Emotion Discovery and Reasoning its Flip in Conversation (EDiReF)
NLP Course Project

Fatemeh Bozorgi, Reza Shatery and **Kankana Gosh**
Master's Degree in Artificial Intelligence, University of Bologna
{ Fatemeh.Bozorgi, Reza.Shatery, Kankana.Gosh}@studio.unibo.it

Abstract

SemEval 2024 Task 10, "EDiReF," aims to advance human communication understanding by analyzing emotional shifts in English dialogues. By analyzing a dataset comprising 4,000 dialogues, researchers aim to identify the factors that influence emotional shifts. a BERT model is put to the test in two configurations: one in its initial state and another after undergoing extensive training. These BERT models are compared against simpler classifiers like random and majority classifiers. The evaluation process involves using metrics to measure how well these models can recognize emotions within dialogues. The findings underscore the effectiveness of the BERT model, particularly after fine-tuning, in discerning the intricate nuances of emotions within conversational contexts. Through rigorous analysis, it becomes evident that BERT outperforms the simpler classifiers, achieving higher accuracy and demonstrating superior performance metrics. This underscores the efficacy of BERT in complex emotion recognition tasks within conversational contexts.

1 Introduction

In the fascinating world of artificial intelligence (AI), one of the big goals is to make machines behave more like humans. A key part of this is getting them to understand human emotions (Ekman, 1992). Emotion Recognition in Conversation (ERC) has become a lively area of research in the field of Natural Language Processing (NLP). It's all about teaching computers to recognize emotions during conversations, which becomes especially important when emotions suddenly change.

But just recognizing when emotions change isn't enough. We also need to know why they change, so we can make better decisions. For example, in customer service, if a customer goes from feeling happy to feeling upset because of something said by a computer system, it's crucial to understand what was said that caused this change. This knowledge can help improve future interactions.

Emotion-Flip Reasoning (EFR), described by (Kumar et al., 2021). (2021), is a new approach that tries to figure out what exactly causes emotions to change during conversations involving multiple people. This project aims to test how well computer systems can handle both recognizing emotions and understanding what causes them to change.

A variety of approaches have been explored to tackle the problem of emotion recognition in dialogue. Traditional methods often rely on keyword spotting and rule-based systems, which, while straightforward, suffer from a lack of context sensitivity and scalability. Machine learning models offer more nuanced predictions but require extensive labeled datasets and can struggle with the subtleties of human emotion(Kumar et al., 2024).

In this project, alongside the BERT (Bidirectional Encoder Representations from Transformers) model, we also employ random and majority classifiers as baselines for comparison. These simpler classifiers serve to highlight the advanced capabilities of BERT when it comes to understanding the contextual nuances of emotional dialogues. The comparative analysis aims to demonstrate the superior performance of BERT over these baseline methods.

By fine-tuning BERT on a curated dataset of English dialogues, the project aims to not only identify emotions but also discern the underlying triggers for emotional shifts. This approach is inspired by recent advancements in transformer-based models and their success in capturing complex language patterns.

The datasets for this task comprise manually annotated conversations focusing on emotions and triggers for emotion shifts.

Two key metrics stand out in our evaluation process: Sequence F1 and Unrolled Sequence F1. The

Sequence F1 metric calculates the F1-score for each dialogue individually and then computes the average score across all dialogues. This gives us insights into how well our models perform on a dialogue-by-dialogue basis. On the other hand, the Unrolled Sequence F1 takes a different approach. It involves flattening all utterances across dialogues and then computing the F1-score. This provides a more comprehensive overview of our model's performance across the entire dataset, irrespective of individual dialogues. Both of these metrics are computed for emotions and trigger labels, allowing us to assess our models' proficiency not only in recognizing emotions but also in identifying the specific triggers that lead to emotional shifts. By reporting these metrics, we aim to provide a thorough evaluation of our models' capabilities in navigating the complex landscape of emotional understanding in conversations.

2 Background

Recent advancements in deep learning, particularly the development of transformer-based models like BERT, have set new benchmarks in various NLP tasks. BERT's architecture, which allows it to capture bidirectional context, makes it an ideal candidate for tasks requiring a deep understanding of language nuances. However, the application of BERT to emotion recognition in dialogue is relatively unexplored, presenting an opportunity for significant contributions to the field.

This project's focus on emotion triggers in dialogue is rooted in the hypothesis that understanding the cause of an emotional shift is as crucial as recognizing the emotion itself. This perspective is informed by psychological studies that emphasize the role of cognitive appraisal in emotional responses. By integrating BERT with a dataset specifically annotated for emotion triggers, we aim to pioneer a model that not only identifies emotions but also elucidates the reasons behind them.

In the vast landscape of artificial intelligence (AI), understanding human emotions holds a pivotal role in bridging the gap between machines and humans. Emotion Recognition in Conversation (ERC) has emerged as a significant area of research within Natural Language Processing (NLP), aiming to imbue AI systems with the ability to comprehend and respond to human emotions effectively. Our project delves into this domain, focusing on the intricate dynamics of emotional understanding in conversational contexts.

At the heart of our endeavor lies the exploration of BERT, a transformative model in NLP renowned for its prowess in capturing intricate linguistic nuances. Despite its remarkable capabilities, BERT's application in deciphering emotions within conversational exchanges remains relatively underexplored. Motivated by this gap, our project seeks to harness the power of BERT to unravel the complexities of emotional shifts in dialogue.

Drawing inspiration from the concept of Emotion-Flip Reasoning, we embark on a journey to not only recognize emotions but also discern the underlying triggers that precipitate these emotional transitions. This nuanced approach stems from the recognition that understanding the causative factors behind emotional shifts is paramount for AI systems to engage meaningfully in human-like interactions.

Our project is guided by a twofold objective: first, to develop a model proficient in recognizing emotions within conversations, and second, to elucidate the causal mechanisms driving emotional transitions. By integrating advanced deep learning techniques and leveraging meticulously annotated datasets, we aim to pave the way for AI systems capable of not only perceiving emotions but also comprehending the contextual cues that shape human emotional experiences.

Through rigorous experimentation and analysis, we strive to advance the frontier of emotion understanding in AI, with the ultimate goal of fostering empathetic and engaging interactions between humans and machines.

3 System description

Our system comprises several key components:

3.1 Data Preprocessing

- Data Loading: We load the data from the JSON file using the Pandas library.
- Handling Missing Values: We replace any NaN values in the 'triggers' column with 0.0.
- In our data preprocessing pipeline, we transformed the emotion labels (originally represented as strings) into numerical values. Each unique emotion was assigned a unique numeric ID. This conversion allows us to work with emotions in a consistent and computa-

tionally efficient manner during subsequent analysis.

- Train/Validation/Test Split: We split the data into train, validation and test sets based on specified proportions(80/10/10).
- In our model training process, we addressed class imbalance by computing class weights for both emotions and triggers. These weights ensure that underrepresented classes receive the appropriate attention during training.

3.2 Architecture

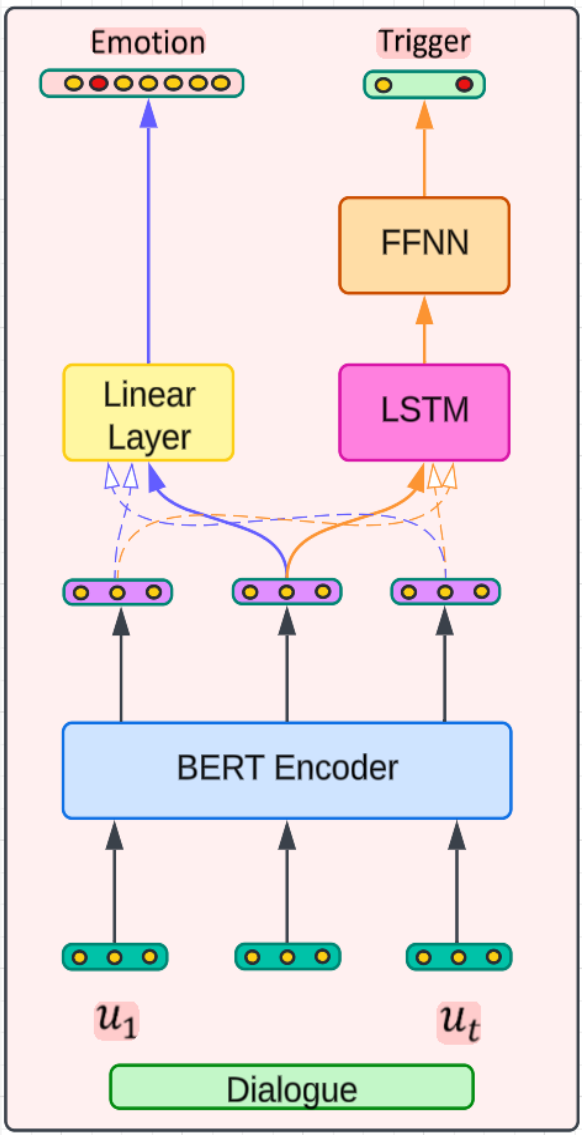

Figure 1: An overview of our proposed architecture. The input dialogue consists of utterances u_1 to u_t, resulting in input size $[t, max\_no\_of\_tokens]$, where t is the number of utterances in each dialogue.

To achieve efficient multitask learning for emotion and trigger classification, we designed a single deep learning model with shared feature extraction and task-specific prediction heads. This architecture leverages the power of a pre-trained BERT model to extract rich contextual representations from the input dialogue utterances. The extracted features are then fed into separate classification heads: one for predicting emotions (multi-label) and another for trigger detection (binary classification). This approach promotes efficient learning by sharing feature representations while allowing for independent optimization of the classification heads tailored to their specific tasks, potentially leading to improved performance on both emotion and trigger classification tasks.

3.2.1 BERT Model and Tokenizer

In our approach, we instantiate one of the fundamental components of the *AutoModel* library, namely *AutoModelForSequenceClassification*. This class serves as a crucial element in our model pipeline, facilitating the acquisition of input text embeddings. Additionally, we employ *AutoTokenizer* to seamlessly load the corresponding tokenizer for the architecture of the chosen model. Specifically, we leverage a pre-trained BERT model initialized from the checkpoint 'bert-base-uncased' to capture rich contextual representations of input utterances. This pre-trained BERT model has been meticulously fine-tuned for the specific task at hand, namely emotion and trigger prediction, ensuring that it possesses the requisite domain knowledge and linguistic understanding to excel in this task.

As part of our experimentation, we explore two distinct model configurations to discern their impact on performance and computational efficiency:

- Freezed: In this setting, we opt to freeze the weights of the BERT embedding layer. By doing so, we preserve the pre-trained contextual embeddings learned by BERT and solely fine-tune the classifier heads appended on top of the model architecture. This approach allows us to leverage the rich semantic representations captured by BERT while focusing our optimization efforts on task-specific classification layers.
- Full: In contrast to the freezed configuration, we adopt a full fine-tuning strategy wherein we fine-tune the entire model architecture, including both the BERT embedding layer and

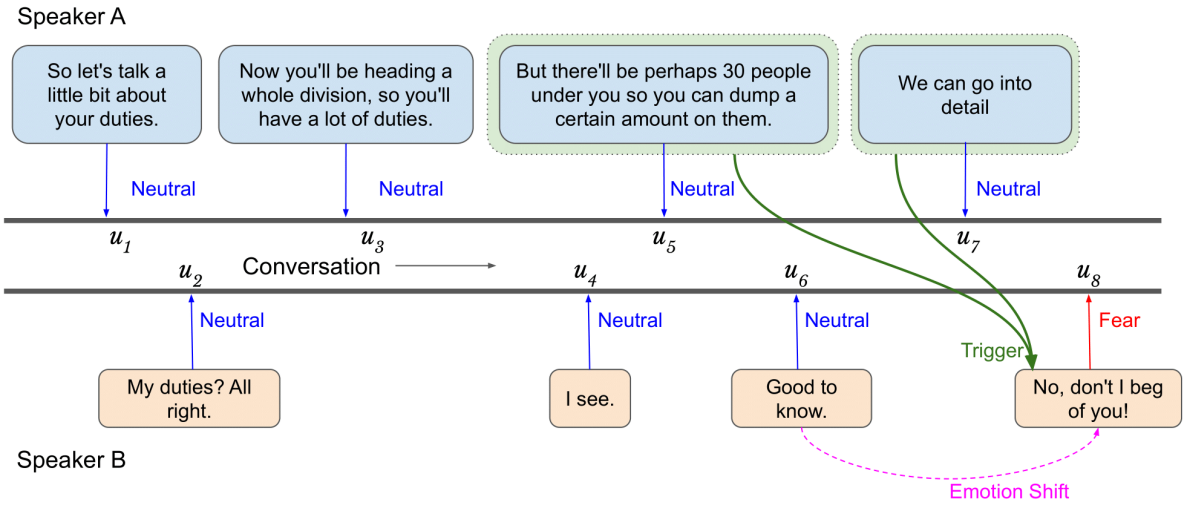

Figure 2: An example scenario depicting the utterance that trigger a flip in emotion of a speaker within a dialog (Kumar et al., 2021)

the classifier heads. This comprehensive fine-tuning approach enables the model to adapt more flexibly to the nuances and intricacies of the target task, potentially leading to improved performance by allowing the model to learn task-specific features at all levels of abstraction.

By systematically exploring these two model settings, we aim to gain deeper insights into the trade-offs between model complexity, computational resources, and predictive performance. This empirical investigation is essential for informed decision making about the optimal configuration of our model for real-world deployment in emotion and trigger prediction tasks.

3.2.2 Task-Specific Classification Heads

Instead of separate models for each task, the model Figure 1, employs two independent classification heads:

- *Emotion Prediction Head*: This head consists of a linear layer that operates on the encoded representation of the first word piece (usually corresponding to the "[CLS]" token) from the final BERT layer output. This focuses on capturing the overall sentiment of the dialogue for emotion classification (multi-label output).

- *Trigger Prediction Head*: This head employs a Long Short-Term Memory (LSTM) network to process the entire sequence of encoded representations from the final BERT layer output. LSTMs are well suited for capturing sequential information, which might be crucial to identify trigger events within the dialogue context. The LSTM outputs are then fed into a feed-forward neural network with multiple layers (linear layer, ReLU activation, another linear layer, and a sigmoid activation function) for trigger classification (binary output).

In summary, the architectural design of the PyTorch model embodies a systematic approach to emotion and trigger prediction in textual data. By integrating advanced deep learning techniques, including BERT-based classification and LSTM-based prediction, the model demonstrates robust performance in jointly predicting emotions and triggers, thereby offering valuable insights into text analysis tasks.

4 Data

The datasets for these tasks comprise manually annotated conversations focusing on emotions and triggers for emotion shifts. This data set is an extension of MELD (Poria et al., 2017),(Kumar et al., 2024), an established Emotion Recognition in Conversation (ERC) dataset that features monolingual English dialogues. This dataset incorporates annotations specifically tailored for emotion-flip reasoning, sourced from the popular TV series F.R.I.E.N.D.S. Each dialogue comprises utterances attributed to individual speakers and labeled with one of seven emotion categories: anger, disgust, fear, sadness, surprise, joy, or neutral.

Emotion-flip reasoning annotations within this dataset focus on identifying trigger utterances responsible for emotional transitions within dialogues. As shown in Figure 2, Trigger utterances are designated with a label of 1 if they directly induce an emotional shift from the speaker's preceding utterance, and 0 if they do not contribute to the transition. Clear guidelines defining triggers were established to guide the annotation process:

- Any utterance directly influencing an emotional change is labeled as a trigger.
- The trigger utterance can originate from the same or a different speaker compared to the target utterance.
- The target utterance itself can qualify as a trigger if it contributes to the emotional transition.

The dataset contains incremental dialogues, i.e, many rows represents the same dialogue with added utterances. Thus assigning unique dialogue_idx to the rows with non incremental utterances. A new dialogue is assumed to start whenever the set of utterances in the current row is not a subset of the set of utterances in the next row. The distribution of dialogs within the training set and the validation set are very different thus we shuffle the dataset making sure that the training and validation sets are distributed uniformly, leading to improve in overall performance of the models from previous not shuffled dataset.

A *context window* is created around each utterance within a dialogue. This window captures a certain number of utterances before the current utterance, here we consider the previous utterance along with the current one, allowing the model to consider the conversational context when predicting emotions and triggers. Including the previous utterance along with the current one shows a little increase around 5.36% in the predicion scores. The max_length is equal to the max no of tokens calculated for window size = 2.

5 Experimental setup and results

5.1 Setup

Architectures/Configurations: BERT (Bidirectional Encoder Representations from Transformers) was used as the backbone architecture for feature extraction.Two simple classifiers, random and a majority classifier, were employed for classification tasks.

We considered a *batch_size=1*, thus processing one dialogue at a time. Each dialogue has multiple utterances, thus tokenized dialogue tensor has shape $[batch\_size, n\_utterances, n\_max\_tokens]$, where *batch_size=1*, hence removed the first dimension and passed input of shape $[n\_utterances, n\_max\_tokens]$. Two seperate losses are computed *emotion_loss* and *trigger_loss* and combined weighting the values, and the parameters are updated through backward propagation. *CrossEntropyLoss()* is used for multi-class Emotion classification and *BCELoss()* Binary Cross Entropy Loss is used for binary Trigger classification.

Through a rigorous hyperparameter tuning process, we achieved optimal performance for the model. This involved standardization through careful selection of hyperparameters. We opted for training the model for 20 epochs, utilizing a learning rate of $1e-5$. The model architecture consisted of an 8-layer Long Short-Term Memory (LSTM) network with a hidden layer size of 128 units. Additionally, a feed-forward neural network with a hidden layer size of 64 units was employed. An AdamW optimizer with weight decay $(1e-4)$ is used for optimization, and a linear learning rate scheduler with warmup is implemented to gradually increase the learning rate during the initial training phase.

To ensure reproducibility of the training results across multiple runs, we iterated the training and validation process over a predefined set of random seeds. For each seed, a unique directory is created to store the trained model and potentially other training artifacts (e.g., logs). The directory name incorporates the seed information for easy tracking.

In our project, we employed a rigorous evaluation framework to assess the performance of our models, utilizing two primary metrics: Sequence F1 and Unrolled Sequence F1. These metrics serve as robust indicators of the effectiveness and accuracy of our dialogue understanding models across various contexts.

- **Sequence F1:** This is a fundamental metric in dialogue understanding tasks, computing the F1-score for each dialogue and subsequently reporting the average score. This metric provides insights into the model's ability to correctly predict emotions and triggers within individual dialogues, thereby offering a com-

Model	Metric	Seq. F1	Unroll. Seq. F1
Majority Classifier	Emotions	18.40	8.92
	Triggers	47.42	45.57
Random Classifier	Emotions	9.10 ± 0.49	11.53 ± 0.49
	Triggers	41.54 ± 0.76	43.73 ± 0.52
BERT freezed	Emotions	21.06 ± 1.17	8.92 ± 0.00
	Triggers	43.37 ± 0.10	45.57 ± 0.00
BERT unfreezed	Emotions	42.65 ± 1.32	42.21 ± 0.81
	Triggers	45.25 ± 0.67	50.68 ± 5.16

Table 1: Summary of Experimental Results

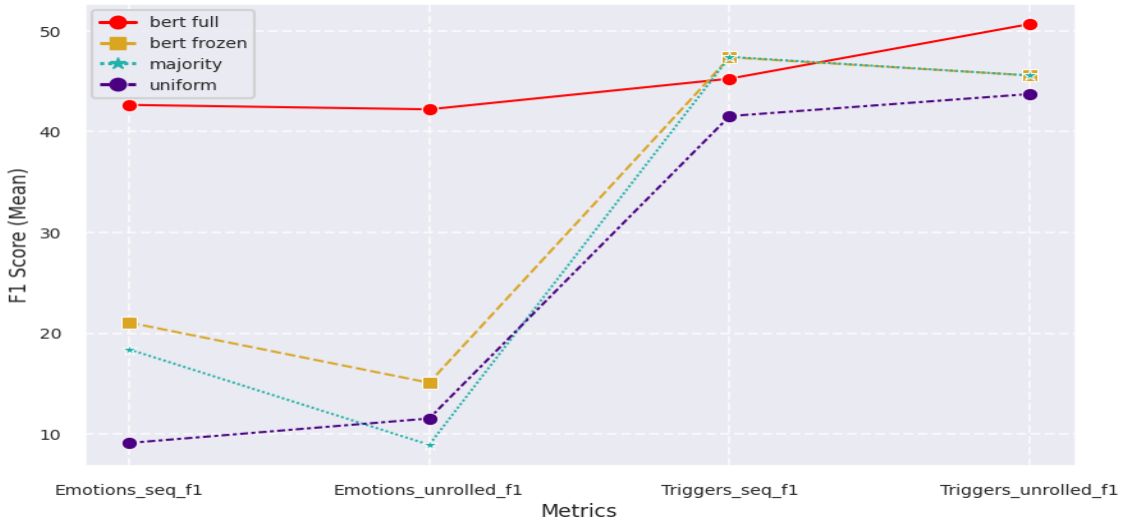

Figure 3: A comparison of the metrics across different models

prehensive assessment of its performance at the dialogue level.

- **Unrolled Sequence F1:** This metric involves flattening all utterances within dialogues and computing the F1-score across the entire dataset. By considering the dialogues as a continuous sequence of utterances, this metric offers a holistic evaluation of the model's performance, capturing its proficiency in comprehensively understanding and predicting emotions and triggers across various dialogue structures and lengths.

Both metrics were computed for both emotions and triggers labels, allowing us to evaluate the model's performance across multiple dimensions of dialogue understanding. This comprehensive evaluation approach ensures a thorough assessment of our models' capabilities and provides valuable insights into their effectiveness in real-world applications.

6 Discussion

In Table 3 findings from the validation and test phases of the Majority Classifier, Random Classifier, BERT freezer, and BERT unfreeze, for Emotions and Triggers are shown.

6.1 Quantitative Results

The *unfrozen BERT* model achieved the best performance for emotion classification, with an average sequence F1 score of 42. 65%. This is significantly higher than the other models, with *frozen BERT* achieving 21.06% and the baseline models ('majority' and 'uniform') falling below 20%. Although *unfrozen BERT* also performed well for trigger classification (average sequence F1 score of 45. 25%), there was less variation between models on this task. Interestingly, the baseline models achieved similar or even slightly better scores for trigger classification compared to the *BERT* models.

The unfrozen BERT Model demonstrated exceptional performance across all metrics, outperforming other models significantly. Its success can be

Emotion	Precision	Recall	F1-Score	Support
neutral	0.72	0.66	0.69	1522
surprise	0.52	0.77	0.62	432
fear	0.07	0.20	0.11	71
sadness	0.49	0.48	0.49	365
joy	0.52	0.48	0.50	603
disgust	0.25	0.29	0.27	145
anger	0.54	0.33	0.41	478
Accuracy			0.56	3616
Macro Avg	0.44	0.46	0.44	3616
Weighted Avg	0.58	0.56	0.56	3616

Table 2: Emotion Classification Metrics *unfrozen BERT*

Label	Precision	Recall	F1-Score	Support
0	0.90	0.82	0.86	3088
1	0.30	0.47	0.37	528
Accuracy			0.77	3616
Macro Avg	0.60	0.64	0.61	3616
Weighted Avg	0.81	0.77	0.79	3616

Table 3: Trigger Classification Metrics *unfrozen BERT*

attributed to several key factors. First, BERT captures contextual information by considering both left context during pre-training. This contextual understanding allows BERT to represent complex relationships within the data. Second, fine-tuning on our specific task further adapts the pre-trained BERT to our domain, making it more effective for our target problem. Third, BERT's extensive pre-training on a massive amount of text data enables it to learn rich representations that generalize well to various downstream tasks. Additionally, BERT's attention mechanism attends to relevant parts of the input sequence, capturing long-range dependencies and identifying crucial context for making predictions. In summary, the combination of contextual embeddings, fine-tuning, and transfer learning contributes to the *unfrozen BERT* Model's superior performance. Further investigation of hyperparameters and domain-specific features could enhance its effectiveness even more. The weakness of classifiers in this task lies in their simplistic approach, inability to capture context, and reliance on majority voting, especially in large languages.

6.2 Error Analysis

To analyze the prediction errors made by our models, we generated a classification report and confusion matrices. The report indicates that our model struggles when predicting a tag associated with a class having a small number of supports (label 0). We've also observed significant imbalance in the dataset, with the target variable having more observations,in Emotions Statistics, 'natural' and in Triggers Statistics, (label 1).

Our *unfrozen BERT* is the best performing model, demonstrating a strong ability to classify emotions across the dataset (Table 2). High accuracy and precision scores for various emotions indicate the model's effectiveness in discerning nuanced emotional states. Notably, the precision and recall scores for the "neutral" emotion are particularly high, signifying the model's proficiency in identifying instances lacking significant emotional content.

In contrast, the trigger classification results for best-performing model in Table 3 reveal a clear performance disparity between non-trigger and trigger instances. Precision, recall, and F1 scores for triggers (label 1) are significantly lower compared to non-triggers (label 0). This imbalance suggests challenges in distinguishing triggers within dialogues. Potential reasons include class imbalance in the data (fewer trigger instances) and limitations in capturing the context or nuance needed for accurate trigger identification. Further refinement of the model or the incorporation of more contextual features may be necessary to improve trigger sensitivity. Examining the confusion matrix results allows us to identify which tags are most frequently misclassified and understand the specific errors made by our classifier. For instance, some emotions classified as natural like joy, anger and fear.

7 Conclusion

To sum up, our project explored how well computers can understand human emotions in conversations. We tested different models, focusing on one called BERT, to see how accurately they could recognize emotions and what triggered those changes.

Our main finding was that the *unfrozen BERT* model performed the best, showing the highest accuracy in predicting both emotions and triggers. This means it is good at understanding the subtle nuances of human emotions in conversations.

However, we also found some areas to improve. We noticed that the models struggled with certain types of emotions and triggers and there were still challenges in handling different style of conversation. Looking ahead, we believe there's more to learn and improve. We can refine the models further, create better datasets, and work with experts from different fields to deepen our understanding of emotions in conversations. Although the model showed some ability to identify emotions and triggers, overall performance remained below average. Experimentation with alternative approaches like teacher forcing and RoBERTa (outside the scope of this project) did not yield significant improvements.

These findings suggest that current models may not fully capture the complexities of emotion and trigger classification in this domain. Looking ahead, exploring the potential of Large Language Models (LLMs) for this task presents a promising avenue for future work, as their ability to handle complex language structures might prove advantageous. In short, the project takes us a step closer to making computers understand human emotions better, which could lead to more intelligent and empathetic AI systems in the future.

8 Links to external resources

Link to the GitHub repository: EDiRef_BERT

References

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3):169–200.

Shivani Kumar, Md Shad Akhtar, Erik Cambria, and Tanmoy Chakraborty. 2024. Semeval 2024 – task 10: Emotion discovery and reasoning its flip in conversation (ediref).

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.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. pages 873–883.