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

Riccardo Marvasi, Edoardo Saturno, Lucia Gasperini and **Arianna Albertazzi**
Master's Degree in Artificial Intelligence, University of Bologna
{ riccardo.marvasi, edoardo.saturno, lucia.gasperini, arianna.albertazzi3 }@studio.unibo.it

Abstract

In humans' conversations, a speaker's cognitive state is subject to change, often prompted by specific preceding utterances, potentially resulting in an emotional state flip. This report delves into the domains of Emotion Recognition in Conversations (*ERC*) and Emotion-Flip Reasoning (*EFR*). ERC seeks to decipher emotions expressed in a sequence of utterances, while EFR focuses on identifying those utterances triggering a speaker's emotional flip. Our approach employs a BERT-based model with an LSTM component, specifically tailored for both tasks and applied to the MELD dataset, a recognized benchmark for emotion recognition in multi-party conversations. Through a meticulous comparison with four baseline models, our findings underscore the model's remarkable efficiency, particularly in excelling at the task of EFR.

1 Introduction

The Emotion Recognition in Conversations (*ERC*) task plays a pivotal role in natural language processing, impacting applications like e-commerce (Narendra Gupta, 2010), social media (Nicoletta Calzolari, 2016) (Akhtar et al., 2020), and healthcare (Ellen Riloff, 2018). Unlike traditional emotion recognition in standalone text, *ERC* delves into dialogues, revealing emotional dynamics within conversations. This extension proves crucial for understanding interpersonal communication, enhancing human-computer interaction, and fostering empathetic conversational agents (Shivani Kumar, 2021). *ERC*'s unique challenge lies in deciphering emotions within a sequence of utterances, involving multiple speakers shaping the evolving emotional landscape.

Expanding this task to Emotion-Flip Reasoning (*EFR*) adds another layer of significance. *EFR* identifies trigger utterances inducing changes in emotional trajectories during a conversation. This aspect is fundamental for explicating intricate shifts in emotional states, providing valuable insights into the underlying mechanisms that shape conversational dynamics. In essence, *ERC* and its extension, *EFR*, play vital roles in unraveling the complexities of human emotions within dynamic conversational contexts.

The Emotion Recognition in Conversations (*ERC*) task is approached through various methods. *Lexicon-Based* Approaches utilize predefined emotional word lists, but they often struggle with nuances and context. *Machine Learning Models* employ algorithms for emotion classification based on features, while *Hybrid Approaches* combine lexicons and machine learning (Mayur Wankhade, 2022) (Svetlana Kiritchenko, 2014). Due to limitations in lexicon-based methods, neural models, particularly deep learning approaches, have gained prominence for their ability to capture complex contextual nuances and improve *ERC* performance.

Exploiting the contextual embeddings provided by *BERT* (Jacob Devlin, 2019), our model aims to capture intricate emotional nuances within dialogues. We opt for BERT due to its proven effectiveness in understanding complex language contexts, which is crucial for discerning sentiment patterns. Our model adeptly addresses both *ERC* and *EFR* tasks, leveraging the richer representation generated by BERT. For emotion recognition, a fully connected neural network operates on top of BERT's contextual embeddings, while the trigger recognition task employs a bidirectional *LSTM* and another feed-forward neural network. The strategic incorporation of both *BERT* and *LSTM* components enhances the model's capacity to capture complex relationships within dialogues. To prove the necessity of a model able to consider the full contest, in order to correctly identify the trigger utterances, we constructed a baseline in which the trigger classifier is a straightforward linear classifier applied on the top of BERT embeddings. Also,

in our study, we will prove the significance of a well-crafted embedding representation for achieving optimal results in *EFR* and *ERC* tasks. In order to do this we implement an additional BERT baseline where all parameters, excluding those of the the linear classifiers, are frozen. This setup allows for a comparative analysis between a model where BERT can fine-tune its embeddings and one where it cannot, elucidating the great impact of embedding preparation on overall performance.

We performed a comprehensive series of experiments utilizing dialogues from MELD, an acronym for *Multimodal EmotionLines Dataset*, for both training and evaluation purposes. To ensure the robustness and generalization of our model, we employed five different seeds. Within our specific setup, the model and baselines were trained on *3200* dialogues of varying lengths. This configuration was replicated for both the evaluation and test sets, each comprising *400* dialogues. Our experimental findings highlight the effectiveness of our proposed approach in addressing the challenges posed by both *ERC* and, notably, *EFR* within the MELD dataset. In the following sections, we will delve deeper into the specifics of our approach, detailing the implementation steps and justifications for the choices made.

2 System description

In this section, we detail the systematic steps taken to implement our strategy. We begin by sourcing the MELD dataset from a JSON file, followed by a series of preprocessing steps. These include encoding categorical labels for emotions, as well as zero-transforming *NaN* elements. The dataset is further manipulated following a dialogue-centric approach to capture contextual dependencies and is subsequently divided into training, validation, and test sets for model evaluation and error analysis. For a comprehensive understanding of the dataset and detailed insights into data preprocessing, readers are encouraged to refer to the *Data* section below.

We created individual *DataLoader* objects for each dataset, serving as input data during training. The training process was uniform across all BERT models and was defined through a dedicated external custom function. This approach proved beneficial, allowing us to concentrate on refining the architecture's structure when defining the model classes.

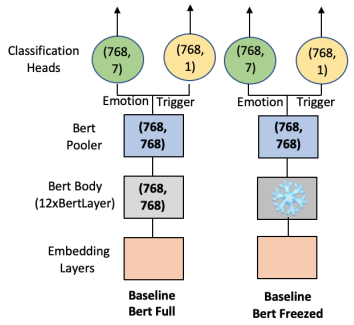

Figure 1: Simplified sketch of the baseline models (Bert Full and Bert Freeze). In the actual structure, there would be three embedding layers, followed by a LayerNorm (768,) with dropout (0.1): word embedding (30522,768), position embedding (512, 768), token type embedding (2, 768).

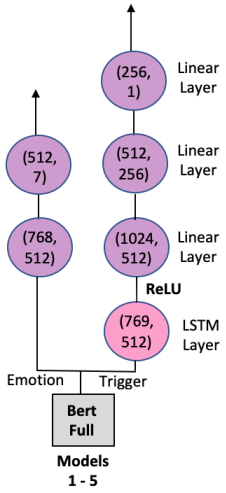

Figure 2: Simplified sketch of the models from 1 to 5.

The defined models consist of four baseline models and five custom models, each associated with a distinct random seed. Two of the baselines are a *Majority Classifier*, and a *Random Classifier*. The other baselines (*figure 1*) are more complex: one, named *BERT Freeze*, is a BERT model in which all parameters in the BERT body are freezed during training, so only the classification heads are allowed to fine-tune. The other, *BERT Full*, shares the same structure as *BERT Freeze*, but in this case both the classification heads and the BERT body are trainable, making it a more sophisticated model with the ability to learn intricate data complexities. To address both *ERC* and *EFR* tasks within a single model/baseline, we incorporate two classifier heads on top of the BERT embedding.

The architecture of the custom BERT model (*figure 2*), named *MyBert*, deviates from the *BERT Full*. The key distinction lies in the use of a weighted loss function and more complex classification heads.

Unlike the *BERT Full*, which employs a single linear layer for each emotion and trigger, our BERT model incorporates two linear layers for emotion, an *LSTM* layer and three linear layers for triggers.

Going deeper into the architecture's details, the trigger prediction component integrates both an *LSTM* layer and a feed-forward neural network. The *LSTM* layer processes inputs coming from the BERT model and incorporates the predicted emotion from the emotion branch, providing bidirectional context for trigger prediction. By concatenating the emotion prediction with the output of the BERT body, the model gains additional information for reasoning whether the given utterance is a trigger. The Feed Forward neural network that uses the output of the *LSTM* layer is structured in a pyramidal way: three Fully Connected Layers (*FCL*) with decreasing number of hidden units preceded by a *ReLU* activation function. The final output ultimately produces a single value. This configuration aims to introduce non-linearity and enhance the model's capacity to capture intricate patterns in the data. Regarding emotions, the chosen structure is again a Feed Forward neural network composed of two *FCL*, again with decreasing number of hidden units. This repeated choice, consistent for both triggers and emotions, aims to facilitate better generalization by reducing the dimensional representation of the data. During model training, we collected essential information to plot the validation loss behavior for each model. This was facilitated by the custom training loop structure. Specifically, at intervals of *400* iterations, the loss values were recorded and stored in a list, later utilized for plotting curves.

Following model training, custom class methods were employed to generate predictions for both emotions and triggers on the test set. These predictions were then used to construct confusion matrices for the various models, offering valuable insights into their behavior beyond F1 score performance. Subsequently, we visualized selected utterances from the test set, presenting actual labels for triggers and emotions alongside predictions from the BERT baselines and *MyBert*. This additional step contributed valuable insights for understanding the way of acting of each model.

3 Data

The dataset used in this study is derived from MELD (Soujanya Poria, 2019), a resource designed for in-depth analysis of emotions and sentiments in dialogues extracted from movies. This dataset encompasses textual, visual, and acoustic modalities, with a specific emphasis on the textual component. The particular dataset we utilized for ERC and EFR tasks consists of 4000 short English dialogues extracted from the TV series Friends.

Each utterance within these dialogues includes information such as the speaker's name, the corresponding emotion categorized into one of seven labels (*joy, anger, neutral, surprise, disgust, sadness, fear*), and a binary label ("*1*" or "*0*") indicating whether the utterance served as a trigger for the dialogue. Preliminary analyses revealed a significant imbalance in the dataset. The majority of utterances were labeled with the emotion "neutral" (*15.263* instances), while the other six emotions combined for a total of *19.737*. Furthermore, a substantial majority of the utterances were classified as "non-trigger" (*29.425* instances compared to *5.575* trigger utterances).

The data was extracted and shuffled at dialogue level and then structured in a Pandas Dataframe. In particular, in order to be able to perform preprocessing more easily, this dataframe was composed by a single utterance (and contextual information) per row. In this way "*NaN*" values handling process and the encoding of the emotion labels into numeric format could be performed quickly. Also, for a nuanced understanding of the context, a new column named "*Text*" was created, by concatenating the speaker's name with the corresponding utterance.

Following the preprocessing steps, we proceeded to tokenize the "Text" column and grouped all the tokenized utterances and labels belonging to the same dialogue. In this way, we generated a new dataframe, this time at the dialogue level, comprising the "input indexes" and "attention mask" values obtained through tokenization.

Following this, the dataset was divided into training, validation, and test sets using predefined percentages, in particular 80 percent for the training set and the remaining 20 percent divided equally between validation and test sets. This splitting guaranteed a balanced representation of data in each set, contributing to the robustness of the model. To minimize the necessity for padding, we opted to truncate each utterance after a specific number of words. This decision was informed by analyzing a boxplot of the length distribution for each utterance. Notably, *99%* of the data fell below the

threshold of *30* words. Consequently, we selected this value as the maximum sequence length, ensuring that truncation at this point would not lead to significant information loss. We need to specify that for tokenization, the *Hugging Face Transformers* (Thomas Wolf, 2020) library was employed, specifically the autotokenizer of the chosen BERT architecture ("*bert-base-uncased*").

In summary, pre-processing steps addressed handling missing values, encoding categorical labels, structuring the dataset, and preparing text data for BERT-based tokenization.

4 Experimental setup and results

The architectures we used were of five different kinds. They included *BERT Freeze*, *BERT Full*, *Majority Classifier*, *Random Classifier*, and five custom models labeled from *Model1* to *Model5*. All BERT parts were of "bert-base-uncased" type, in accordance with the tokenizer.

As elaborated in Section 2, *Bert Freeze* incorporates two parallel classification heads atop its BERT body, whereas *Bert Full* shares the same structure but with trainable weights for all model parameters. The *Random* and *Majority* classifiers' names are self-explanatory, while the custom BERT model mirrors the general BERT architecture but features more complex classification heads and a nuanced forward pass. Importantly, *MyBert* model employs a slightly different approach to the loss function. Initially, the ERC task employed *CrossEntropyLoss*, while the EFR task used *BCEWithLogitsLoss*. However, this approach presented challenges as the value of *BCEWithLogitsLoss* significantly surpassed that of *CrossEntropyLoss*, resulting in a skewed training towards triggers and decreased performance. Consequently, we modified the forward pass of the model to utilize simple *BCELoss*, aligning its scale with *CrossEntropyLoss*. Additionally, to prevent skewing towards *CrossEntropyLoss*, we multiplied *BCELoss* by two.

We maintained uniformity in the hyperparameters across models. This approach contributes to the reliability and comparability of the results obtained from different models. Throughout the training process, model optimization utilizes the *AdamW* optimizer, initialized with a learning rate of $1 \times e^{-5}$. This choice is grounded in the adaptability of Adam to automatically adjust the learning rate, mitigating the necessity for meticulous manual fine-tuning. The selection of *4* epochs strikes a balance between efficient training time and achieving optimal performance. Notably, training multiple models, including 7 BERT models, can pose challenges with an excessive number of epochs.

For assessing the models' performance, F1 scores are utilized for both emotion and trigger predictions. Two approaches are employed for F1 score computation: the first one involves calculating the F1 score for each dialogue and then averaging across all dialogues; the second one is the "unrolled" F1 score, which evaluates utterances individually. Notably, the unrolled F1 score tends to be higher than the average F1 score, as it is independent of error distribution among dialogues. At evaluation time, we obtained the results reported in *Table 1*.

In conclusion, the detailed results presented in this study contribute to a deeper understanding of the strengths and weaknesses inherent in each architecture. This comprehensive evaluation establishes a robust foundation for informed decision-making about future model architectures types and enhances our understanding of the models' capabilities.

	Emotion Unrolled	**Emotion Average**	**Trigger Unrolled**	**Trigger Average**
Bert Freezed	0.205	0.286	0.457	0.482
Bert Full	0.918	0.886	0.536	0.508
Mayority Classifier		0.130		0.423
Random Classifier		0.085		0.457
Model 1	0.900	0.854	0.688	0.638
Model 2	0.901	0.860	0.670	0.618
Model 3	0.892	0.849	0.682	0.629
Model 4	0.906	0.862	0.683	0.635
Model 5	0.900	0.848	0.676	0.631
Average across seed	0.8998	0.8546	0.6798	0.6302

Table 1: F1 scores for each category for the Baseline Model and for the Model that beat the baseline (trained for the 5 seeds). For Random and Majority Classifiers is considered the Macro F1.

5 Discussion

The baseline models, specifically the Majority Classifier and Random Classifier, served as foundational benchmarks. The *Majority Classifier* achieved a Macro F1 score of *0.13* for Emotion, indicating its mediocre performance in identifying

the majority class, while the Random Classifier yielded a Macro F1 score of *0.085* for Emotion, highlighting the poor results obtained by random predictions in this task.

Despite undergoing a moderate adaptation, *BERT Freeze* demonstrated inferior performance compared to other BERT variants, as reflected in its F1 scores. The Emotion Unrolled F1 score, registering at *0.205*, underscores the significance of actively training the BERT parameters to achieve a more refined representation of the data. Similarly, the Trigger Unrolled F1 score, at *0.457*, suggests opportunities for optimization in trigger prediction performance.

In stark contrast, the *BERT Full* model, demonstrated good performance. By undergoing training on the complete architecture, the model showcased a remarkable proficiency in capturing intricate patterns within the data. The achieved F1 scores were *0.918* for Emotion Unrolled, which is quite good, and *0.536* for Trigger Unrolled, which has room for improvement. These F1 scores, in particular for emotion, highlight the effectiveness of *Bert Full* in handling the complexity of the given tasks and show its potential as a robust and competitive baseline to beat.

Moving to the custom models a consistent performance trend emerged across the five seeds. Averaging the results across seeds, these models demonstrated robust Emotion and Trigger predictions, as evidenced by values such as *0.8998* for Emotion Unrolled and *0.6798* for Trigger Unrolled. This consistency shows reliability and generalization capability across various data instances and training scenarios.

Now focusing on a more qualitative analysis of the models we can observe that *BERT freeze* exhibits limited effectiveness across both the classification tasks. Although there appears to be a relative performance increase for triggers, a closer examination of the confusion matrix exposes its behavior as a majority classifier, categorizing all examples as "*0*". Similarly, in emotion classification, a majority classifier behavior is evident, particularly for the "*neutral*" emotion which, in this case, represents the majority class. On the contrary, *Bert Full* model excels in emotion classification to such an extent that no custom model surpasses its performance in this domain. Turning our attention to trigger classification, it is noteworthy that again a majority classifier behavior emerges, predominantly influenced by the imbalanced distribution of class "0" instances in the dataset. Regarding random and majority classifiers both have comparable scores, with respect to the BERT baselines, in trigger classification. However they have the lowest performance in emotion classification. The custom model demonstrates the ability to diversify its learning by simultaneously addressing both classification tasks. It achieves comparable performance to *Bert Full* in ERC and significantly outperforms it in EFR. This accomplishment is attributed to the strategic choices aimed at contextualizing triggers data that was fed inside the architecture.

It's noteworthy to mention that the custom model exhibits significantly faster convergence compared to any other baseline model. For instance, if the training is limited to just one or two epochs (which might be considered in situations with time constraints), the emotion F1-score for the custom model reaches a range of 0.7-0.8. In contrast, the *BERT Full* model achieves only a 0.6 F1-score within the same training limitations.

Concerning the analysis of errors within the outcomes of custom models, it's essential to highlight the following key observations:

- *Trigger Classification*: The confusion matrix shows that many examples are misclassified but this is less concerning for the "0" labels, as over *90* percentage of examples are correctly classified. A different situation can be observed for "1" labels where the misclassified examples are more than the correct ones. In order to solve this, more complex classification heads can be implemented or the weight of the "1" class can be increased in the loss function.

- *Emotion Classification*: Overall, the model demonstrates good performance in emotion classification. However, there is room for improvement in the "*Neutral*" and "*Joy*" classes, while the error rate for other classes is slightly lower. The instances where prior utterances set a context, and the subsequent ones change it, thereby altering the perceived sentiment of a sentence, contribute to a higher error rate in the "Joy" category compared to other classes. During our error analysis we realized how some of the misclassifications may be due to "confusing" labels in the data. An example illustrating this phenomenon can be observed by examining the following dialogue:

- *Ross: She said what?*
- *Emily: She said, "If I'm not gonna be happy getting married somewhere that we find in a day, well then we should just postpone it."*
- *Ross: Postpone it?*
- *Ross: Emily, do you think Monica realises how much our parents spent on this wedding?*
- *Ross: Do you my sister's teeny-tiny little brain comprehends that people took time out of their lives to fly thousands of miles to be here, huh?*
- *Ross: This isn't right.*

For a human observer it is quite obvious that Ross is nervous and is getting angry at the idea that his sister is post-poning the wedding. However the ground truth tag of the "*postpone it*?" utterance is "*Joy*" which can result confusing. The custom model assigned the "Neutral" tag to the utterance. This is to demonstrate that the data can actually contain errors and this will impact the performance of the model even if slightly.

Summarizing, in order to improve the results in the future, we could point out:

- *Error-Specific Analysis*: Conducting a detailed analysis of misclassifications, with a focus on examples where models struggle, may help in giving creative ideas on how to "squeeze out" as much performance as possible from the categories with the most errors.

- *Ensemble Approaches*: Investigate ensemble approaches by combining predictions from multiple models. This strategy may further boost overall performance and mitigate the impact of seed variability. (Jacqueline Kazmaier, 2022)

- *Larger Models*: The use of models like "RoBERTa" could also improve performance by being able to learn even more complex data patterns than BERT (Inhan Liu, 2019). More than that these kind of models, that represent evolution of the original BERT architecture, fix specific weaknesses of the original model.

6 Conclusion

In this study, we embarked on a comprehensive exploration of emotion recognition and trigger identification within dialogues. Our main findings highlight the success of *MyBert* in surpassing the challenges posed by the BERT Full baseline, showcasing enhanced capabilities in capturing nuanced patterns related to emotion and trigger classification. We observed that the integration of trigger predictions within BERT embeddings and complex classification heads, as well as introducing weighted loss functions played a crucial role in achieving promising performance across both the classification tasks. However, it is important to acknowledge the limitations of our current implementation. One notable challenge lies in the classification of triggers, because even with all the strategic choices implemented in *MyBert*, model performances could not top *0.70* of F1 score. This is due to the inherent difficulty of this kind of task. Future work could focus on increasing the size of the chosen model or increase the complexity of the classification heads.

7 Links to external resources

SemEval 2024 Task 10: Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) https://lcs2.in/SemEval2024-EDiReF/

MELD Dataset for EDiRef https://drive.google.com/drive/folders/1YgUU9nwFr9UiJKmGbFS9ByuL5fQWp8MO

AutoTokenizer - Hugging Face https://huggingface.co/docs/transformers/model_doc/auto

References

Md Shad Akhtar, Asif Ekbal, and Erik Cambria. 2020. How intense are you? predicting intensities of emotions and sentiments using stacked ensemble [application notes]. *IEEE Computational Intelligence Magazine*, 15(1):64–75.

Julia Hockenmaier Jun'ichi Tsujii Ellen Riloff, David Chiang. 2018. Fine-grained emotion detection in health-related online posts. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1160–1166, Brussels, Belgium. Association for Computational Linguistics.

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

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

Jan H. van Vuuren Jacqueline Kazmaier. 2022. The power of ensemble learning in sentiment analysis. *Expert Systems with Applications*, 187.

Chaitanya Kulkarni Mayur Wankhade, Annavarapu Chandra Sekhara Rao. 2022. A survey on sentiment analysis methods, applications, and challenges. 55:5731–5780.

Giuseppe Di Fabbrizio Narendra Gupta, Mazin Gilbert. 2010. Emotion detection in email customer care. *NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 10–16.

Thierry Declerck Sara Goggi Marko Grobelnik Bente Maegaard Joseph Mariani Helene Mazo Asuncion Moreno Jan Odijk Stelios Piperidis Nicoletta Calzolari, Khalid Choukri. 2016. Emotion analysis on twitter: The hidden challenge. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16):3953–3958.

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

Navonil Majumder Gautam Naik Erik Cambria Rada Mihalcea Soujanya Poria, Devamanyu Hazarika. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations.

Saif M. Mohammad Svetlana Kiritchenko, Xiaodan Zhu. 2014. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Victor Sanh Julien Chaumond Clement Delangue Anthony Moi Pierric Cistac Tim Rault Remi Louf Morgan Funtowicz Joe Davison Sam Shleifer Patrick von Platen Clara Ma Yacine Jernite Julien Plu Canwen Xu Teven Le Scao Sylvain Gugger Mariama Drame Quentin Lhoest Alexander M. Rush Thomas Wolf, Lysandre Debut. 2020. Transformers: State-of-the-art natural language processing.