden information which is precious to predict the emotion flip triggers for the last utterance.

The above hypothesis can be verified from the results on the validation dataset as shown below.

	Model 3	**Model 4**
Validation F1 scores	0.5813509828901917	0.7210937052935437
Training F1 scores	0.5829718756921273	0.7490111759176963

Contribution:

Task 1 - Manav Mittal, Utkarsh Venaik, Lakshay Kumar, Akash Kushwaha

Task 2 - Manav Mittal, Utkarsh Venaik, Lakshay Kumar, Akash Kushwaha

Task Report - Manav Mittal, Utkarsh Venaik, Lakshay Kumar, Akash Kushwaha