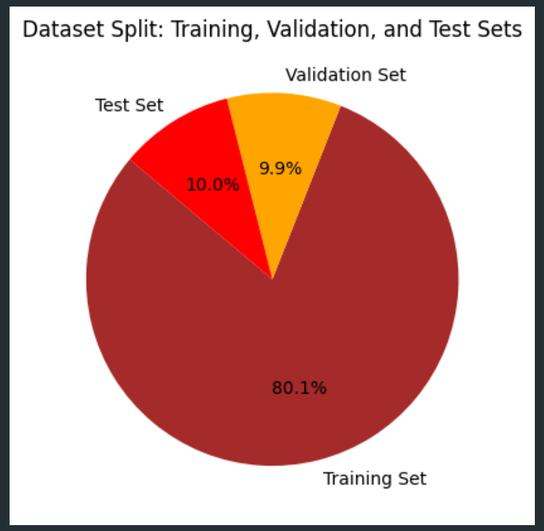
ERC and EFR in English Conversations

G. Castelli, A. Fratini, M. Mone, F. Pigliapoco

NLP COURSE PROJECT - AA 2023-2024

Preprocessing and Data Imbalance

To maintain consistency and coherence, overlapping dialogues were kept together within the same dataset splits during the splitting process.



PADDING:

- **MAX_LENGTH = 256**
- We choose this dimension to handle most dialogues efficiently and reduce extra padding

ONE-HOT ENCODING FOR LABELS:

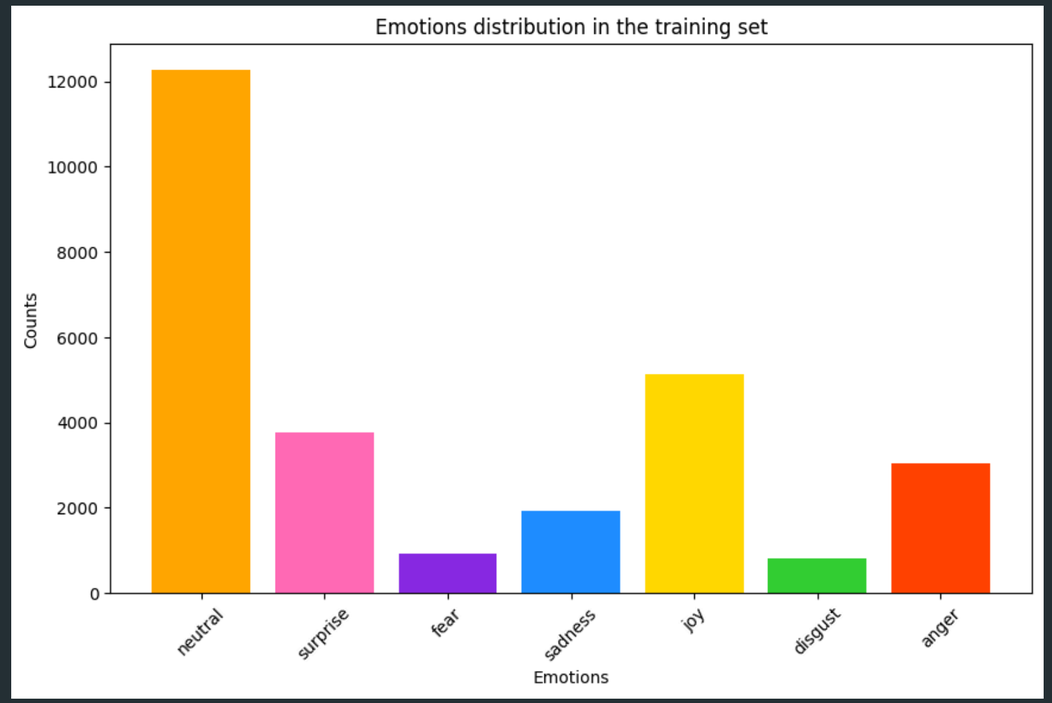
- **Emotions**: Encoded into a 7-dimensional one-hot vector, where each position represents one emotion.
- **Triggers**: Encoded into a binary vector (2-dimensional), where [0,1] represents the presence of a trigger and [1,0] represents no trigger

Preprocessing and Data Imbalance

The dataset is heavily imbalanced with "neutral" emotion being the most frequent

Class Weights: We applied class weighting during training to ensure that minority classes like "disgust" and "fear" received higher weights in the loss function, helping the model learn more effectively from these underrepresented classes.

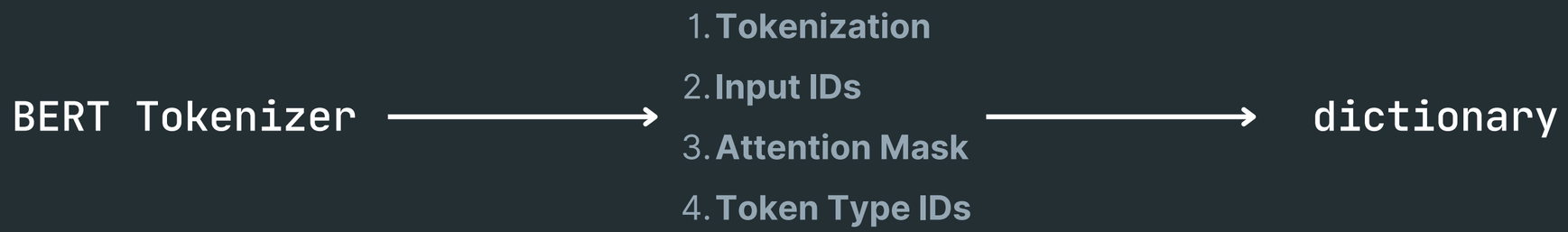
$$\text{weight(j)} = \frac{\text{total samples}}{\text{n of emotions} \times \text{count(j)}}$$



Tokenization

Before tokenization, the dataset is preprocessed by splitting entire dialogues into individual utterances. Each utterance is paired with its corresponding emotion and trigger labels. This transformation creates a new dataset where each row represents one utterance with its associated labels, preparing it for tokenization and model training.

These utterances are then passed to the BERT Tokenizer which use **encode_plus( )** method.



Model Architecture

Emotion Prediction head

- Dropout layer (p=0.3) for regularization.
- Fully connected (Linear) layer: transforms BERT embeddings to a 7-class output for emotion classification.

Base Model: bert-base-uncased

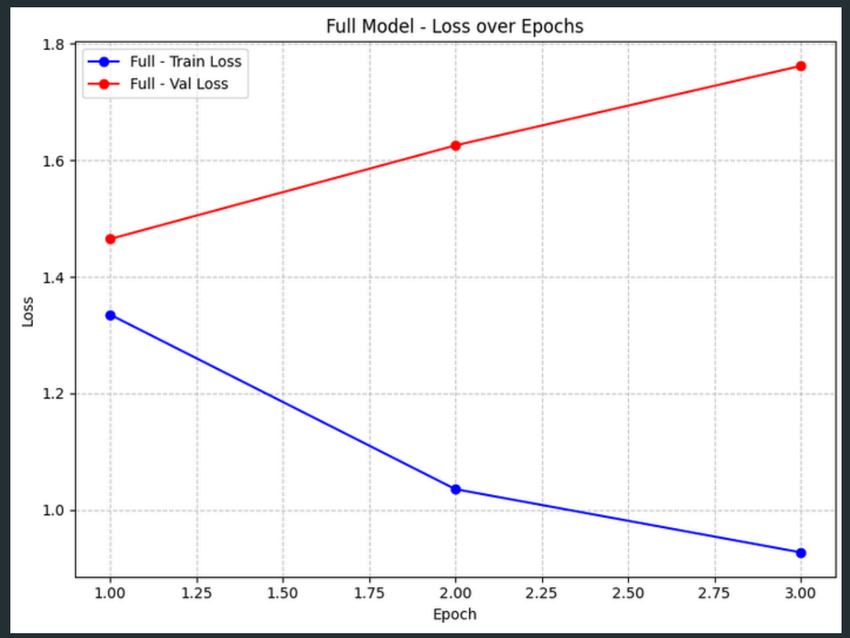
- Pre-trained BERT model.
- Generates contextual embeddings from input tokens
- Transfers knowledge from large-scale textual data.

Trigger Prediction head

- Dropout layer (p=0.3) for regularization.
- Fully connected (Linear) layer: transforms BERT embeddings to a 2-class output for trigger classification.

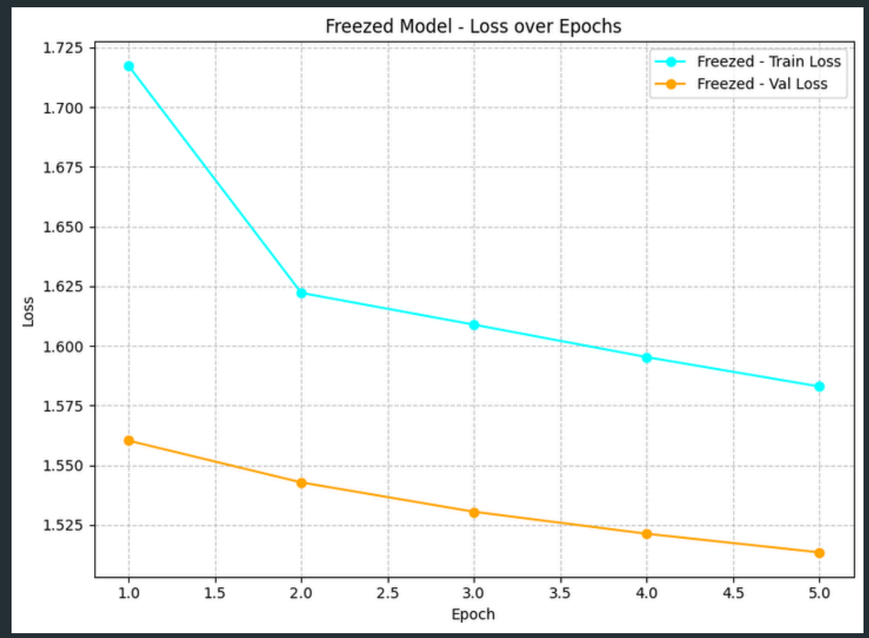
Freezed vs Full

FULL



All weights, including the pre-trained BERT model, are updated during training for greater flexibility.

FREEZED



Pre-trained BERT weights are frozen; only the final classifier layers are fine-tuned, reducing overfitting

Experimental Setup

1

INTRODUCTION

We evaluated two baseline models for comparison:

- **Random** Classifier
- **Majority** Classifier

These models provided **low F1 scores** and poor performance, especially on minority emotions.

2

EVALUATION METRICS

- **F1 Score**: primary metric to address class imbalance (balances precision and recall). Is more effective for evaluating performance on minority classes.
- **Accuracy**: secondary metric, less reliable due, it can mask poor performance on minority emotions.

3

TUNING HYPERPARAMETERS

A GridSearch to optimize:

- Learning rates: 1e-5, 5e-5
- Epochs: 3, 5
- Patience: early stopping set to 2

The optimal configuration achieved better F1 scores, especially for minority classes.

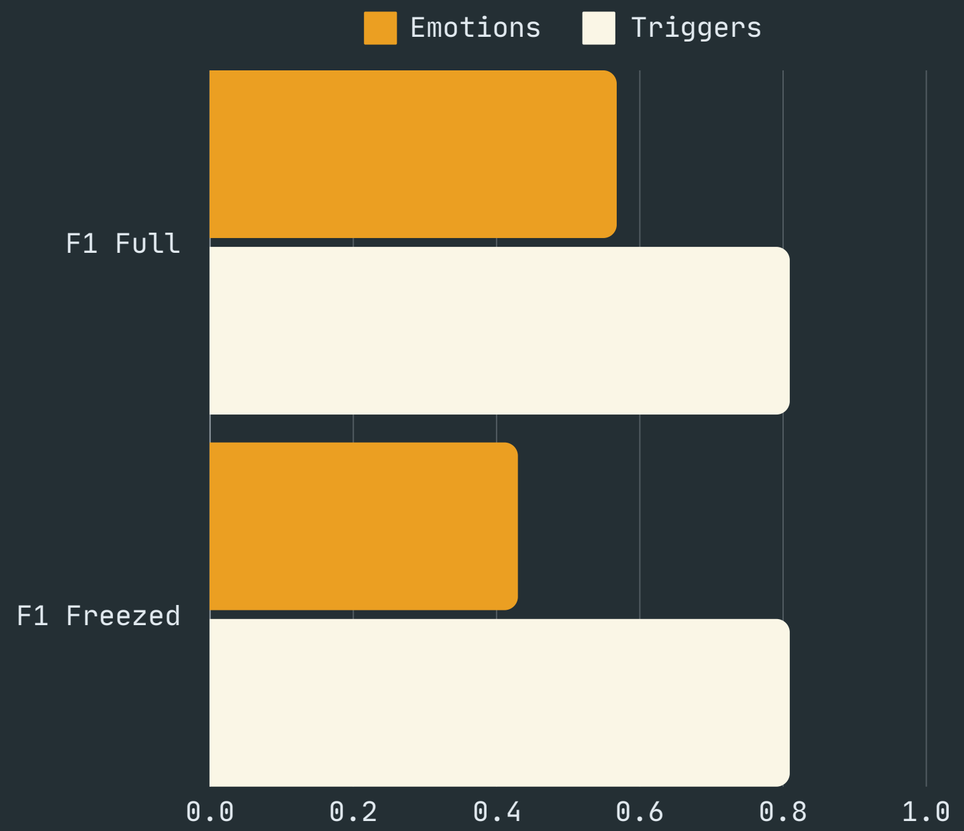
seed = 42, epochs = 5, learning_rate = 5e-5, patience = 2

Results and Error Analysis

Sequence F1 score

Full Model:
Average F1 Score for Emotions: 0.567740
Average F1 Score for Triggers: 0.809320

Freezed Model:
Average F1 Score for Emotions: 0.429787
Average F1 Score for Triggers: 0.809320



Unrolled F1 Score

Full Model:

Trigger	Unrolled F1
0	1.0
1	0.0

The models perform perfectly in identifying the absence of triggers but completely fail to detect their presence. Highlighting the data imbalance

Freezed Model:

Trigger	Unrolled F1
0	1.0
1	0.0

Full Model:

Global Unrolled F1 Score for Emotions: 0.550530

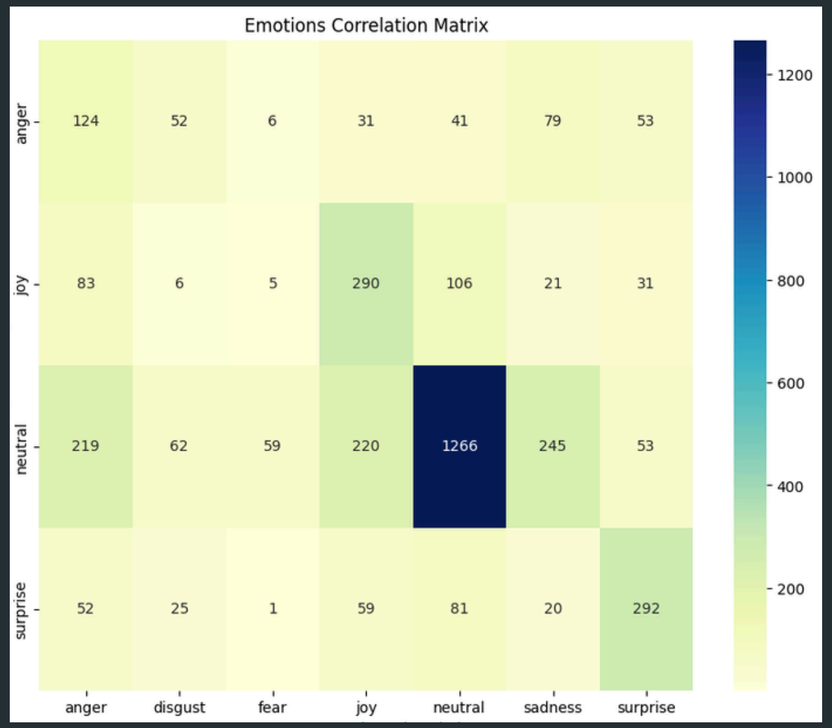
Global Unrolled F1 Score for Triggers: 0.854829

Freezed Model:

Global Unrolled F1 Score for Emotions: 0.431881

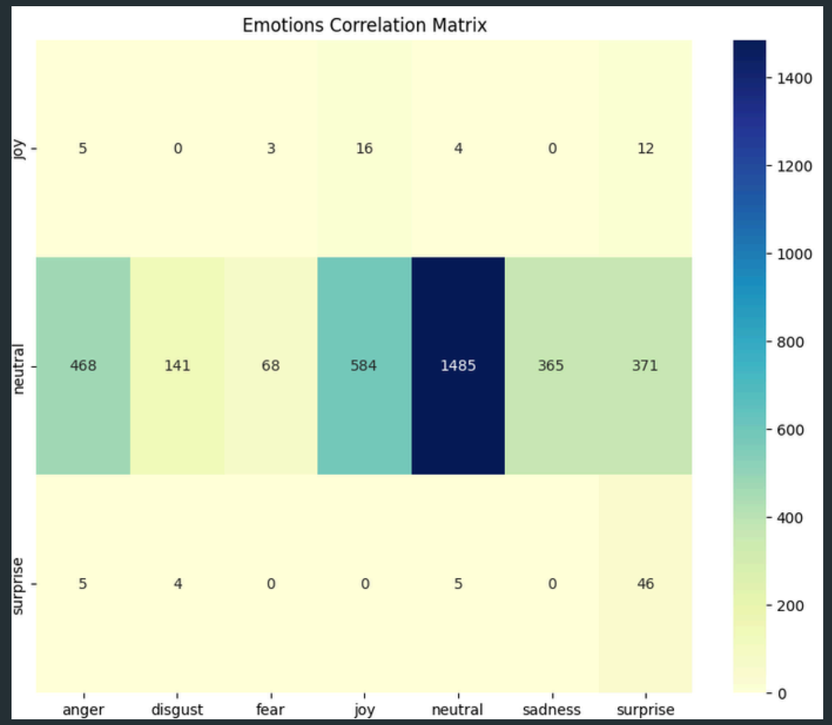
Global Unrolled F1 Score for Triggers: 0.854829

Confusion Matrix



Minority Class Challenge:

Emotions with fewer samples, such as fear and sadness, aren't predicted, highlighting the challenge of handling underrepresented classes in imbalanced datasets.



Miss-classification example

Full Model Emotion Miss-Classification:

emotions	utterances
[disgust, sadness, sadness]	[Oh God, I hate my job, I hate it, I hate my job, I hate it, I hate it...]

Target Emotion: disgust

Predicted Emotion: anger

Freezed Model Emotion Miss-Classification:

emotions	utterances
[neutral, anger]	[Okay, here's batch 22., Oh, maybe these'll take a little longer...]

Target Emotion: anger

Predicted Emotion: neutral

Freezed and Full Model Trigger Miss-Classification:

utterances	triggers
[Last stop, Montreal. This stop is Montreal., ..., ..., ..., ..., ..., ..., ...]	[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0]

Target Trigger: 1

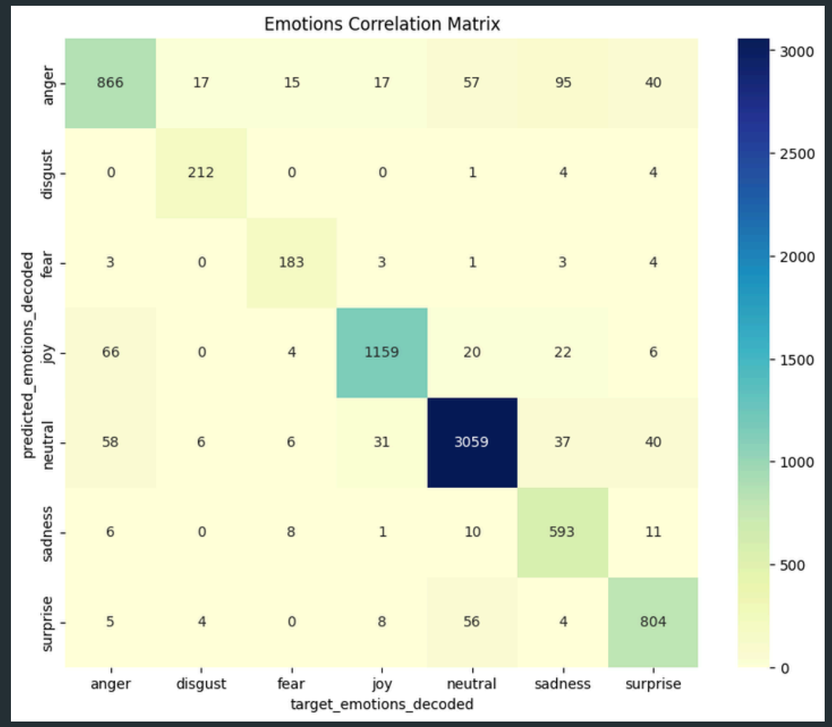
Predicted Trigger: 0

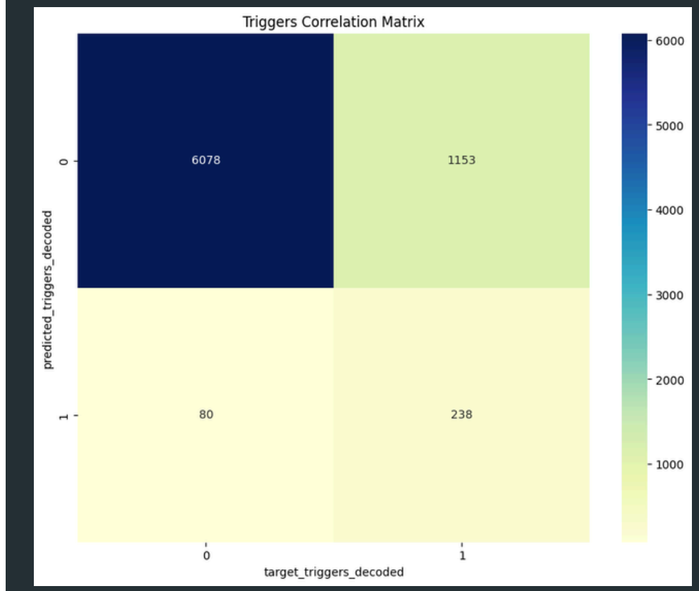
Further Experiments

1 DATA AUGMENTATION

An additional data augmentation step to try to deal better with the imbalanced dataset.

The process creates variations in the text data by shuffling two random words within each sentence of any dialogue having a minority class emotion or trigger, providing additional training examples with slightly different word orders to improve model generalization.





Further Experiments

2 ENTIRE DIALOGUES

Added to the previous data augmentation step it redefines the input data

It redefines the input data by feeding entire dialogues rather than single sentences into the tokenizer and, then, the model. This approach augments the previous experiment, aiming to highlight the role of context in improving emotion classification accuracy.