NLP Course Project

Pelinsu Acar, Rubin Carkaxhia, Calin Diaconu and **Ruud Johannes Wilhelmus Korsten**
Master's Degree in Artificial Intelligence, University of Bologna
{ pelinsu.acar, rubin.carkaxhia, calin.diaconu, ruudjohannes.korsten }@studio.unibo.it

Abstract

The Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) task poses a significant challenge in understanding emotional dynamics within conversational dialogue. In this report, we present an investigation into the development of a system aimed at identifying emotions and detecting emotional shifts in dialogues. Leveraging a BERT-based architecture with specialized classification heads for triggers and emotions, our objective is to enhance emotional discourse analysis. We compare our approach against baseline models, including random and majority classifiers, to assess its effectiveness. Our experiments involve frozen and full fine-tuning configurations of the BERT model. The main findings highlight the BERT full model's superior performance in trigger detection, while challenges persist in accurately classifying emotions. Class-wise analysis and error examination show model biases and limitations, providing insights for the future improvements of our model.

1 Introduction

The Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) task addresses the inexplicit nature of emotional discourse within conversational dialogue, a critical aspect of natural language understanding. Understanding emotional dynamics within conversations is essential for developing empathetic conversational agents and enhancing human-computer interaction. Besides empathetic response generation, identifying triggers for emotional flips holds potential in impact monitoring, enabling informed decisions for downstream tasks by organizations or individuals. Traditional approaches to this task often rely on statistical methods or simplistic classifiers such as Bayesian Networks (BN) or the Maximum Likelihood Principle (MLP), and Support Vector Machine (SVM), which may fail to capture the nuanced nature of emotions in dialogue. Therefore, our study explores the efficacy of a BERT-based architecture, inspired by recent advancements in natural language processing, to tackle the EDiReF task.

Our approach is motivated by the contextual understanding capabilities of BERT, which have shown promise in capturing complex linguistic patterns. By fine-tuning BERT for the specific requirements of the EDiReF task, we aim to leverage its contextual embeddings to improve emotional discourse analysis. Our experimental design includes the evaluation of the BERT model against baseline classifiers, providing benchmarks for performance comparison.

In our experimental setup, we observed training and validation loss over 10 epochs in total and evaluate the model on the test set for both baseline models and the BERT model with frozen and fine-tuned settings using sequence and unrolled f1 scores. Here are the main results of our work:

- Bert model with fine-tuning shows a significant improvement for the triggers comparing to the baselines. However, the model fails to discriminate different emotions and it over-predicts the most presented label in the dataset.

- Data cleaning and weighting for the imbalanced classes may lead to a higher f1 scores. Moreover, our model needs definitely a longer training to understand the emotional context better.

2 Background

The solution described in this report aims to solve the Emotion Discovery and Reasoning its Flip in Conversation (EDiReF) task, as described in [5]. It is actually part of the 10th problem of the SemEval 2024 competition [3], with an English

dataset (the only one of interest for the current project) of 4,000 entries provided by the organizers.

The purpose of this task is to design a system capable of working on conversational dialogue and understand its emotional discourse. As mentioned in [4], a response generation mechanism could use the detected triggers as feedback for conversation steering. This would be another step towards the creation of human-like conversational agents.

The problem can be seen as made up of 2 parts: the first one is emotion identification in each utterance, while the second phase consists in identifying between which of these utterances an emotion flip happened. This process and the labeling of the dataset are backed by research done in the domain of Psychology, with publications like [9], [7], and [6], covering topics related to emotion recognition in conversations.

3 System Description

3.1 Baseline Models

Two classifiers that don't require actual training, but give results based on statistics of the dataset, are implemented in order to have baseline values for the metrics.

DummyClassifier objects are used, both of them implemented in the scikit-learn library [8]. The first one gives follows a "uniform" strategy, where the output class is sampled from a uniform distribution. The second one is based on the "most_frequent" strategy, which needs to be initialized with the training data, and on any evaluation input it will return the same class, the one that was the most frequent in the training set.

3.2 BERT Models

A single BERT[1]-inspired architecture is used over three different experiments, that employ different layer-freezing configurations. These configurations will be explored in the "Experimental Setup and Results" section of the current report, while here the differences in architecture from the basic BERT will be described.

The structure of the entire model follows the one in the requirements, as illustrated in Figure 1. To achieve it, the implementation of BertForSequenceClassification from the Hugging Face library [10] is extended with a solution similar to a public one, made available in the MultiModal-Toolkit package [2]. The input of the model consists of the concatenated utterances of a data entry, and on top of the BERT model there is a classification head.

This head consists of 2 separate classifiers, which are actually fully connected layers, each with its own sigmoid activation, to limit the output values in the range 0-1. The first one is in charge of the triggers, while the second one is in charge of the emotions. The reason to keep these 2 classifiers separate is to be able to weight differently the losses for each of the 2 tasks. The output of these 2 classifiers is supposed to follow a one-hot encoding, with 24 outputs for the trigger (24 being the maximum number of utterances for an entry, from the entire dataset), and 24 * 7 (where 7 is the number of classes for the emotions). Since the outputs are actually, with the input following a one-hot encoding, the loss that will be used is Binary Cross-Entropy.

4 Data

The dataset used for training the models previously described originates from the competition mentioned in the Background section. It consists of a set of dialogues, called episodes, split into utterances. Each episode has a unique name, it contains lists of the names of the speakers associated to each utterance, the labels for each utterance's emotion (given as string, one of "neutral", "surprise", "fear", "sadness", "joy", "anger", and "disgust", but later transformed by the current solution into numerical classes) and trigger (with true - false values marked with 1.0 - 0.0 floating point values), and the utterances themselves, given as lists of strings. The set seems to originate from the script of the "Friends" TV series, being actually dialogues from some episodes of it.

Only the training set of the competition is used in the current solution. It contains 4,000 episodes, with some of them having non-unique dialogues (some episodes contain a set of utterances that is also included - but in an extended form, by adding new utterances - in the set of utterances of another episode). 833 episodes with completely unique sets of utterances have been identified. The distribution of the number of non-unique dialogues for each of these episodes is plotted in Figure 2, where it can be noticed that more than 84% of them have at least one subset of their dialogue in another episode.

This problem is addressed in the data split, where no episode containing a subset of utterances from another one will be separated from it in a differ-

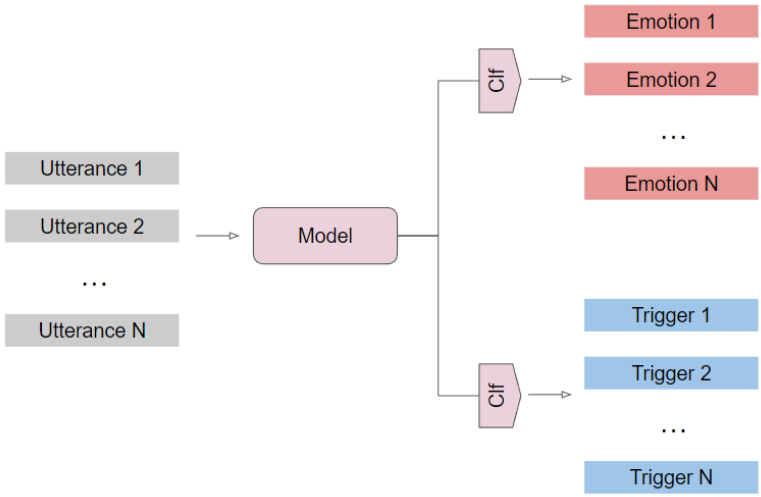

Figure 1: Model Structure

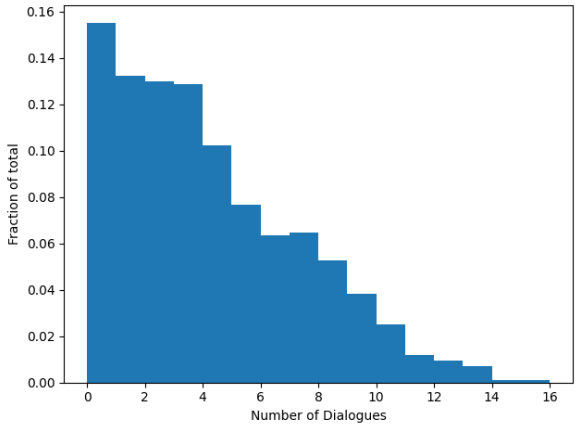

Figure 2: Plot of the number of non-unique episodes for each unique one

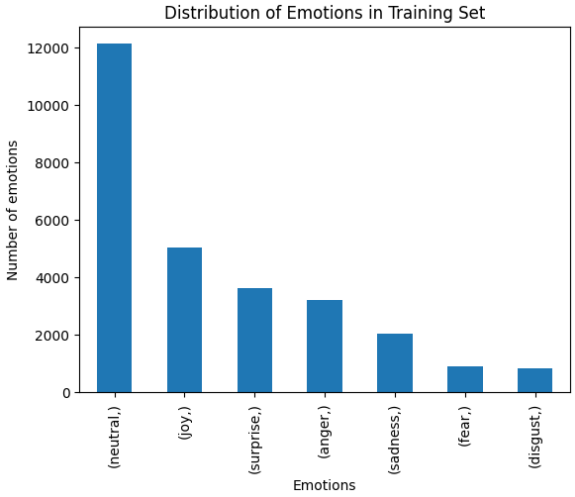

Figure 3: Plot of the distribution of emotions in training set

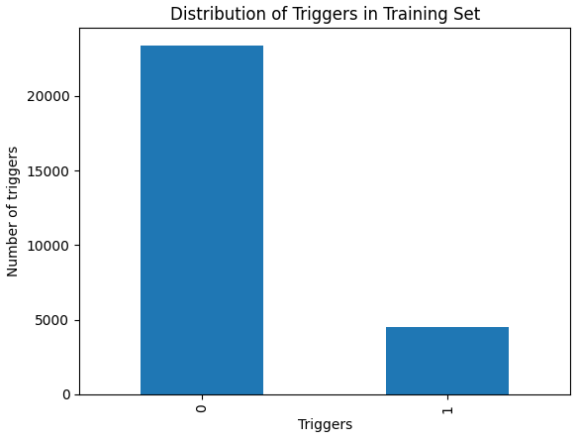

Figure 4: Plot of the distribution of triggers in training set

ent split. Since a 80-10-10 train-validation-test set distribution is aimed, the train set will contain 667 unique episodes, while validation and test will have 83 each. The number of non-unique sub-dialogues is not taken into consideration since, due to randomness, a similar distribution is expected among all 3 subsets. Thus, the definitive split will be into 3200 total episodes for the training subset, 412 for validation, and 388 for test.

We also analyzed the distribution of labels in our training set to see if we have any class imbalance. It can be noticed from Figure 3 and Figure 4 that we have a significant class imbalance in both emotions and triggers.

In terms of preprocessing, the default tokenizer of BERT is used, with the utterances of an episode being concatenated, with separator tags between them. The labels are padded to reach the maximum

number of utterances in an episode (which is 24), with the value -1, that will be ignored by the loss. The input is to be padded by a pre-built collator.

5 Experimental Setup and Results

Other than the changes regarding the classification head, the default, pre-trained, BERT models are loaded from the Hugging Face library, using the base uncased version. Each model is trained 5 times, with a different seed each time, having values from 1 to 5. The average over the 5 experiments of all metrics is reported in the current report. Both training and evaluation are performed in the Google Colab environment, having the following hardware resources allocated: Intel Xeon CPU, working at 2.20 GHz, 13GB of RAM, 100GB of disk space, and a T4 NVIDIA GPU.

On the topic of layer freezing, 3 configurations are used. The first one freezes the backbone, so all the BERT layers, fine-tuning only the classification head, the second configuration fine-tunes the entire network, and the third one is a mix of the 2, where it starts with the entire backbone frozen, and after that it gradually unfreezes layers, starting from the ones closer to the input.

The hyper-parameters are set as follows: small learning rate, of 2e-5, since a pretrained model is involved, batch size of 1, and weight decay of 0.01. The models are all trained over 10 epochs, and optimized with the default Adam.

As specified in the requirements, the metrics used for evaluation revolve around the F1 score, with "sequence" and "unrolled" variations reported. "Sequence" computes the F1 score over each episode, while "unrolled" evaluates individual utterances.

6 Discussion

6.1 Quantitative Analysis

According to the table 1, we can compare our models based on their f1 score and accuracy. Our evaluation mainly focuses on the sequence f1 score (micro average of f1 scores over each dialogue) and the unrolled f1 score (micro average of f1 scores flattening all utterances). The random classifiers and majority classifiers generally perform poorly compared to Bert models. They have low F1 scores across both triggers and emotions. Both Bert frozen and Bert full models outperform the random and majority classifiers in terms of micro f1 scores and accuracy for emotions. However, Bert frozen is not able to perform better than the baselines for triggers.

	Baseline Random		**Baseline Majority**		**Bert Frozen**		**Bert Full**	
	Triggers	Emotions	Triggers	Emotions	Triggers	Emotions	Triggers	Emotions
Sequence f1 score	0.6982	0.2692	0.7224	0.2524	0.5058	0.4747	0.7735	0.4725
Unrolled f1 score	0.7197	0.2230	0.7602	0.1966	0.4976	0.4211	0.8375	0.4176
Accuracy	0.72	0.22	0.76	0.20	0.50	0.42	0.84	0.42

Table 1: Triggers and emotions numerical results for all models, measuring the sequence, unrolled F1 score and the accuracy.

Emotions	**Baseline Random**	**Baseline Majority**	**Bert Frozen**	**Bert Full**
Anger	0.14	0.17	0.04	0.06
Disgust	0.00	0.00	0.02	0.00
Fear	0.01	0.00	0.22	0.00
Joy	0.15	0.13	0.04	0.03
Neutral	0.38	0.33	0.61	0.61
Sadness	0.06	0.00	0.01	0.01
Surprise	0.12	0.00	0.05	0.00

Table 2: Class-wise f1 scores of emotions for all models

Bert Full gives the best f1 scores and the accuracy for binary label of triggers. Considering multi-class emotions, Bert frozen and Bert full perform almost the same and they both struggle to effectively classify emotions, as indicated by low f1 scores and accuracy.

To understand the low f1 scores of emotions, we can analyze the class-wise scores for each model. Table 2 shows that both Bert full and Bert frozen perform better for the most presented class in the dataset which is *neutral*. While the baseline models completely ignore the most underrepresented emotions which are *disgust* and *fear*, Bert frozen performs less biased result assigning these labels to some utterances.

6.2 Error Analysis

In both Bert freeze and Bert full models, we observe instances where certain emotion categories are overpredicted compared to others. Here are some insights:

Triggers

- Considering the confusion matrices in tables 3 and 4, Bert frozen wrongly predicted triggers when there were none.
- This could be due to the model's tendency to over-predict triggers, possibly because of the imbalanced class (number of 0's is five times more than the number of ones).
- Although Bert full has also some issues detecting triggers, the number of false positives decreased significantly compared to Bert

freeze, indicating that full fine-tuning helped the model reduce this type of error. However, despite the decrease in false positives, false negatives increased.

- This might indicate that the model's focus on reducing false positives came at the cost of missing some actual triggers.

Emotions

- The most challenging pair of emotions are sadness and neutral, 88 samples belonging to neutral misclassified as sadness and 295 samples belonging to sadness misclassified as neutral.

- As it is seen from tables 5 and 6, Class 4 (neutral) appears to be overpredicted for both models, as indicated by the relatively high number of instances classified into this category

- This overprediction might suggest that the model is biased towards class 4 or cannot distinguish between class 4 and other emotion categories effectively.

- Indeed, as we have observed in the data analysis, class 4 is the most presented label.

Actual \ Predicted	Non-trigger	Trigger
Non-trigger	1451	1519
Trigger	247	298

Table 3: Bert Frozen - Confusion Matrix of Trigger

Actual \ Predicted	Non-trigger	Trigger
Non-trigger	2930	40
Trigger	531	14

Table 4: Bert Full - Confusion Matrix of Trigger

Actual \ Predicted	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	8	10	5	9	**308**	1	11
Disgust	0	1	1	3	62	0	4
Fear	0	0	6	2	54	0	4
Joy	10	13	4	16	**581**	2	18
Neutral	37	10	25	36	**1434**	24	58
Sadness	18	0	4	5	**292**	2	9
Surprise	8	12	0	7	**385**	3	13

Table 5: Bert Frozen - Confusion Matrix of Emotions

Actual \ Predicted	Anger	Disgust	Fear	Joy	Neutral	Sadness	Surprise
Anger	14	2	0	0	**321**	15	0
Disgust	3	0	0	0	57	11	0
Fear	1	0	0	3	59	3	0
Joy	18	2	0	10	**593**	21	0
Neutral	70	18	0	7	**1441**	88	0
Sadness	23	0	0	9	**295**	3	0
Surprise	10	2	0	18	**364**	34	0

Table 6: Bert Full - Confusion Matrix of Emotions

7 Conclusion

In this project, we tested and compared our custom Bert-base-uncased model with the baselines for the Emotion Discovery and Reasoning its Flip in Conversation. We modified Bert model adding another classification head for the emotions and triggers and evaluated it in two different settings (freezed and fine-tuned).

The Bert full model outperforms the Bert frozen model for triggers showing that fine-tuning all layers in the Bert full model allows for better adaptation to the specific task, resulting in improved performance compared to the frozen Bert model. On the other hand, for emotion detection, the pattern is similar to Bert freeze. The model's f1 scores for most emotions remain low, indicating that fine-tuning the entire model did not significantly improve its performance in emotion prediction.

Possible Improvements

- Collecting more balanced datasets that include a diverse range of examples for each emotion category to prevent the model from favoring certain classes over others.

- Experimenting with different fine-tuning techniques, such as adjusting learning rates or using class weights, to encourage the model to pay equal attention to all emotion categories during training.

8 Links to External Resources

More information on the topic of the competition can be found on their official website: https://lcs2.in/SemEval2024-EDiReF/ . Some ideas in the current solution were inspired by discussions and questions answered by the organizers on the official Google Groups channel: https://groups.google.com/g/ediref2024_group .

The notebook with the developed solution can be accessed at https://bit.ly/3OwdToG , with the weights stored at https://bit.ly/3Uypesd .

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[2] Ken Gu and Akshay Budhkar. 2021. A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico. Association for Computational Linguistics.

[3] Shivani Kumar, Shad Akhtar, Tanmoy Chakraborty, and Erik Cambria. 2023. Semeval 2024 task 10: Emotion discovery and reasoning its flip in conversation (ediref).

[4] Shivani Kumar, Shubham Dudeja, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Emotion flip reasoning in multiparty conversations.

[5] 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.

[6] Richard S Lazarus and Susan Folkman. 1984. *Stress, appraisal, and coping*. Springer publishing company.

[7] JHM Mooren and IAMH Van Krogten. 1993. Contributions to the history of psychology: Cxii. magda b. arnold revisited: 1991. *Psychological reports*, 72(1):67–84.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[9] Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances.

[10] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.