EFR AND ERC IN ENGLISH CONVERSATIONS

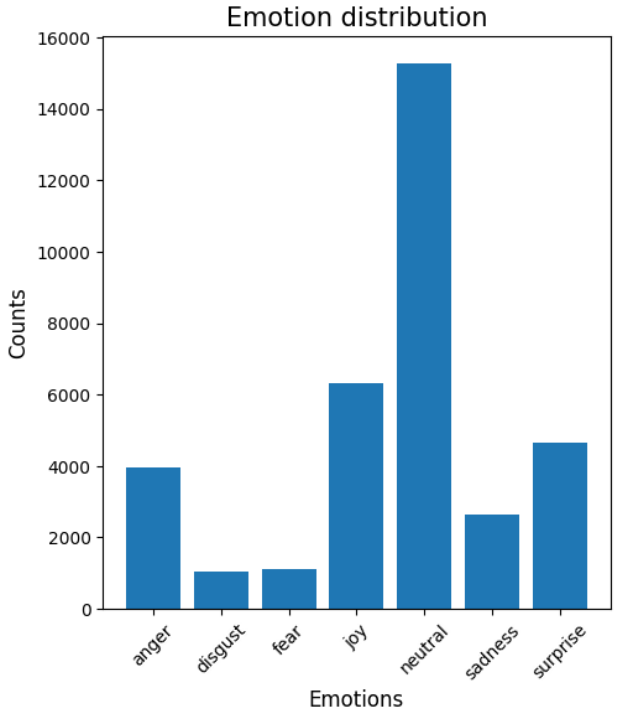
M. Dore - R. Murgia - G.M. Marongiu - G. Gaiani

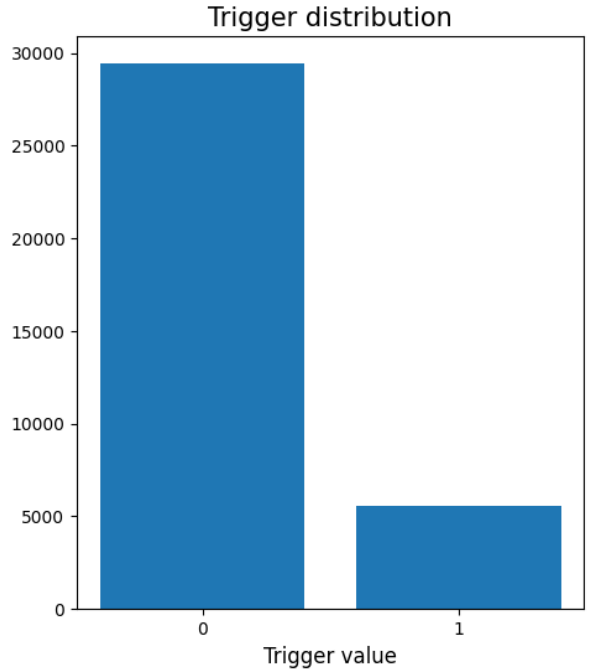
A.Y. 2023/2024

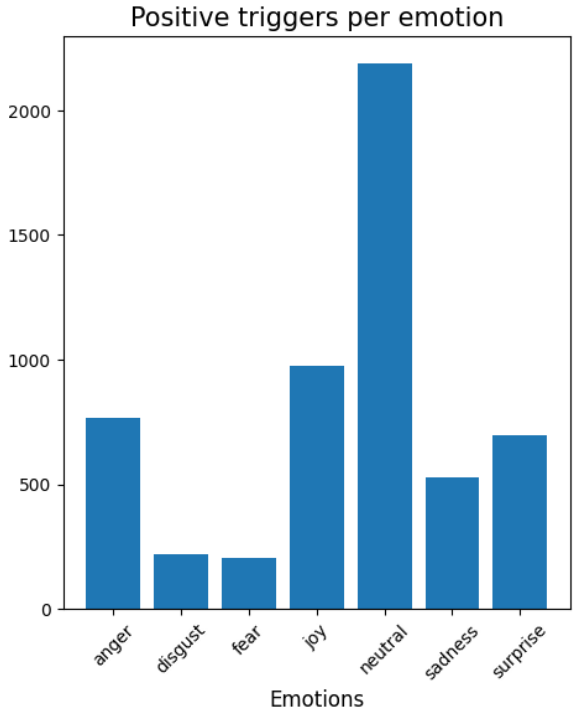
NLP Course Project

DATASET ANALYSIS

DATA DISTRIBUTION: EMOTION AND TRIGGER







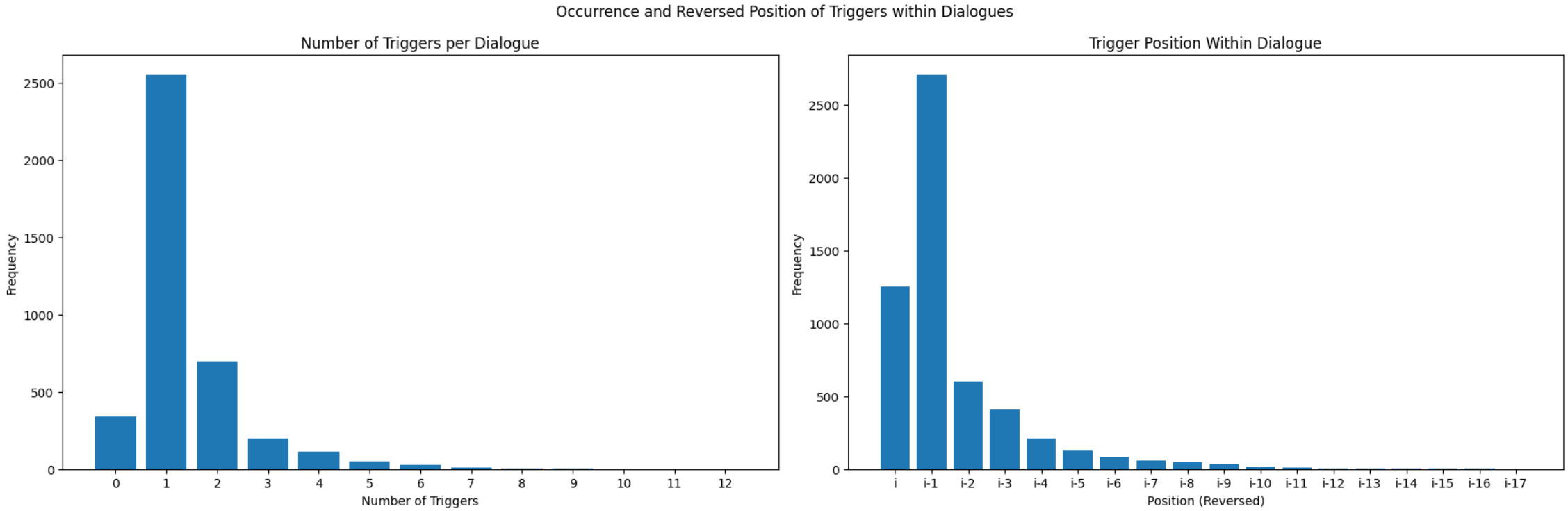
Class Imbalance:

- Significant imbalance in emotions and triggers.
- "Neutral" emotion and "0" trigger are predominant.
- Define a class weights to balance this distribution.

Emotion Flips:

- No specific emotion triggers a flip more than others.
- Equal occurrences of emotion flips for each emotion.

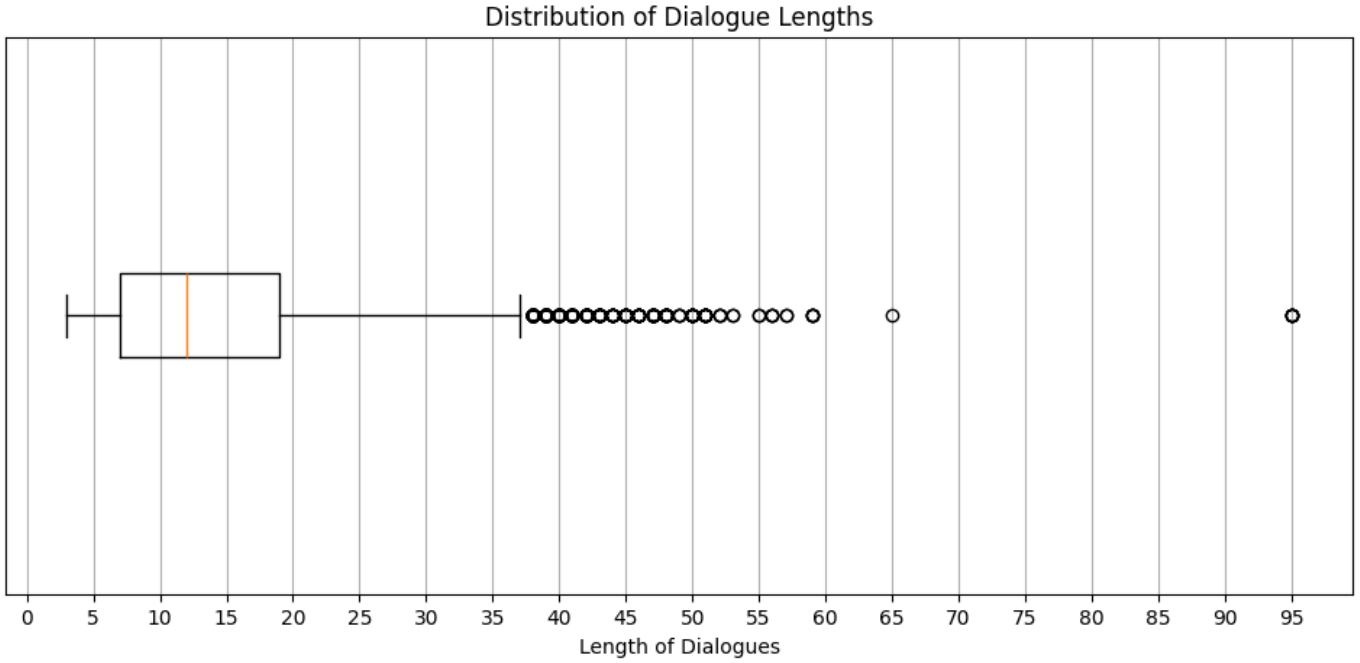
DATA DISTRIBUTION: TRIGGERS WITHING DIALOGUES



Positive Triggers:

- Mostly appear in the final part of dialogues.
- Peak in second-to-last utterance.

DATA DISTRIBUTION: DIALOGUE LENGTHS



Utterance Length:

- Tokenized utterance range in lengths.
- Performance slowdowns depending on the padding policy.

PRE-PROCESSING

Dialogue	Speakers	Emotions	Utterances	Triggers
	Ross	neutral	Yes that's right.	0
	The Instructor	surprise	Why?	0
	Ross	neutral	I tired attacking two women, did not work.	0
	The Instructor	surprise	What?!	0
	Ross	neutral	No, I mean it's okay, I mean, they're-they're ...	0
	Ross	anger	In fact, I-I-I was married to one of them.	0
	The Instructor	anger	Let me get this straight man, you attacked you...	0
	Ross	surprise	Oh, no!	0
	Ross	surprise	No-no!	0
	Ross	surprise	No. I tired!	1
	Ross	surprise	But I couldn't.	0

$[Dialogue_1, \dots, Dialogue_D]$

Padding
All dialogues same length
All utterances same number of tokens

Tokenization
Using the *bert-base-uncased* tokenizer

[CLS], $utterance_{1,1}$, [SEP], ..., [SEP], $utterance_{S,1}$

⋮

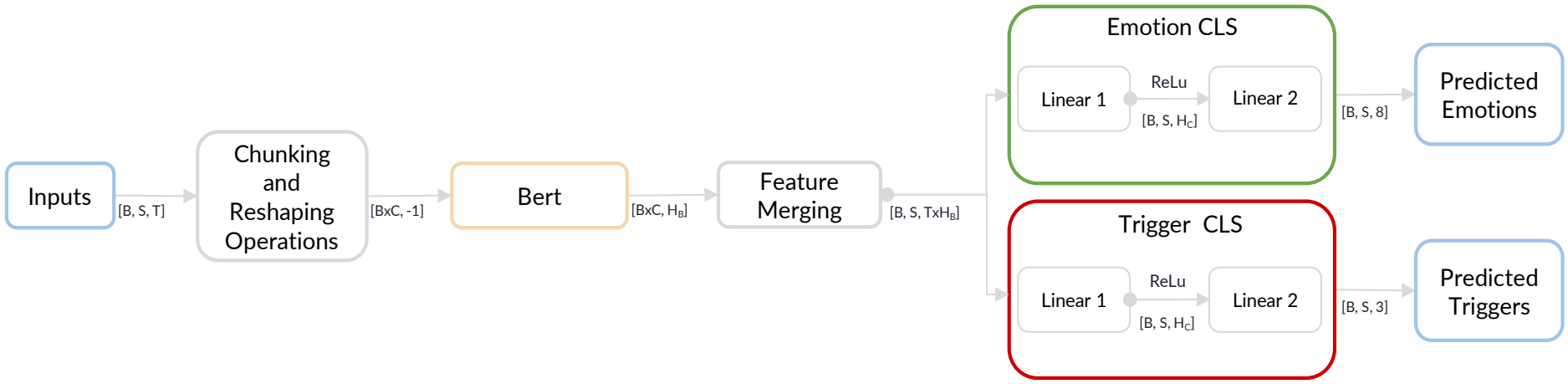
[CLS], $utterance_{1,D}$, [SEP], ..., [SEP], $utterance_{S,D}$

```
Input={
        "input_ids": [B, S, T]
        "attention_mask": [B, S, T]
        "token_type_ids": [B, S, T]
        }
```

B: batch size, S: number of utterances in each dialogue, T: length of each utterance

BERT-BASED MODELS

BERT-BASED MODELS STRUCTURE



B: batch size, S: number of utterances in each dialogue, C: number of chunks, T: length of each utterance, H_b = 768, H_C = 512

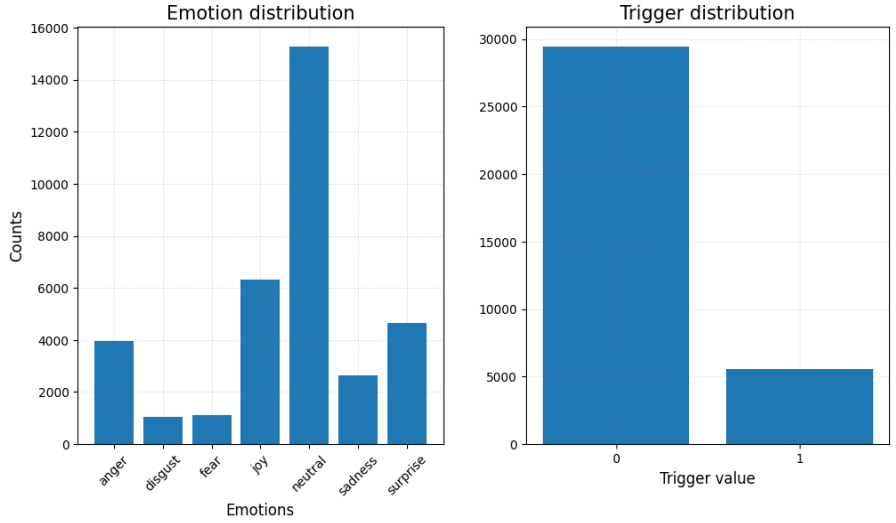
- The input dialogues are chunked into smaller sections to ensure they fit within the Bert's input limit of 512 tokens.
  1. **Method**: Split dialogues into 'C' parts ensuring no sentence breakage.
  2. **Automatic Calculation**: Based on token count, maximum sentences, and BERT's token capacity.
- Two distinct models based on BERT base uncased from Hugging Face as a feature extractor:
  1. **Bert Freezed:** Utilizes the standard, out-of-the-box BERT model.
  2. **Bert Full:** Fine-tuned on our specific data to better suit our task requirements.
- Then reconstructed through feature merging ensures that features from the same dialogue are combined post-chunking.
- The Emotion and Trigger classifiers are composed of a sequence of two Linear Layer, with a ReLU function applied between them.
- The final Output is a organized dictionary containing the Logits produced by classifiers.

EXPERIMENTAL SETUP

- **Introduction**
  - Definition of baselines: Random Classifier and Majority Classifier.
  - Compare Bert based model with baselines.

- **Evaluation Metrics**
  - Unrolled Sequence F1: calculated across all utterances.
  - Sequence F1: calculated individually for each dialogue and averaged.

- **Training Methodology**
  - Used a combination of two Cross Entropy Loss.
  - Experimenting with additional class weights to enhance model performance while addressing the phenomenon of class imbalance:



$$Loss_{WhitClassWeights} = \alpha \cdot \text{CrossEntropy}_{\text{emotion}}(y_{\text{true}}, y_{\text{pred}}, w_{\text{emotion}}) + \beta \cdot \text{CrossEntropy}_{\text{trigger}}(y_{\text{true}}, y_{\text{pred}}, w_{\text{trigger}})$$

$$Loss_{WhitoutClassWeights} = \alpha \cdot \text{CrossEntropy}_{\text{emotion}}(y_{\text{true}}, y_{\text{pred}}) + \beta \cdot \text{CrossEntropy}_{\text{trigger}}(y_{\text{true}}, y_{\text{pred}})$$

$$\text{with } \alpha = \beta = 1$$

$$w_i = \frac{N}{K \cdot N_i}$$

w_i: weight of i-th class
N: total number of samples
K: number of classes
N_i: sample of class i

EARLY STOPPING STRATEGY

- **Overview**
  1. Differentiated training monitoring for Freezed Bert and Full Bert models using tailored Early Stopper approaches.

- **Bert Freezed Model Monitoring**
  1. **Early Stopper Application:**
     1. Monitored using a specialized Early Stopper.
     2. Evaluates Unrolled F1 scores for each label individually.
  1. **Weight Management:**
     1. Reloads weights for network portions specific to each label's prediction.
     2. Freezes these network portions post-reloading to isolate and refine other network areas

- **Full Bert Model Monitoring**
  1. **Challenges:**
     1. Full Bert model allows ongoing updates to BERT-associated weights during training, making elective freezing impractical.
  2. **Adopted Strategy:**
     1. Instead of freezing, monitors the aggregation of Unrolled F1 scores calculated from the outputs of individual classifiers.

$$\text{Agg\_Unrolled\_F1} = w_1/2 \cdot f1(y_{\text{true\_unrolled}}, y_{\text{pred\_unrolled}}) + w_2/2 \cdot F1(y_{\text{true\_unrolled}}, y_{\text{pred\_unrolled}})$$

$$\text{with } w_1 = w_2 = 1$$

GRID SEARCH RESULTS FOR BERT FREEZED

Grid Search Methodology:

- Training and validation of **Bert Freezed** model with and without the use of class weights.
- Use of a set of 5 different seeds for each configuration and compute statistics to enhance the robustness of the results.

Loss Name	Class	Unrolled_F1		Sequence_F1	
		Statistics			
		Mean	Std	Mean	Std
CE_with_CW	Trigger	0.525	0.026	0.510	0.012
	Emotion	0.730	0.016	0.370	0.01
CE_without_CW	Trigger	0.577	0.013	0.526	0.016
	Emotion	0.760	0.012	0.367	0.009

Bert Freezed metrics statistics computed on the validation setover the five seeds

Loss Name	Seed	Unrolled_F1		Sequence_F1	
		Emotion	Trigger	Emotion	Trigger
CE_with_CW	42	0.7419	0.5312	0.3787	0.5214
CE_with_CW	69	0.7494	0.4802	0.3670	0.4903
CE_with_CW	90	0.7098	0.5491	0.3680	0.5220
CE_with_CW	1	0.7196	0.5335	0.3562	0.5077
CE_with_CW	77	0.7337	0.5355	0.3820	0.5102
CE_without_CW	42	0.7600	0.5789	0.3577	0.5194
CE_without_CW	69	0.7591	0.5826	0.3731	0.5369
CE_without_CW	90	0.7806	0.5909	0.3803	0.5464
CE_without_CW	1	0.7545	0.5553	0.3588	0.5048
CE_without_CW	77	0.7467	0.5811	0.3690	0.5245

Extract of grid search results of Bert Freezed on the validation set

GRID SEARCH RESULTS FOR BERT FULL

Grid Search Methodology:

- Training and validation of **Bert Full** model with and without the use of class weights.
- Use of a set of 5 different seeds for each configuration and compute statistics to enhance the robustness of the results.

Loss Name	Class	Unrolled_f1		Sequence_F1	
		Statistics			
		Mean	Std	Mean	Std
CE_with_CW	Trigger	0.545	0.033	0.517	0.021
	Emotion	0.718	0.035	0.367	0.009
CE_without_CW	Trigger	0.570	0.005	0.526	0.011
	Emotion	0.775	0.012	0.387	0.005

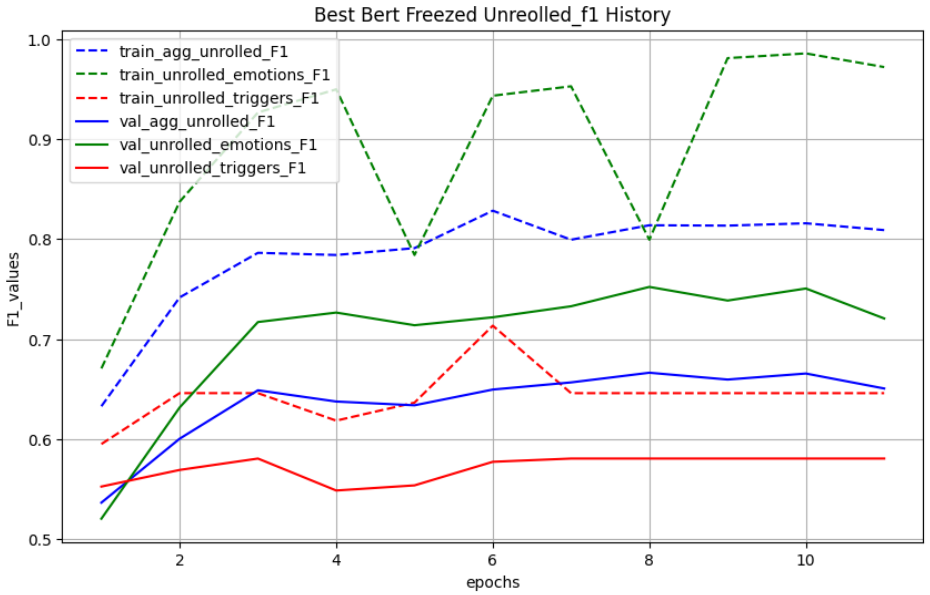
Bert full metrics statistics computed on the validation setover the five seeds

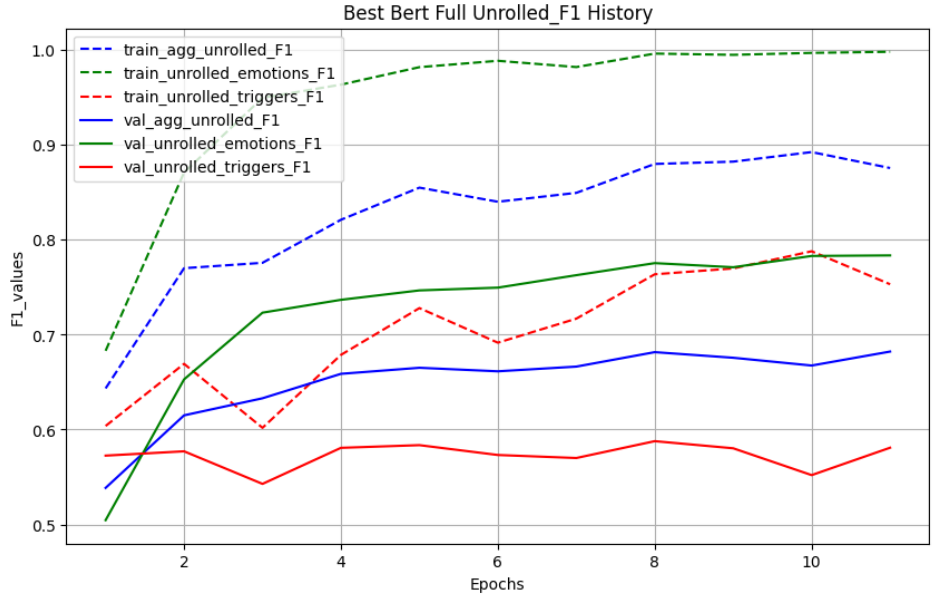
Loss Name	Seed	Unrolled_F1		Sequence_F1	
		Emotion	Trigger	Emotion	Trigger
CE_with_CW	42	0.7035	0.5585	0.3598	0.5218
CE_with_CW	69	0.7212	0.5578	0.3741	0.5222
CE_with_CW	90	0.7590	0.4909	0.3792	0.4813
CE_with_CW	1	0.6678	0.5376	0.3572	0.5247
CE_with_CW	77	0.7422	0.5806	0.3662	0.5381
CE_without_CW	42	0.7955	0.5651	0.3949	0.5205
CE_without_CW	69	0.7720	0.5665	0.3846	0.5096
CE_without_CW	90	0.7807	0.5665	0.3846	0.5096
CE_without_CW	1	0.7674	0.5718	0.3908	0.5349
CE_without_CW	77	0.7626	0.5781	0.3822	0.5386

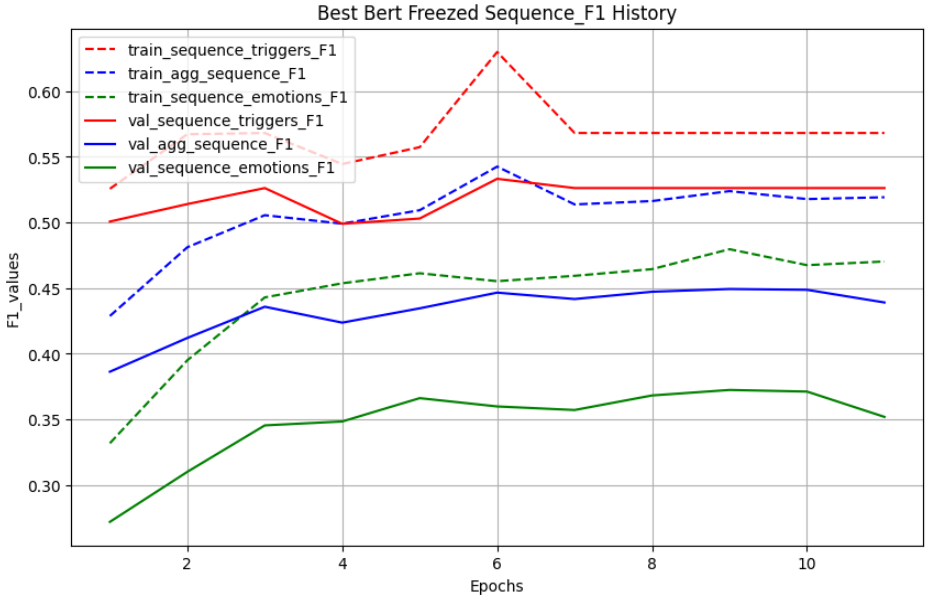
Extract of grid search results of Bert Full on the validation set

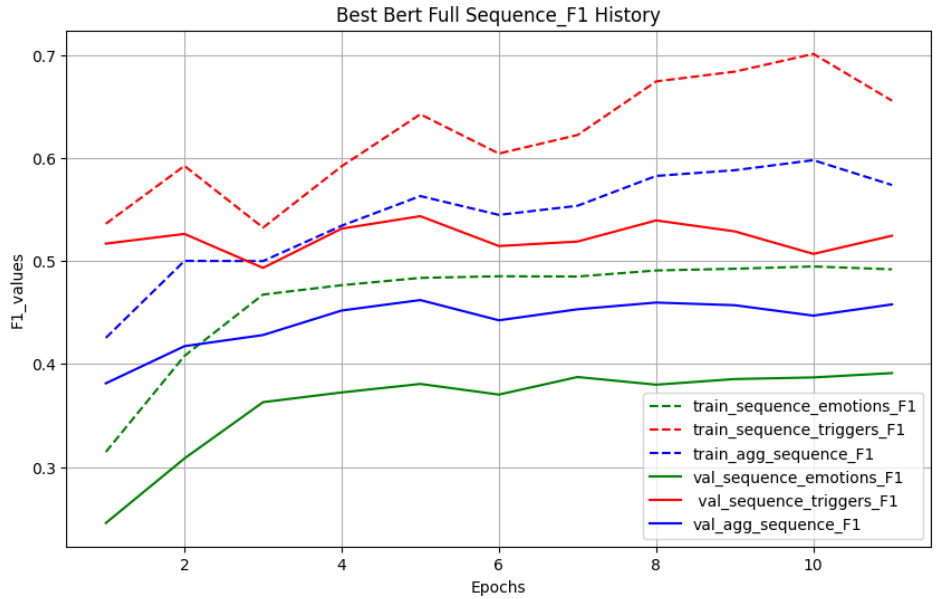
MODEL TRAINING

- As emerged from the executions of the two grid searches The addition of the class weights did not bring significant improvements in the performace of the Bert based models.
- We chose to not consider them and train the models using the Loss funcition that doesn't use it.









RESULTS AND ERROR ANALYSIS

BASELINE MODELS: EMOTION CLASSIFICATION

Emotion classification:

- Poor results especially due to the number of possible values.

- Random classifier does better in unrolled F1.
- Majority classifier does better in sequence F1.

- Maximum scores obtained on emotion classification: **0.128** (unrolled) and **0.174** (sequence)

Model	Class	Unrolled F1	Sequence F1
Random clf	Trigger	0.43227	0.41118
	Emotion	0.12880	0.09196
Majority clf	Trigger	0.45653	0.48501
	Emotion	0.08553	0.17486

Random and Majority classifier Unrolled F1 scores

Emotions	Random clf Scores	Majority clf Scores
Anger	0.142549	0.000000
Disgust	0.063492	0.000000
Fear	0.068536	0.000000
Joy	0.156977	0.000000
Neutral	0.198405	0.598759
Sadness	0.109804	0.000000
Surprise	0.161838	0.000000

Random and Majority classifier unrolled F1 scores on the single emotions

BASELINE MODELS: TRIGGER CLASSIFICATION

Trigger classification:

- Good performance compared to emotion scores.
- Majority classifier does better in every F1 configuration.
- Maximum scores obtained on trigger classification: **0.456** (unrolled) and **0.485** (sequence)

Model	Class	Unrolled F1	Sequence F1
Random clf	Trigger	0.43227	0.41118
	Emotion	0.12880	0.09196
Majority clf	Trigger	0.45653	0.48501
	Emotion	0.08553	0.17486

Random and Majority classifier Unrolled F1 scores

Triggers	Random clf Scores	Majority clf Scores
0	0.621842	0.91307
1	0.242710	0.000000

Random and Majority classifier Unrolled F1 scores on the single triggers

BERT-BASED MODELS RESULTS

- Both Bert Freezed and Bert Full **outperform the baselines** in almost every metric.
- On the validation set the Bert-based models are very similar in terms of overall performance
- Results on the validation set (in **green the best ones** for each metric):

Model	Class	Unrolled Sequence F1	Standard dev.	Sequence F1	Standard dev.
Bert Freezed	Trigger	0.57781	0.01336	0.52643	0.01603
	Emotion	0.76021	0.01259	0.36782	0.00962
Bert Full	Trigger	0.57002	0.00521	0.52645	0.01163
	Emotion	0.77570	0.01295	0.38742	0.00527

Freezed and Full Bert F1 scores on the validation set, the standard deviation is computed averaging the results of 5 seeds

BERT-BASED MODELS RESULTS

- Both Bert Freezed and Bert Full **outperform the baselines** in almost every metric.
- Only sequence F1 on triggers (computed with majority clf) is comparable.
- The Bert-based models are very similar in terms of overall performance. **Full Bert tends to be slightly better**.

Model	Class	Unrolled Sequence F1	Sequence F1
Bert Freezed	Trigger	0.44621	0.43106
	Emotion	0.63317	0.56274
Bert Full	Trigger	0.51193	0.48246
	Emotion	0.64009	0.56666

Freezed and Full Bert F1 scores on the Test set

Class	Unrolled F1	Sequence F1
Trigger	0.456	0.485
Emotion	0.128	0.174

Best results obtained by the two baselines (test set)

Class	Unrolled F1	Sequence F1
Trigger	0.511	0.482
Emotion	0.640	0.566

Best results obtained by Full Bert (test set)

Class	Unrolled F1	Sequence F1
Trigger	0.570	0.485
Emotion	0.775	0.174

Best results obtained by Full Bert (validation set)

EMOTION SCORES

Observations:

- **Simplest emotions to classify**:
  - "**neutral**" due to its high occurrence,
  - "**surprise**" usually associated with exuberant punctuation (e.g., "!!!").

- Hardest emotions to classify (few occurrences and probably more related to body expressiveness):
  - "**disgust**"
  - "**sadness**"

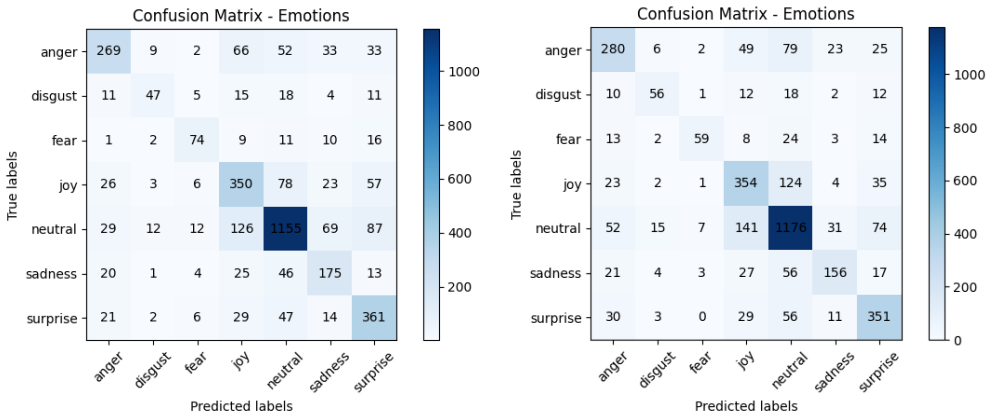
- Occasional difficulties in distinguishing between emotions that might also be confused by a human reader if there is a lack of context. For example:

Utterance	Real Emotion	Full Bert prediction
The ring is gone!	Surprise	Sadness
And you're not supposed to be gossiping!!	Disgust	Anger
I'd love it too. Shoot, I gotta go. So, I'll t...	Neutral	Joy

Prediciton examples (Bert Full) from utterance «1016», «1028» and «1678», respectively

Emotions	Bert Freezed Scores	Bert Full Scores
Anger	0.639715	0.627100
Disgust	0.502674	0.562814
Fear	0.637931	0.602041
Joy	0.601892	0.608770
Neutral	0.795729	0.776494
Sadness	0.571895	0.607004
Surprise	0.682420	0.696429

Bert Freezed and Full unrolled F1 scores on the single emotions



Freezed (left) and Full Bert (right) confusion matrices for emotion classification

TRIGGER SCORES

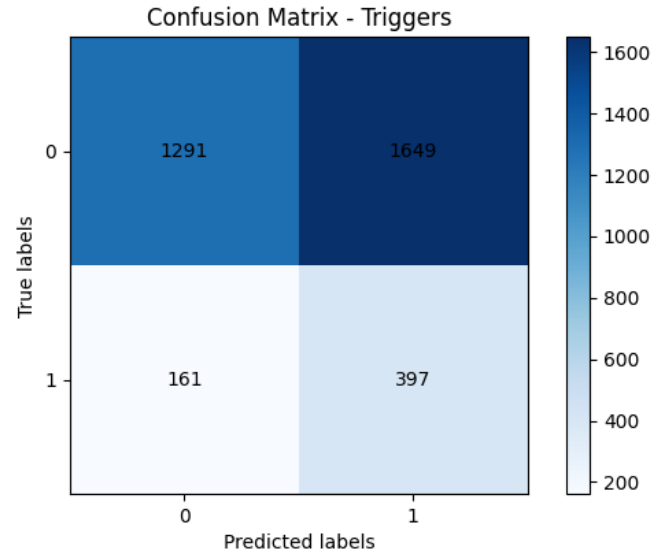
Observations:

- Bert Freezed < Majority Classifier < Bert Full
- Mistakes in detecting a positive trigger may be due to the chunking step of the model pipeline, that may lead to a loss of context.

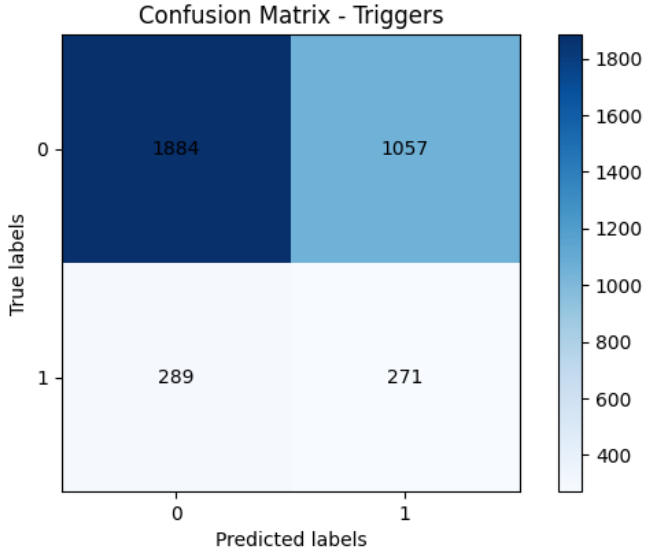
Triggers	Bert Freezed Scores	Bert Full Scores
0	0.587753	0.736801
1	0.303682	0.287076
	0.445762	**0.511938**

Freezed and Full Bert F1 scores on the single triggers (with average)

Bert Freezed

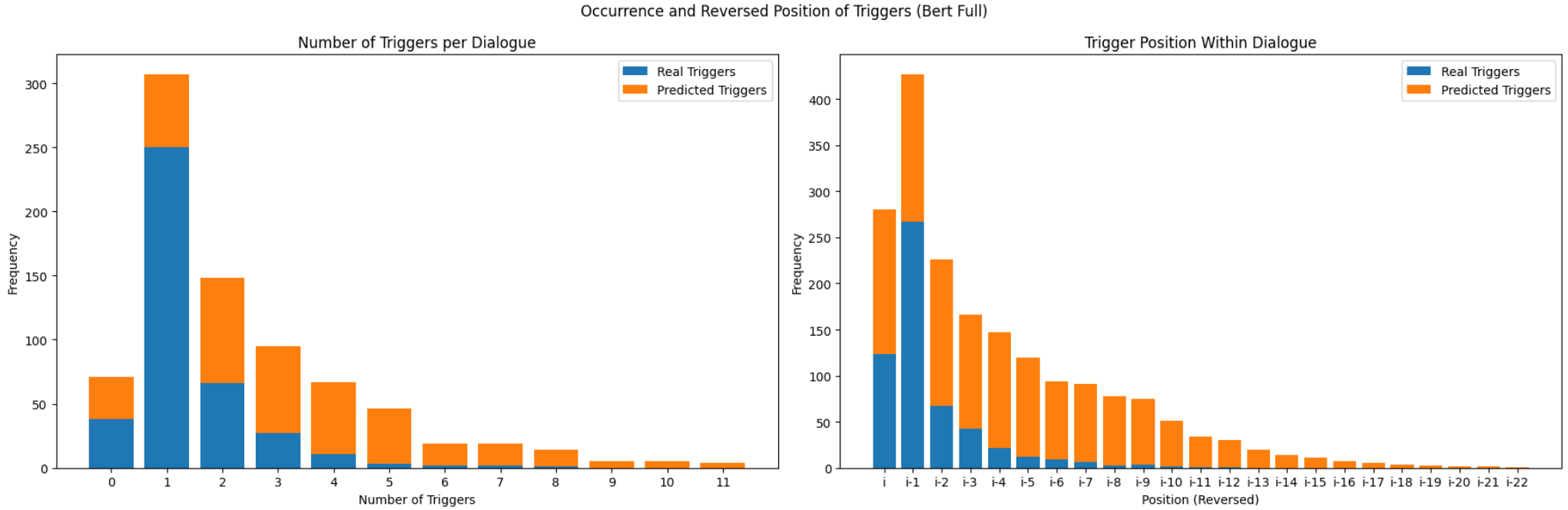


Bert Full



LIMITS AND FUTURE IMPROVEMENTS

TRIGGER CLASSIFICATION LIMITATIONS



Trigger classification limitations:

- Implicit triggers
- Chunking Problem
- Classification structure
- Number of positive examples

Possible improvements:

- Rule-based or machine learning techniques instead of an LLM-based architecture
- Improve the classification mechanism to keep track of the global dependencies in the input sequence

(Kumar et al, 2024)

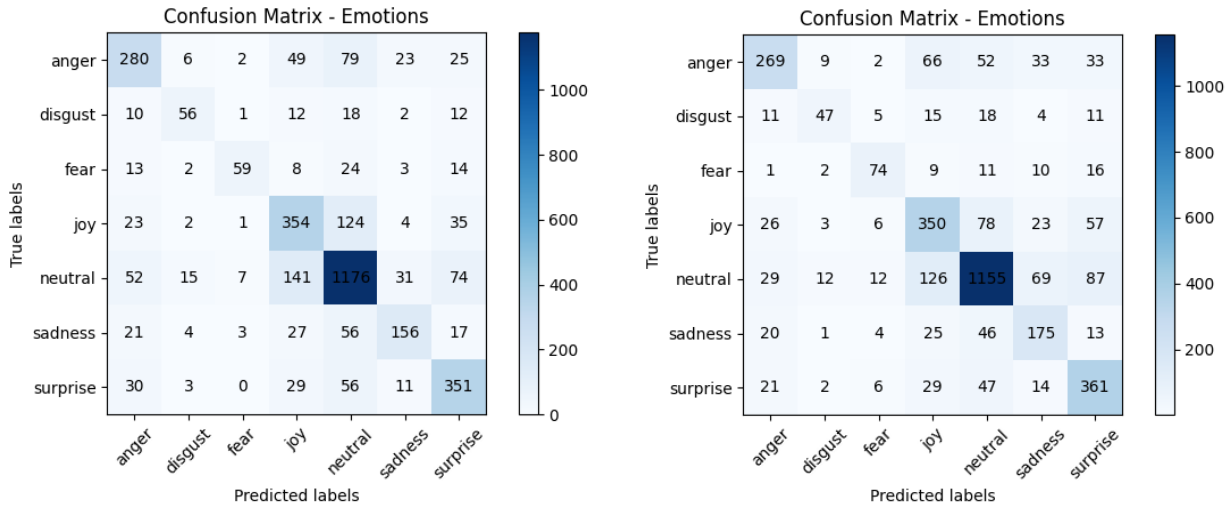
EMOTION CLASSIFICATION LIMITATIONS

Emotion classification limitations:

- Emotion ambiguity
- Chunking problem
- Classification structure

Possible improvements:

- Give the model additional information about the context
- Decrease size of the dataloader
- Increase the input size of the BERT model
- Improve the classification mechanism to keep track of the global dependencies in the input sequence

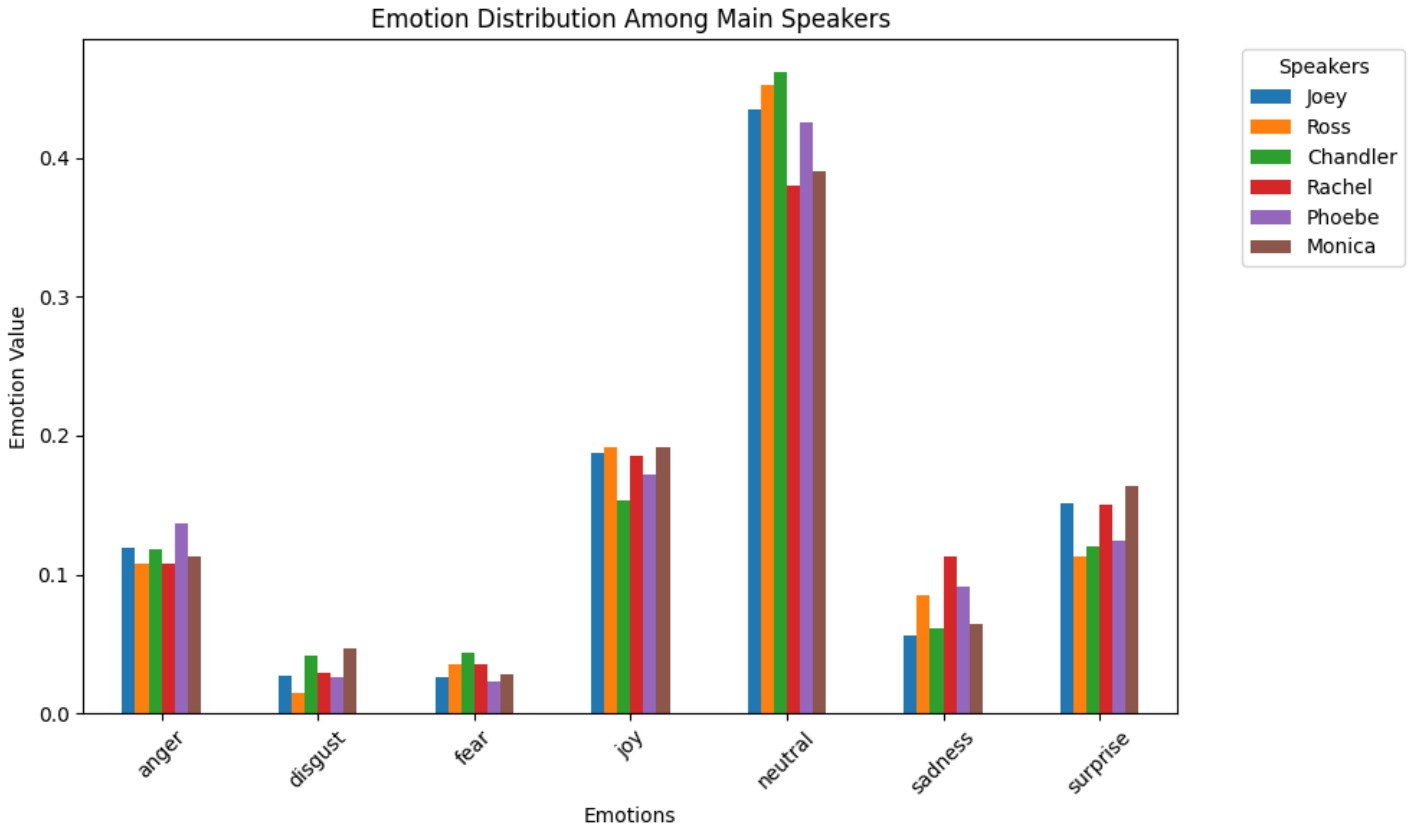


Freezed (left) and Full Bert (right) confusion matrices for emotion classification

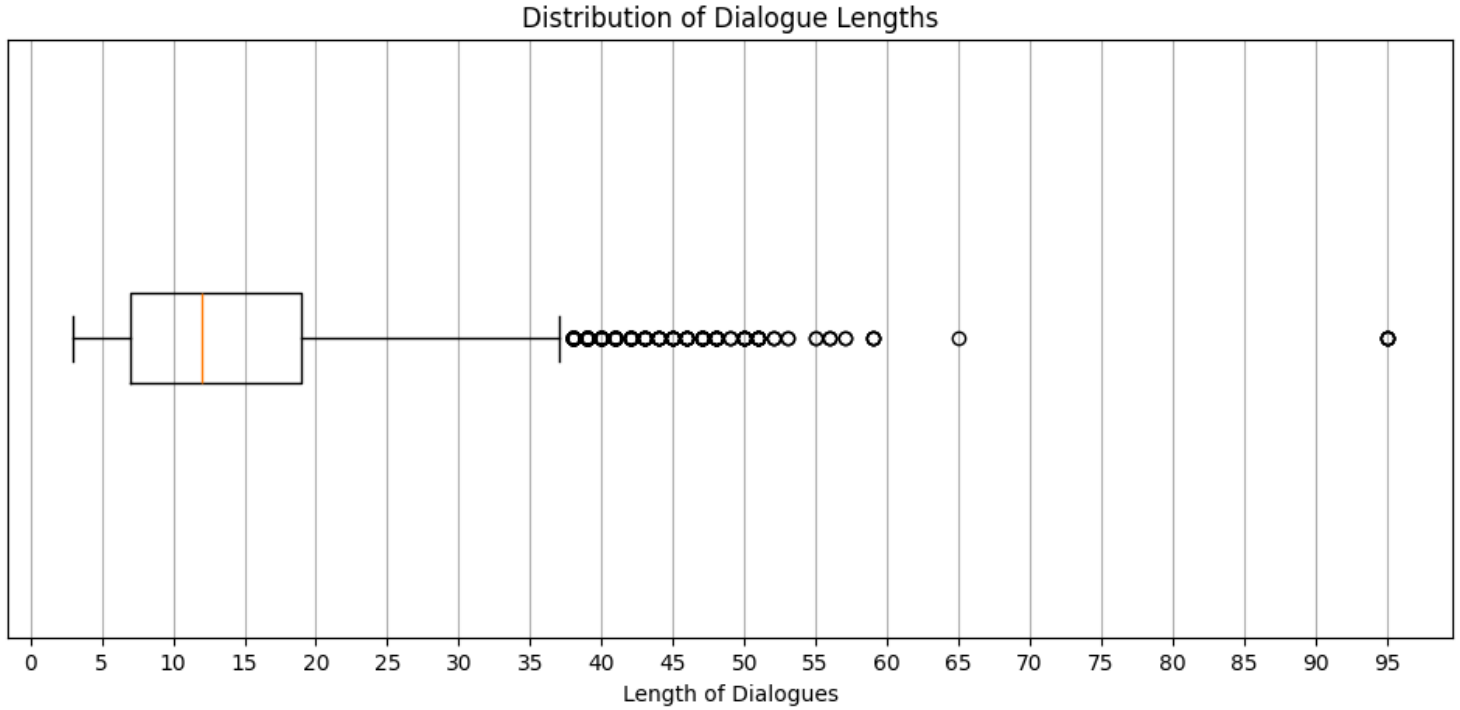
POSSIBLE SOLUTION – SPEAKER INFORMATION

Possible effect:
Encode the information about the speakers into the input sequence in order to give the model supplementary information about the context

Counterpart:
The additional data may lead to the generation of biases, as it introduces the risk of the model making generalizations based on speaker name more than the utterance itself.



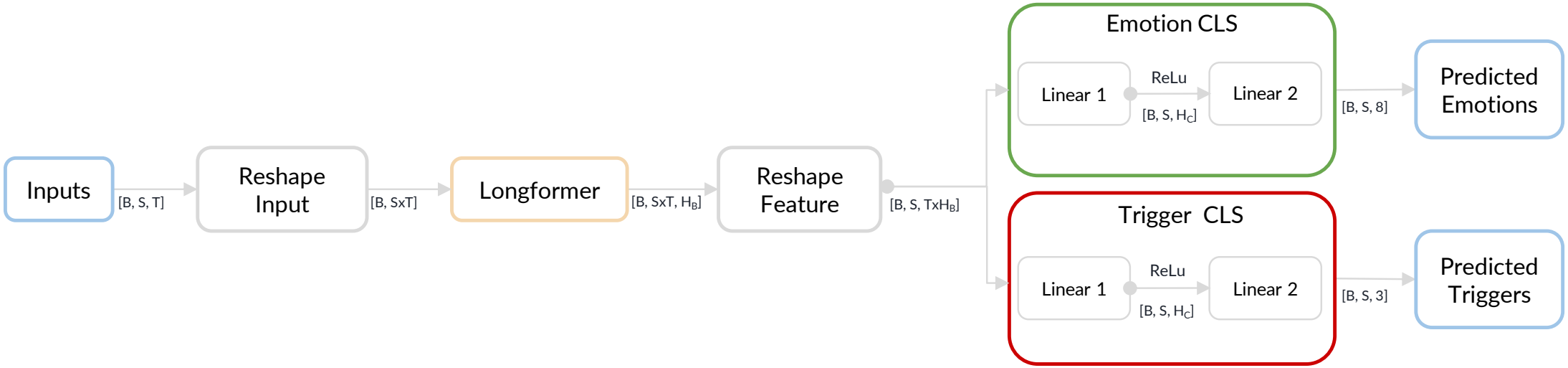
CHUNKING AND CONTEXT PROBLEM



Attenuate the problem: developing a model which takes as input a set of dialogues with a shorter padding.
Having a smaller input sequence decreases the number of chunks, so that the model analyzes a bigger portion of the dialogue.

Among all utterances in the dataset, only 4 have a total length exceeding 95, while every other utterance falls below 65.

CHUNKING AND CONTEXT PROBLEM – LONGFORMER



Partially solve the problem:

- Longformer is a tansformer based on RoBERTa with an input size up to 4096 token.
- The model showed significant improvements in terms of both Unrolled and Sequence F1.
- Without hardware limitation, it would be possible to train Longformer layer, similarly to the Bert Full mode (Beltagy et al. 2020)

	Class	Unrolled F1	Sequence F1
Big BertOne	Trigger	0.54995	0.52998
	Emotion	0.87445	0.81443

Big Bertone performance on the test set

Longformer: The Long-Document Transformer. Beltagy, I., Peters, M. E., & Cohan, A. Allen Institute for Artificial Intelligence, Seattle, WA, USA.

THANK YOU!

CHUNCKING OPERATION

•The number of chunks, denoted as *C*, is calculated to determine how to divide the input.

• *C* is chosen as the smallest divisor of the number of sentences, such that dividing the product of the number of sentences and the number of tokens by C, the result does not exceed the maximum token capacity allowed by BERT.

• We ensure that each chunk contains a manageable number of tokens, complying with the limitations of BERT's input capacity.

$$\mathrm{C} = \min \left\{ c : (\mathrm{n\_sentence} \bmod \mathrm{c}) = 0,\ \frac{\mathrm{n\_sentence} \times \mathrm{n\_token}}{c} \leq \mathrm{bert\_token\_capacity},\ 1 \leq d \leq \mathrm{n\_sentence} - 1 \right\}$$