Multi-Head Neural Network for Emotion Classification and Trigger Detection in Multi-Speaker Dialogues
NLP Course Project & Project Work

Lorenzo Balzani, Alessia Deana and **Thomas Guizzetti**
Master's Degree in Artificial Intelligence, University of Bologna
{ lorenzo.balzani, alessia.deana, thomas.guizzetti}@studio.unibo.it

Abstract

In this project, we tackle the Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) task, which focuses on predicting the emotion of each sentence in a dialogue and identifying any emotional shifts of speakers, or "triggers". Our methodology began with creating a BERT-based multi-head neural network, which we then aimed to refine by incorporating a concatenation method, a transformer, and a large language model. The outcomes demonstrated below average capabilities in emotion prediction and the performance in identifying emotional triggers was less satisfactory. Through our efforts to enhance the model with various improvements, we observed a marginal but significant improvement in the large language model's ability to detect triggers. Still, more fine-tuning is required to reliably recognize emotional changes within dialogues.

1 Introduction

The Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) task aims to identify and predict emotional changes within dialogues, a critical aspect of enhancing conversational AI and understanding human emotional dynamics. This challenge is pivotal for applications ranging from mental health support to customer service, where recognizing emotional shifts can dramatically improve interactions and interventions.

Approaches to EDiReF have evolved from traditional machine learning techniques, such as SVMs and Random Forests, to advanced deep learning models like CNNs, RNNs, and, notably, transformer-based architectures like BERT. These advanced models excel in capturing textual context and emotional nuances but demand significant data and computational resources, posing scalability challenges. Additionally, the subjective nature of emotions complicates data labeling, affecting model training and performance.

We addressed the challenge by developing five distinct models: four multi-head models integrating a BERT-based sequence classification framework with specialized classification heads for both emotion and trigger detection. These models varied in their configurations, including both frozen and unfrozen BERT layers, combinations of emotion and trigger data, and the inclusion of a transformer. The fifth model was a large language model (LLM) fine-tuned through prompt training. Each model was trained on five different seeds on a consistent dataset comprised of dialogues annotated with corresponding emotion and trigger labels, and the F1 score was used to evaluate each model for the emotions and triggers labels. The results are compared to two baselines (majority and random classifiers).

Our findings indicate that while all models performed well in emotion classification, surpassing their efficacy in trigger detection, the standout was the multi-head model featuring an unfrozen BERT base and the most straightforward implementation of classification heads—a feed-forward neural network for emotions and a recurrent neural network for triggers.

2 Background

Research on Emotion Recognition in Conversations reveals avenues beyond mere emotion identification, emphasizing the explicability of emotional dynamics within dialogues. Human emotions frequently shift due to implicit (external) and explicit (internal) factors, the latter being more identifiable. Understanding these emotional transitions has practical applications, including feedback for dialogue agents and affect monitoring. Studies have explored empathetic response generation by adjusting responses based on captured emotions, and others have developed generative models to positively influence user sentiment. Recognizing emotional triggers can inform decisions in various domains. The project addresses the nuanced task of understanding emotional dynamics within conversations, focusing on a range of emotions including 'neutral', 'surprise', 'fear', 'sadness', 'joy', 'disgust', and 'anger', alongside the triggers causing these emotional shifts. Examples illustrate various scenarios, including instances where emotion flips are self-triggered by the speaker or caused by others, with some situations involving multiple triggers or speakers. This concept underscores the complexity of emotional dynamics in dialogues, highlighting the need for nuanced understanding and identification of factors leading to emotional transitions. To achieve this, the project employs a combination of cutting-edge techniques, including both BERT and ROBERTA models, with variations like frozen and unfrozen encoders, and advanced methods like teacher forcing and dynamic weight loss. Additionally, the project innovates by fine-tuning a large language model, GPT-3.5-Turbo, using specific prompts to tackle the task at hand. Through these diverse methodologies, the project seeks to create a nuanced model capable of accurately identifying and predicting emotional shifts and triggers within conversations, thereby paving the way for more empathetic and understanding conversational AI systems.

3 System description

To assess the most effective approach for our task, we implemented several distinct models alongside two baselines, aiming to explore a range of strategies. The baselines consist of simple majority and random classification techniques, while the remaining models, excluding the Large Language Model (LLM) solution, share a core methodology. Specifically, each model processes sentences within dialogues, encoding them through the BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019) language representation frameworks. Despite this common foundation, the models diverge in their utilization of the language models' outputs. They primarily focus on the last hidden state, especially the CLS token, which captures the semantic essence of each sentence. This output is then directed towards distinct classification mechanisms for both emotion recognition and trigger detection tasks. For emotion classification, a straightforward Feed Forward neural network is employed, whereas trigger classification leverages an LSTM. Activation functions employed across these models include Rectified Linear Unit (ReLU), Sigmoid, and Sigmoid Linear Unit (SiLU), with the recurrent neural network elements incorporating a Dropout rate of 20% [Figure 1] .

The models explored are listed below, highlighting their principal variations:

- **BERT unfreezed encoder**:
  - It utilizes a unfreezed pre-trained BERT model (AutoModelForSequenceClassification) from checkpoint "bert-base-uncased" for initial text embedding.
  - It features a custom feed-forward neural network (FFNN) for emotion prediction, comprising a sequence of linear, ReLU, and softmax layers for classifying multiple emotions.
  - It employs a bidirectional LSTM layer, informed by BERT's output, followed by a second FFNN with a linear, ReLU, and sigmoid sequence to predict binary trigger presence.
  - It takes as input the sentences uttered in the dialog, specifying the individual speakers.
- **BERT freezed encoder**:
  - Same as *BERT unfreezed encoder*, but with the BERT layer freezed.
- **RoBERTa** :
  - Same as *BERT unfreezed encoder*, but from checkpoint "roberta-base", this model processes input consisting of sentences spoken in the dialogue, with the names of the individual speakers.
- **RoBERTa teacher forcing**:
  - Same as *RoBERTa*, but the a bidirectional LSTM layer for trigger classification is informed by both BERT's output and the true emotion labels, by concatenating the true emotion labels with with the CLS tokens from BERT [Figure 2].
- **RoBERTa weighed losses**:
  - Same as *RoBERTa*, but using a dynamic weight loss method that weighs multiple loss functions by considering the homoscedastic uncertainty of both the

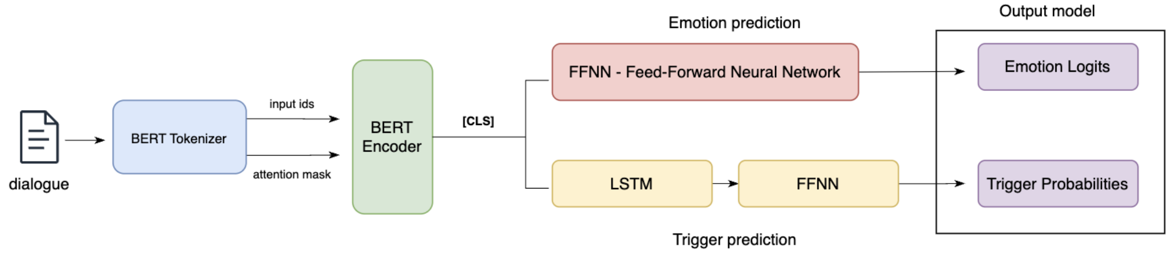


Figure 1: High level overview of our architecture

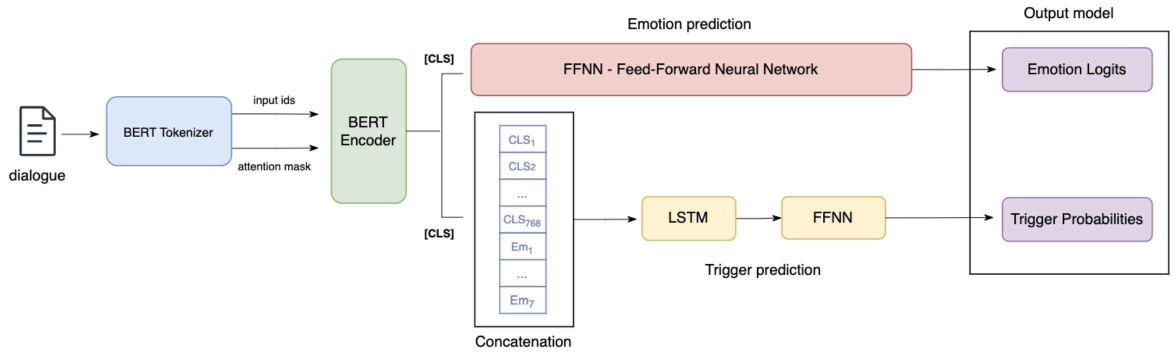


Figure 2: High level overview of our architecture with concatenation

emotion and trigger classification task. (Kendall et al., 2017)

- **RoBERTa Transformer** :
  - Same as *RoBERTa*, but the a bidirectional LSTM layer for trigger classification is replaced by 16 Transformer Encoder layers.
- **Fine-tuned LLM**:
  - It consists of a fine-tuned GPT-3.5-Turbo LLM with prompts designed to solve the task. The inputs consist of sentences spoken in the dialogue concatenated with the names of the individual speakers.

4 Data

The dataset comprises 4,000 short dialogues in English, each tagged with an emotion and a trigger label. The emotion label categorizes the sentiment of the utterance into one of seven categories: ’neutral’, ’surprise’, ’fear’, ’sadness’, ’joy’, ’disgust’, or ’anger’. Meanwhile, the trigger label, defined as a boolean variable, indicates whether an utterance marks an emotional shift within the dialogue. Notably, some trigger labels were initially misformatted as NaN; these have been converted to zero to maintain data integrity.

The dialogue structure encapsulates the interplay between two speakers, capturing the fluidity of human emotion through their exchanges. The analytical challenge lies in not only categorizing the emotional content of each individual utterance but also in detecting the points at which these emotions pivot - these are the moments tagged as triggers. For example, as depicted in the accompanying image (see Figure 3), an exchange might begin with a neutral sentiment, progress through joy, and then shift to sadness, marked by specific trigger points that indicate these emotional transitions.

For the segmentation of the dataset into training, validation, and test sets, the first step entailed assigning an index to each dialogue. Due to the dataset’s structure of incremental dialogues, the process involved determining if each row was part of the dialogue sequence of the immediately preceding row with an added sentence. This method

Trigger	Percentage (%)
False	84.07
True	15.93

Table 1: Percentage Distribution of Triggers

Emotion	Percentage (%)
Neutral	43.61
Joy	18.05
Surprise	13.27
Anger	11.33
Sadness	7.57
Fear	3.18
Disgust	3.00

Table 2: Percentage Distribution of Emotions

identified a total of 833 unique dialogues. For the distribution of these dialogues into different sets, 80% of the 833 dialogues were designated for the training set and 10% for each validation and test set.

The distribution of emotion and trigger labels exhibits a degree of class imbalance, a factor we will account for in our analysis. The specifics of this distribution are detailed in Table 1 and Table 2.

5 Experimental setup and results

We set out to perform the Emotion and Triggers classification with our seven models, by training, evaluating and testing on the dataset provided on 5 different seeds. We evaluate our models using both:

Sequence F1: The average F1-score calculated for each dialogue.

Unrolled Sequence F1: The F1-score computed across a concatenated set of all utterances.

The architectural and hyperparameter configurations of our models varied, tailored to optimize performance across different structures. Following preliminary experimentation, we standardized our training regimen to include 10 epochs, a learning rate of 2×10^{-5}, an LSTM with 16 layers and a hidden size of 256, and a feedforward neural network with a hidden size of 128. We implemented early stopping with a patience of 1 and a minimum delta of 1×10^{-2} to mitigate overfitting.

Our loss function strategy differentiated between emotion and trigger classifications, utilizing Cross Entropy and Binary Cross Entropy, respectively. The composite loss function was formulated as:

$$\text{Total Loss} = \begin{bmatrix} 0.8 & 0.2 \end{bmatrix} \times \begin{bmatrix} emotion\_loss \\ trigger\_loss \end{bmatrix} \quad (1)$$

This allocation was manually determined to prioritize emotion classification accuracy due to its comparatively higher loss magnitude.

Specifically for the *RoBERTa weighed loss* model, we adopted a dynamic weighting approach to balance emotion and trigger losses, leveraging the homoscedastic uncertainty of both tasks as described by (Kendall et al., 2017) . This was facilitated through the `MultiTaskLoss` class in PyTorch, which modulates each task's loss contribution based on its variance. Such dynamic adjustment promotes balanced learning by scaling losses according to learnable log variance parameters, with support for both sum and mean reductions in computing the aggregate multi-task loss:

$$\text{MultiTask Loss} = \sum \left(\frac{1}{2\sigma^2} \times \text{Loss} + \log \sigma \right) \quad (2)$$

where σ represents the standard deviation associated with each task's loss, indicating the degree of uncertainty or variance.

Incorporating the *RoBERTa teacher forcing* approach, we leveraged the final CLS token output from BERT alongside the actual emotion labels, after one-hot encoding, to enhance the bidirectional LSTM layer's performance in trigger classification. This method presupposes that accurate emotion identification can significantly improve trigger detection, given their inherent correlation.

For leveraging LLMs, GPT-3.5-Turbo-0613 via the OpenAI API was our choice, employing both one-shot and few-shot techniques, complemented by dataset fine-tuning, to refine our outcomes. An illustrative prompt format we used was: "Classify the following sentences into emotions: [neutral, surprise, fear, sadness, joy, disgust, anger], and identify whether the sentences are triggers for an emotion-flip of a speaker (generally caused by another speaker) by classifying into these two options: [TRIGGER, NOT A TRIGGER]." for our Fine-tuned LLM with speakers. We also defined our GPT System to be "You are an emotion classifier, that can classify the emotion of sentences in a dialog and the sentences that trigger an emotion-flip of speakers.". This setup served to guide the model in accurately categorizing sentences based on emotional content and trigger identification. The model

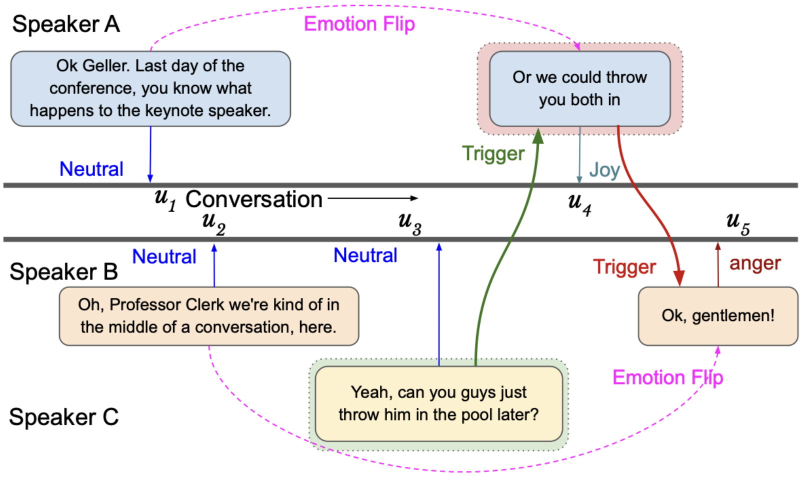

Speaker A
Emotion Flip
Ok Geller. Last day of the conference, you know what happens to the keynote speaker.
Or we could throw you both in
Trigger
Neutral
Joy
u_1 Conversation
u_2
u_3
u_4
u_5
Speaker B
Neutral
Neutral
Trigger
anger
Oh, Professor Clerk we're kind of in the middle of a conversation, here.
Ok, gentlemen!
Emotion Flip
Speaker C
Yeah, can you guys just throw him in the pool later?

Figure 3: Example of dialogue with Emotion-flip(Kumar et al., 2021)

was fine-tuned over 2 epochs over the training set using the OpenAI fine-tuning funtionality. At inference, the temperature of the text generation (which regulates how creative the model is) was set to 0.

The AdamW optimizer was the choice for all models, aligning with best practices for managing weight decay and stabilizing training dynamics.

The initial results, encompassing both the frozen and unfrozen BERT models alongside the baseline comparisons, are presented in Table 3. Results from further experiments involving the remainder of the models are detailed in Table 4.

6 Discussion

The overall results from all models indicate below average performance in emotion recognition, with the unfrozen BERT model, leveraging either the "bert-base-uncased" or "roberta-base" checkpoints from the Hugging Face library. Specifically, it reached a Sequence F1 Score of around 40% and an Unrolled Sequence F1 Score of 39%. The fine-tuned LLMs instead excelled in trigger classification, recording a Sequence F1 Score of approximately 78% and an Unrolled Sequence F1 Score of 77%. This performance aligns with our expectations, positing emotion recognition and trigger classification as complex optimization challenge. The achieved scores surpass those of the Majority and Random classifiers on emotion classification, that indicate that patterns are being learned. However for triggers, the models perform just as well as the baselines.

As anticipated, there's a substantial discrepancy in emotion classification outcomes between BERT models with frozen layers during training and those with unfrozen layers, with the latter significantly outperforming the former. However, for trigger classification, the models performed similarly, suggesting that fully fine-tuning the network may not offer a significant advantage over employing the pre-trained embeddings with a custom classifier layer for this task.

It's notable that none of our additional exploratory models significantly surpassed the performance of the initial models, an unexpected finding that suggests we might have reached the peak achievable score with our current multi-task approach, given our setup and hardware constraints.

Interestingly, BERT and RoBERTa yielded similar results. Despite RoBERTa's training on a vastly larger corpus and its exclusion of the next sentence prediction (NSP) task—focusing instead solely on the masked language model (MLM) task with extensive hyperparameter tuning and longer training durations—its advanced training procedure and larger dataset typically ensure its superiority over BERT in various NLP tasks. However, this advantage did not significantly impact this particular task.

It should also be emphasized that during our experimentation, we noticed that incorporating speaker information into our dataset without altering the model's architecture or parameters, marginally improves the emotion recognition task but does not enhance the overall trigger classification results. This indicates that the manner in which our model integrates speaker information, which is crucial for determining triggers, does not

Model	Metric	Seq. F1	Unroll. Seq. F1
Majority Classifier	Emotions	18.20	8.69
	Triggers	48.17	45.68
Random Classifier	Emotions	8.95 ± 0.34	11.22 ± 0.34
	Triggers	42.03 ± 0.60	43.82 ± 0.43
BERT freezed	Emotions	22.11 ± 3.90	15.31 ± 2.67
	Triggers	47.42 ± 0.00	45.57 ± 0.00
BERT unfreezed	Emotions	40.48 ± 0.93	39.46 ± 1.37
	Triggers	46.35 ± 2.13	47.48 ± 3.80

Table 3: F1-Score Comparison of Initial Models

Model	Metric	Seq. F1	Unroll. Seq. F1
RoBERTa	Emotions	39.83	38.00
	Triggers	47.42	45.57
RoBERTa $_{\text{teacher forcing}}$	Emotions	37.83	38.71
	Triggers	39.97	54.25
RoBERTa $_{\text{weighed losses}}$	Emotions	42.76	37.34
	Triggers	47.42	45.57
RoBERTa $_{\text{transform}}$	Emotions	36.44	34.70
	Triggers	47.42	45.57
Fine-tuned LLM	Emotions	52.20	51.44
	Triggers	78.53	77.92

Table 4: F1-Score Comparison of Further Exploration Models

add significant value for trigger classification.

Furthermore, integrating emotion labels with BERT's outputs for the Trigger classification in the *RoBERTa teacher forcing* model did not enhance performance. This might be due to the dimensionality of the emotions, based on dialogue length, being overshadowed by the dimensionality of BERT's outputs (712).

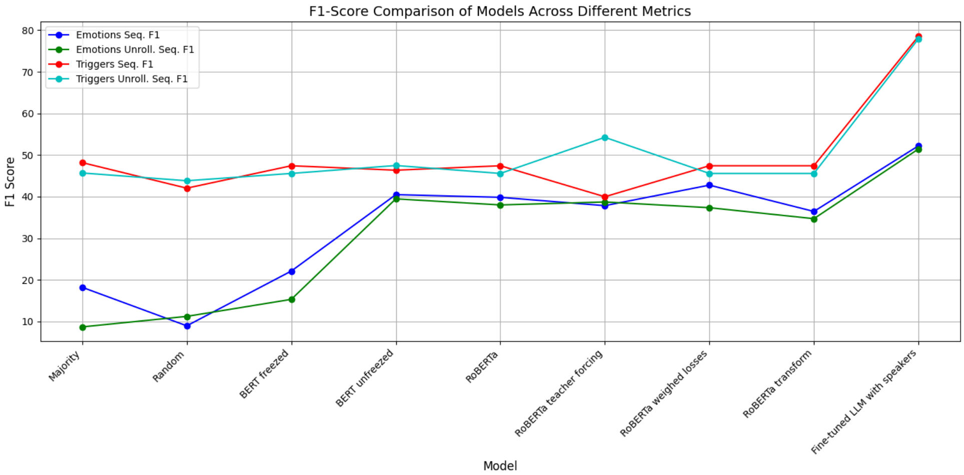
Employing a dynamic multi-task loss in *RoBERTa weighted losses* also failed to produce better outcomes, and in some cases, resulted in slightly worse results. This indicates that the original manually configured multi-task loss was already well-suited for our training and optimization needs.

The introduction of a transformer model in the *RoBERTa transformer* configuration led to poorer outcomes, suggesting that adding this level of complexity does not contribute positively to the model.

Regarding the fine-tuned Large Language Model (LLM), we can see that the performance of the model on Emotion recognition outperforms the standard set by the other models. Moreover, the performance in Trigger classification surpasses that of other models, with *Fine-tuned LLM with speakers* averaging an F1 score of 77%, indicating that a Large Language Model has achieved a deeper understanding of sentence meanings, thereby effectively identifying trigger sentences. This is undeniably a positive outcome that merits further investigation in future research. The model also exhibited some hallucinations (e.g. assigning an emotion that was not part of our initial label set to a sentence) and the inability to respond to some queries, however the number of these issues was small compared to the test data. When hallucination occurred, the model is set up to randomly assign an emotion or a trigger value to a sentence.

Analyzing the results of the emotion classification with our best performing model *BERT unfreezed*, as presented in Table 5 reveals a high degree of model accuracy and precision across various emotions, indicating a robust ability to discern nuanced emotional states within the dataset. The precision and recall scores for 'Neutral' emotion are particularly high, suggesting that the model is highly effective at identifying instances with no significant emotional content.

In contrast, the trigger classification results for *BERT unfreezed*, as presented in Table 6, exhibit



a clear disparity between the performance on non-trigger instances (label 0) and trigger instances (label 1), with precision, recall, and f1-scores for label 1 being notably lower than those for label 0. This imbalance, potentially exacerbated by class imbalance where triggers are less represented, highlights the challenges in distinguishing triggers within dialogues, necessitating further model refinement or the incorporation of more contextual or nuanced features to improve sensitivity towards trigger identification.

On the other side the best best performances were reached were reached by the LLM model. Table 7 presents the precision, recall, F1-Score, and support for various emotions. Joy has the highest F1-Score at 0.75, followed by sadness and anger, whereas surprise and fear have the lowest scores, indicating a challenge in their identification. The weighted average scores are higher, reflecting that classes with more occurrences, like joy, positively influence the overall results.

Table 8, instead, shows metrics for a binary classification task with significantly higher scores for label 0, suggesting that the model may be particularly well-suited for this classification or that the task is more clearly defined.

In summary, GPT-3.5-Turbo performs better in binary classification than in emotion classification, with varying effectiveness across different emotions.

7 Conclusion

Our study into emotion and trigger classification using various advanced models, including BERT, RoBERTa and GPT-3.5-Turbo, reveals significant insights into the computational understanding of nuanced emotional states and trigger events in dialogues. Our exploration, which spanned from leveraging pre-trained embeddings to fine-tuning large language models (LLMs), showcased a below-average performance in emotion recognition tasks. The unfrozen LLM emerged as a frontrunner for the task of emotion recognition and of trigger classification, achieving good Sequence and Unrolled Sequence F1 Scores.

Despite these results, trigger classification encountered predictable challenges, underscored by the models' uniform performance across this task. This observation suggests that while pre-trained models provide a robust foundation for emotion recognition, their application to trigger classification requires a nuanced approach that possibly extends beyond the scope of mere fine-tuning. Interestingly, none of the models, including those based on RoBERTa's extensive corpus and sophisticated training regimen, managed to significantly exceed the benchmarks set by our initial models for emotion recognition, hinting at a potential ceiling in performance dictated by the inherent complexity of the task and perhaps our current methodological approach.

It is important to highlight that the most promising results for trigger classification, though still requiring validation through additional experiments, were achieved with the Large Language Model employing GPT-3.5-Turbo-0613. This intriguing outcome suggests that further research could concentrate on expanding the application of Large Language Models for solving the task, experimenting with various models.

In conclusion, our findings affirm the potential of using deep learning models for understanding emotional nuances and triggers in text, with the

Emotion	Precision	Recall	F1-Score	Support
neutral	0.85	0.58	0.69	1718
surprise	0.49	0.59	0.54	432
fear	0.09	0.26	0.13	74
sadness	0.28	0.39	0.33	301
joy	0.47	0.61	0.53	643
disgust	0.14	0.30	0.19	117
anger	0.46	0.35	0.40	497
Accuracy			0.53	3782
Macro Avg	0.40	0.44	0.40	3782
Weighted Avg	0.61	0.53	0.55	3782

Table 5: Emotion Classification Metrics BERT Unfreezed

Label	Precision	Recall	F1-Score	Support
0	0.84	1.00	0.91	3167
1	0.15	0.20	0.17	615
Accuracy			0.84	3782
Macro Avg	0.50	0.60	0.54	3782
Weighted Avg	0.73	0.87	0.79	3782

Table 6: Trigger Classification Metrics BERT Unfreezed

Emotion	Precision	Recall	F1-Score	Support
neutral	0.54	0.54	0.54	497
surprise	0.51	0.40	0.45	117
fear	0.16	0.11	0.13	74
sadness	0.61	0.71	0.66	643
joy	0.78	0.73	0.75	1718
disgust	0.43	0.47	0.45	301
anger	0.61	0.64	0.62	432
Accuracy			0.65	3782
Macro Avg	0.52	0.51	0.51	3782
Weighted Avg	0.65	0.65	0.65	3782

Table 7: Emotion Classification Metrics GPT-3.5-Turbo

Label	Precision	Recall	F1-Score	Support
0	0.91	0.97	0.94	3167
1	0.75	0.53	0.62	615
Accuracy			0.90	3782
Macro Avg	0.83	0.75	0.78	3782
Weighted Avg	0.89	0.90	0.89	3782

Table 8: Trigger Classification Metrics GPT-3.5-Turbo

unfrozen BERT (and RoBERTa) model performing well in emotion classification and fine-tuned GPT-3.5-Turbo in trigger classification (and emotion classification). However, the lesser performance in trigger classification underscore the need for continued research in this area, focusing on model refinement and the exploration of more sophisticated multi-task learning frameworks.

8 Links to external resources

Link to the GitHub repository: nlp_projects.git

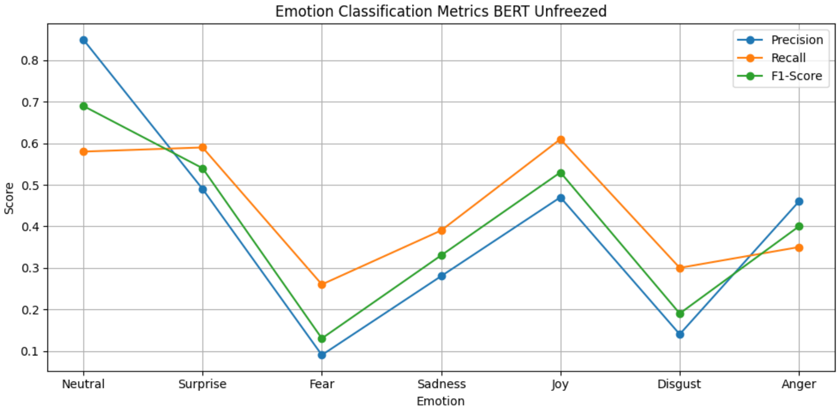
References

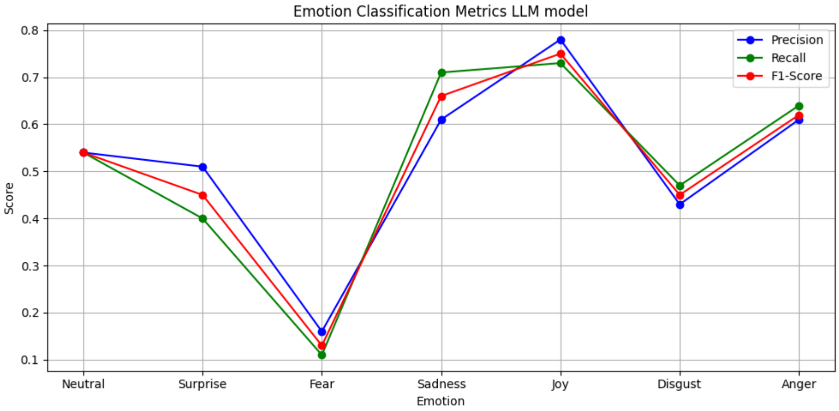
Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2017. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *CoRR*, abs/1705.07115.

Shivani Kumar, Anubhav Shrimal, Md Shad Akhtar, and Tanmoy Chakraborty. 2021. Discovering emotion and reasoning its flip in multi-party conversations using masked memory network and transformer. ArXiv:2103.12360v3.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.


Emotion Classification Metrics BERT Unfreezed
Precision
Recall
F1-Score
Score
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Neutral
Surprise
Fear
Sadness
Joy
Disgust
Anger
Emotion


Emotion Classification Metrics LLM model
Precision
Recall
F1-Score
Score
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Neutral
Surprise
Fear
Sadness
Joy
Disgust
Anger
Emotion