Emotion Discovery and Reasoning its Flip in Conversation NLP Course Project

Marco Panarelli, Jacopo Meglioraldi, Mihail Stamenov
Master's Degree in Artificial Intelligence, University of Bologna
{ marco.panarelli, jacopo.meglioraldi, mihail.stamenov }@studio.unibo.it

Abstract

This paper addresses the challenge of Emotion Recognition in Conversations (ERC) and Emotion Flip Reasoning (ERF) as defined by the EDiReF shared task competition. The aim is to classify emotions in each utterance and identify trigger utterances that cause an emotion shift within dialogues. Leveraging transformer-based models, we compare the performance of BERT and ELECTRA architectures. To retain dialogue contextual information during classification, we concatenated the utterance with the [CLS] token of the dialogue. We experimented with two approaches: processing the utterance independently from the dialogue and extracting it from the dialogue's last hidden state. Our findings show that ELECTRA achieved a notable F1 macro score of 0.8182 for emotion classification, outperforming BERT. However, trigger utterance classification remains challenging, with a maximum F1 score of 0.4581.

1 Introduction

The task of this project follows the EDiReF shared task competition where is required to assign an emotion to each utterance from a fixed set of emotion (ERC) and additionally identify the trigger utterance(s) that causes an emotion-flip in the conversation dialogue (EFR). Emotions are an important part of the human beings often defined as an individual's mental state associated with thoughts, feelings and behaviour. Natural language is often indicative of individual's emotion and NLP joined the research over this topic in the last years thanks to the availability of new dataset collected purposefully from social media and other web recordings such as "Emotionlines" (Chen et al., 2018). Solving emotion recognition have a wide scenarios of application such as opinion mining, recommender systems, health-care, legal trials and so on. Although several approaches have been considered, the problem is intrinsically difficult to master due to lack of contextual information about the individuals, and, as addressed in some part by this task, about dialogue temporality. Known approaches to address this task can be subdivided in three main groups according to (Shaheen et al., 2014): keyword-based detection, learning-based detection and hybrid detection. Keyword-based detection approaches rely on ruled expert knowledge and on the creation of lexical representation of emotion in which keyword are identified and used to classify the whole sentence or part of it. This approaches lack of contextual information from which identical keyword can express different emotions. Learning-based approaches are more effective, depending on the presence of a dataset as a learning base. Early works on learning-based methods used standard machine learning techniques, followed by CNNs and LSTM models. Nowadays, state of the art techniques are based on transformers models, capable of learning contextual information and interdependence between words. BERT (Devlin et al., 2018) is one the most known models to address NLP tasks and this project tries to apply it to emotion recognition comparing its performances against another transformer-based technique called ELECTRA (Clark et al., 2020), which learn better contextual representation from a different pre-training task. Hybrid techniques try to mix the two approaches and theoretically gain the advantages of both. In practice these techniques are still not competitive enough with respect to transformer-based techniques, thus they won't be explored in this project. As required from project assignment, the experiments were run over the EDiReF English dataset of the challenge with BERT model as a baseline in two different setups: freeze and unfreeze. Both of the setups are followed by a linear classifier to evaluate the emotion and trigger classes of the utterances. In order to evaluate the capability of BERT to extract contextual information we fed to the model the whole precedent dialogue in

single utterances classification. We repeated the experiment with ELECTRA as a base model and evaluate the results. We used the average F1 between emotion and triggers as the criterion for the best performing model. Among the different experimented model, ELECTRA with an average F1 of 0.638, scores 0.8182 F1 macro score in the classification of utterances emotions which present a good recognition of emotions in the sentences. We found that the classification of trigger utterances is an hard problem for all our studied models, with ELECTRA scoring 0.4581 and the best one (*BERT nopool*) 0.481. We also found no significant statistical difference in the correct emotion classification of an utterance which is previously followed by an emotion trigger utterance, as the challenge was trying to investigate.

2 Background

We used two base models for our architectures, BERT and ELECTRA. The latter one was chosen to test the impact of a different pretraining approach on this task. The core innovation of ELECTRA lies in its pretraining task: instead of masking random tokens and predicting them, ELECTRA trains a discriminator to distinguish between real tokens and corrupted tokens generated by a separate generator model. Although it may seem similar to the training setting used in GAN the generator is trained with maximum likelihood meaning that its objective is not to fool the discriminator. The generator is typically a smaller language model that predicts tokens in masked positions, so it is tasked with replacing masked tokens with plausible alternatives, creating a set of corrupted text sequences. The discriminator instead is trained to identify which tokens in the input text have been replaced by the generator. This model processes the entire input sequence and outputs a binary classification for each token, indicating whether it is original or replaced. In practice, the generator is first pre-trained alone and then fixed while training the discriminator. This iterative process ensures that the generator produces high-quality token replacements, which in turn improves the discriminator's performance. This approach results in better performances in downstream tasks and it can achieve good results also with less computational resources.

3 System description

When classifying an utterance it is reasonable and necessary to take into consideration the whole context of the dialogue it comes from. To accomplish this we devised three different approaches. All models take the utterance to be categorized and the dialogue they come from as input. Our solutions were inspired from the idea of (Shivani Kumar, 2021) to choose a target utterance and classify it with emotion and trigger, but instead of concatenating the encoding of each utterance with the target utterance we decided to encode the whole dialogue and after that the target utterance. The result are combined in 3 different ways and these are the main differences between the models. A dropout layer is present between the backbone model and the classifier to regularize the training. The models are built using the HuggingFace's transformers module, expanded when necessary for custom behaviour. All the models have two separated fully connected layers which are the classification heads for emotions and triggers. The models use either BERT or ELECTRA for the backbone. For easier references we have named the models **concat**, **nopool**, and **extraction**.

- The first model (named **concat**) takes the **pooler_output** from the encoding of the dialogue and concatenates it to the encoding of the target utterance, so the dimension of the channel becomes the length of the utterance plus 1. This means that two forward pass through the backbone model are performed, one for the entire dialogue and one for the target utterance. In the visualization of the model Fig 1 we can see that the result of the concatenation is passed first to the dropout layer and then to the classification heads. To obtain the final output an average over the second channel is performed. As it is explained in the documentation of BaseModelOutputWithPooling in HuggingFace library the **pooler_output** is "last layer hidden-state of the first token of the sequence (classification token) after further processing through the layers used for the auxiliary pre-training task. E.g. for BERT-family of models, this returns the classification token after processing through a linear layer and a tanh activation function. The linear layer weights are trained from the next sentence prediction (classification) objective during pre-training".

- The second model (named **nopool**) works as the first one with the slight difference that it is not taking the **pooler_output** from the backbone output but the raw first token from the last hidden layer, which is the [CLS] token that should work as a comprehensive representation of the whole dialogue. A visualization of the model can be seen in 1.

- For the third model (named **extraction**) we used a different approach. While in the first two models the idea was to process the utterance independently from the dialogue, here instead we perform a single forward pass of the entire dialogue. The target utterance's tokens are then extracted from the last hidden state and concatenated whit the [CLS] token embedding. In Fig 2 we can see that they are padded to the length of the longest utterance in the batch and concatenated to the classification token of the dialogue encoding, the remaining part of the pipeline is identical to the other two models.

A uniform classifier and a majority classifier are used as a baseline to compare our models.

4 Data

The dataset used for the experiments are data picked from the official challenge web page and consist in 4000 set of utterances divided with an 80/10/10 split, resulting in 3200 for the training set, 400 for the validation set and 400 for the test set). Each utterance row is paired with the context of the previous dialogue and with the emotion labels for all the utterances of the dialogue. Additionally the label of the emotion trigger flip for all the utterances of the dialogue are given too. We first augmented the dataset by splitting each dialogue in a set of new utterances and reconstructed the dialogue till the new point of the conversation. In this way we can focus only on the classification of a single utterance at a time still retaining context information. To reconstruct the dialogue we concatenated all the utterances with the separation token of the corresponding tokenizer. The resulting dataset is composed by 28062 rows for the training set, 3437 rows for the validation set and 3501 for the test. Analysis on the new dataset present some class imbalance: the 'neutral' emotion class having a frequency at least double respect all other emotions and with 'disgust' class being the least represented. We found imbalance in the trigger task classification too, where positive case are far less frequent than negative ones. A combined representation is depicted in Fig. 3. As the last step of pre-processing, each utterance and dialogue was tokenized, using the according tokenizer for each model. Additional information are linked to each row in order to better analyze data at the end, such as indexes to reconstruct dialogues.

5 Experimental setup and results

Our experiment was conducted as follow: for each architecture and architecture configuration, we run the training five times, each with a different initialization seeds and collected results for the best performing one. Additionally we also trained the "*extraction*" model with the BERT backbone frozen, in the end a total of 25 models were trained. The optimizer used is AdamW with learning rate $2e-5$ and weight decay equal to 0.01. The scheduler is a standard cosine scheduler. The problem at hand can be seen as a multi-task learning process where a single model has to concurrently learn two different, although correlated, tasks. We decided to treat it as a simple parallel classification task where the losses for the emotions and triggers classification objectives are simply summed with equal weights. For the emotion classification tasks we used the standard cross entropy loss while for the trigger one the binary cross entropy, in both cases a tensor with weights to address the class imbalance was given to the loss functions. These weights were computed with the *compute_class_weight()* function of the scikit-learn library, in our preliminary experiments these weights did not influence the results significantly but we decided to keep them given that we didn't see a negative impact. Most of the training was performed over an RTX 3070 with 8GB of VRAM and each model took around 3 hours to train the five epochs. Table 1 summarize the score metrics for the best performing model between the 5 seeds for each configuration, while Table 2 shows the F1 computed on the dialogues.

6 Discussion

Our results showed that our implemented architectures have a good capability in predicting emotion and learn contextual information from dialogues, scoring 0.8182 F1 macro score for our best performing model while others still stay above the 0.7 line of F1 score. On the contrary, our architectures

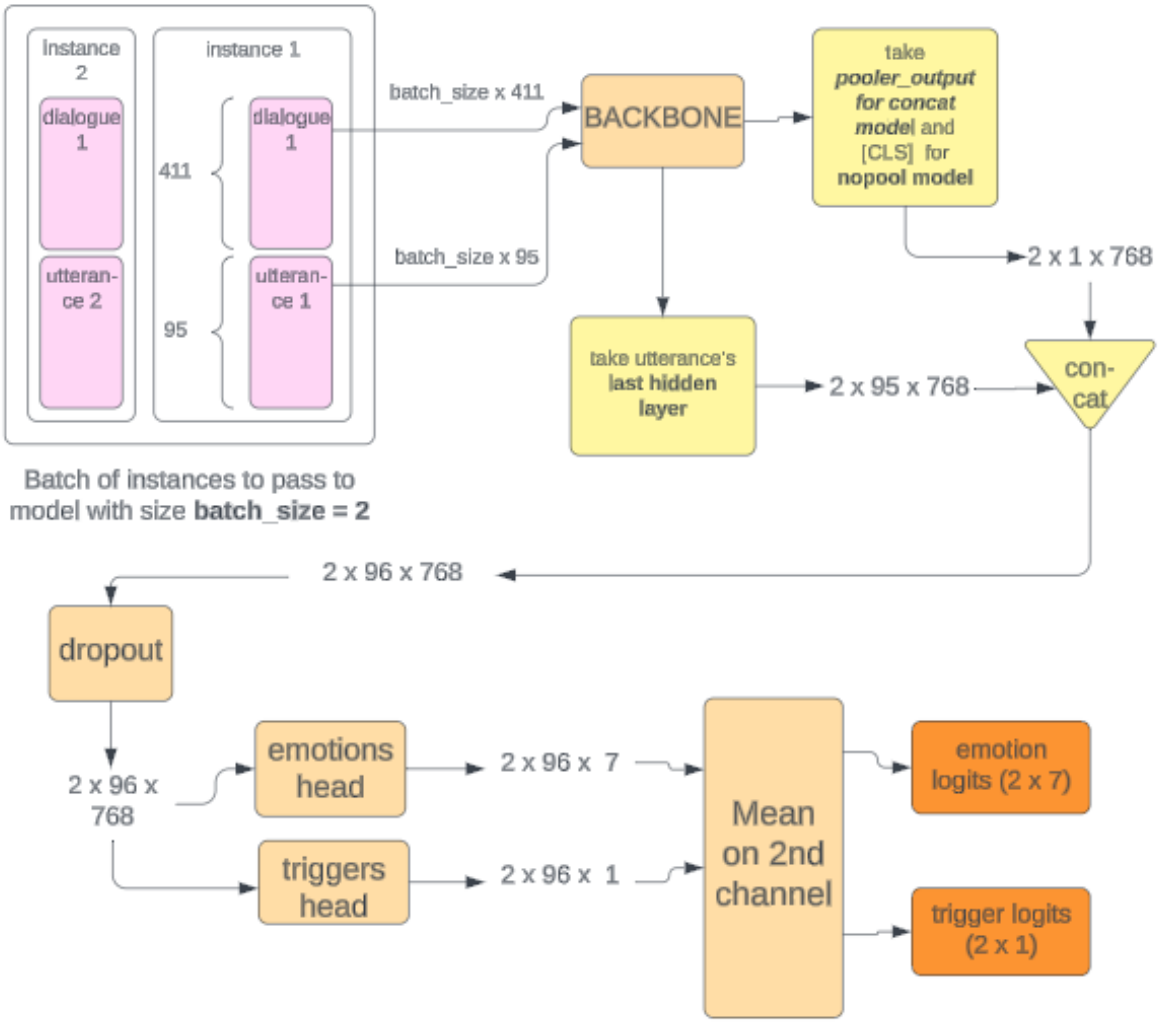

Figure 1: Concat and nopool models architecture with batch of 2.

Model <seed>	Emotion		Trigger		Avg.	
	Acc.	F1	Acc.	F1	Acc.	F1
BERT concat 51	0.860040	0.838365	0.811482	0.185185	0.835761	0.511775
BERT nopool 666	**0.868609**	**0.853210**	0.819766	0.200253	**0.844187**	0.526732
BERT extract freeze 77	0.741217	0.717597	0.815196	**0.481986**	0.778206	0.599791
BERT extract 666	0.815481	0.792455	0.824907	0.478298	0.820194	0.635376
ELECTRA 51	0.836047	0.818277	0.820908	0.458081	0.828478	**0.638179**
random	0.427306	0.085537	**0.840046**	0.000000	0.633676	0.042769
majority	0.145387	0.121880	0.501857	0.239092	0.323622	0.180486

Table 1: Results comparison between the different models, metrics are computed on all the utterances

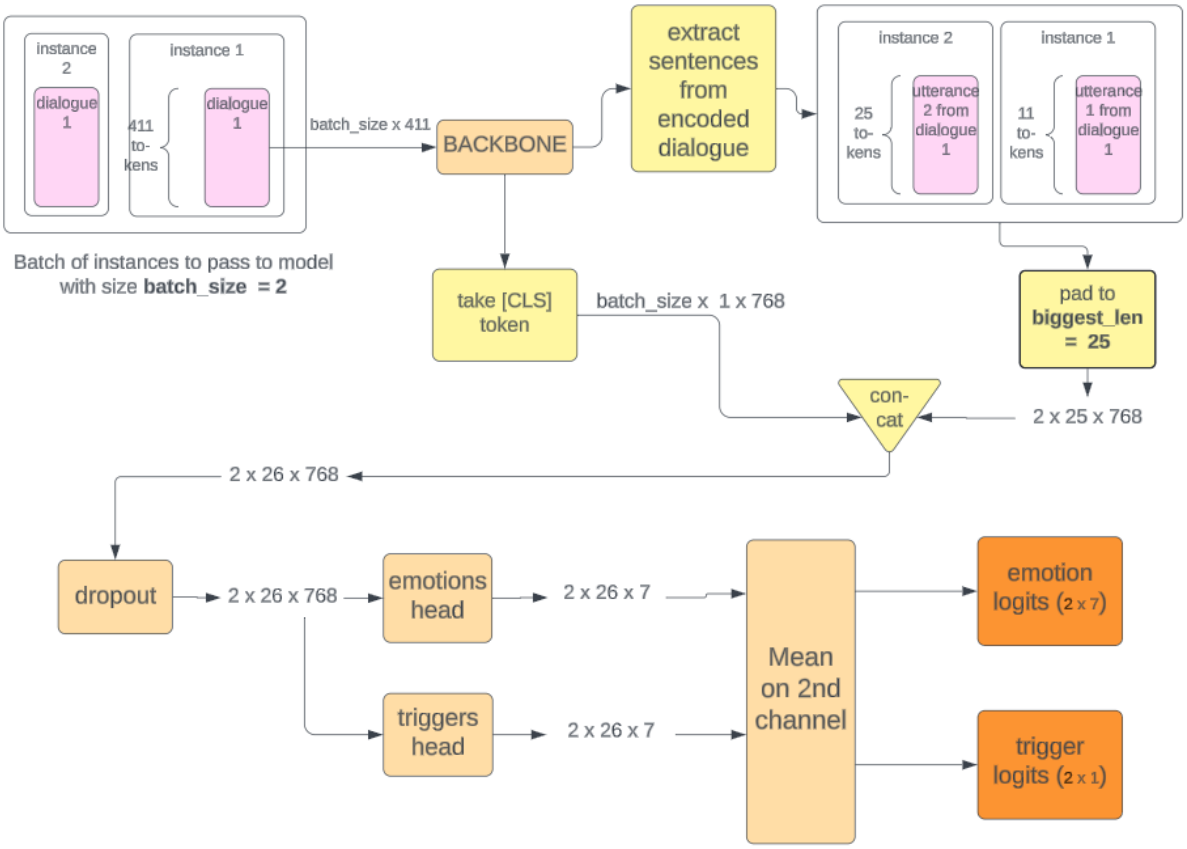

Figure 2: Extraction model architecture with batch of 2 and longest utterance of 25 tokens.

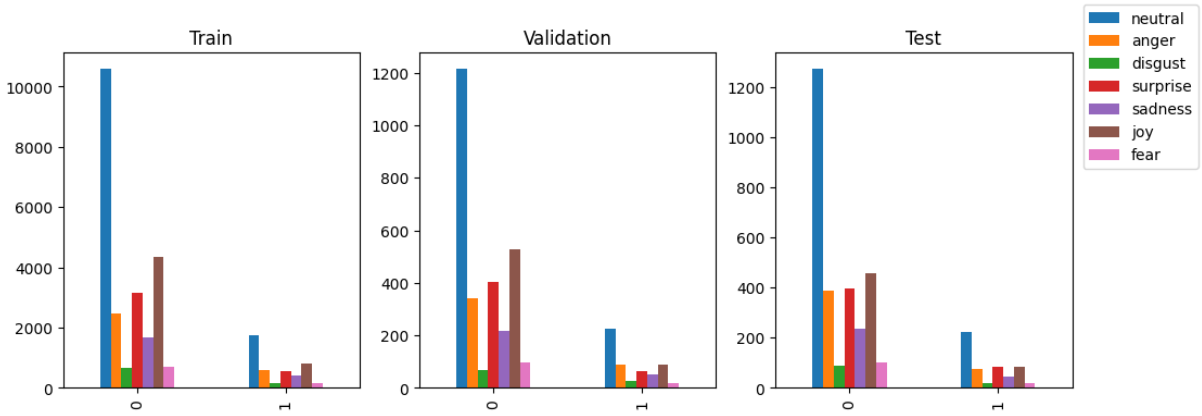

Figure 3: Combined distribution for emotion classes and trigger utterances

Model <seed>	Emotion F1	Trigger F1	Avg. F1
BERT concat 51	0.8498	0.7715	0.81065
BERT nopool 666	**0.8589**	0.7742	**0.81655**
BERT extract freeze 77	0.729	0.7763	0.75265
BERT extract 666	0.8085	0.7841	0.7963
ELECTRA 51	0.8268	**0.7854**	0.8061

Table 2: Average F1 for each dialogue

Model <seed>	Emotion Acc.		Support	
	Correct t	Wrong t	Correct t	Wrong t
BERT concat 51	0.619048	0.744304	54.0	506.0
BERT nopool 666	0.833333	0.753117	52.0	508.0
BERT extract freeze 77	0.643411	0.636364	224.0	336.0
BERT extract 666	0.675214	0.725000	205.0	355.0
ELECTRA 51	0.666667	0.715170	201.0	359.0

Table 3: 'Trigger effect' for best models. This table show the effect that a correct or incorrect trigger classification of ground truth positive trigger utterances have over the next utterance emotion classification accuracy

Utterance id	Dialogue	emotion	(GT)	trigger	(GT)
utterance 3881	Hey, y'know what a really good rainy day game is?	neutral		0	
	What?!	surprise	(neutral)	0	
	I mean naked game. Strip poker, we should totally play strip poker.	joy		0	
	No, no!	neutral	(disgust)	0	
	What are you crazy?!	anger	(disgust)	0	
	Come on! When you go away, you-you have to play, it's like a law!	neutral	(sadness)	0	
	Allll done!	anger	(neutral)	0	
	Aww, thank you.	neutral	(joy)	0	
	Okay, who's next?!	neutral		0	
	No-o-o! No way!	surprise	(disgust)	0	
	Come on, please?! I'm boredddd! You let me do it once before.	anger	(neutral)	1	(0)
	Yeah well, if ah, if that's the rule this weekend... No!	anger	(joy)	1	
utterance 1408	Really Mr. Geller, you don't have to do this.	anger	(neutral)	0	
	Oh come on! Here we go! Stand by for mission countdown!	anger	(joy)	1	(0)
	I'm an alien. I'm an alien.	sadness	(neutral)	0	
	Oh no! An asteroid!	anger	(surprise)	0	
utterance 2214	Okay, look, I-I know what you guys are going to say	fear	(neutral)	1	(0)
	You two will have very hairy children.	sadness	(disgust)	1	
	Okay, I didn't know you would say that.	anger	(surprise)	1	(0)
utterance 1445	Well, that's not something a girl wants to hear.	neutral	(anger)	0	
	No, come on don't start. Ouch!	anger	(joy)	0	
	What?	neutral	(sadness)	0	
	Stupid balls are in the way.	surprise	(sadness)	0	
utterance 1515	Oh good, okay, I can't take it anymore.	neutral		0	
	I can't take it anymore.	fear	(anger)	0	
	So you win, okay?	joy	(anger)	0	
	Here!	surprise	(anger)	0	
	Pheebs?	joy	(anger)	0	
	Flying a jet?	neutral	(surprise)	1	

Table 4: Examples of misclassified dialogues from the test set

perform not so good as trigger utterance classifiers, scoring only 0.4581 of F1 score with some configurations scoring lower than the majority classifier baseline. Results from further analysis lead us to make the hypothesis that trigger utterances are not useful in the dialogue classification task. First, the labeled data are very noisy, with a great unbalance between positive and negative cases. Just by looking at the data, the labels seemed hard to understand from a human prospective: emotions along the dialogue change way more frequently with respect the hypothetical effect of trigger utterances. Our hypothesis is then supported by results. Comparing the performance of "BERT extract freeze 77" and "BERT concat 51" models, we can see that a lower performance in classifying trigger utterances does not lead to a lower performance in the emotion classification task. Instead it is counter intuitive that our best performing model on the trigger task is also the worst in terms of emotion classification. This is probably explainable considering that the loss is split in learning the two tasks jointly. In fact if the trigger classification task does not help the emotion task, it is just wasted effort. A further analysis on the effect of the trigger utterance is reported in Table 3, where the "Trigger effect" is studied. From the challenge directives, trigger utterances were investigated because considered useful as marks along the conversation in which emotion abruptly change. We investigated the effect of our models over the utterances following a ground truth trigger utterance. In theory, a correct classification of the presence of trigger utterance should lead our model to more easily classify the follow utterance by having an additional information, while not recognizing the presence of the trigger should make the classification harder. Results in the tables showed no appreciable difference in the accuracy of the classification task, thus our model either didn't learn any useful representation of trigger utterances or the information given by the trigger utterance is not enough to make a difference. We studied the effect of the three different configuration of processing the language implicit representation learned by the transformer model, presented in Section 3. From a theoretical point of view, the "extraction" configuration using the raw hidden state of the transformer model, should present a stronger capability of learning context and discriminate difficult utterances based on dialogue. As we can see in the results, all the three extraction models presented (BERT extract freeze 77, BERT extract 666 and ELECTRA 51) have a significant higher score over the trigger classification task which seems a lot more context dependant. They perform well in the emotion classification task, but surprisingly, the concat and nopool configuration have a very high score in the emotion classification task. Our hypothesis is that because the configuration rely both on single utterance removed from context and whole dialogue representation from BERT, the learned implicit representation already learned by BERT base model is so strong that could perform better on the tasks without considering too much context, resulting in a lower trigger score. Analyzing instead the effect of the frozen vs unfrozen model, we can see that the unfrozen models perform better in terms of emotion classification, while remaining comparable in the trigger classification task. To conclude our discussion we dove deeper into the analysis of our best performing model, assessing the robustness of its performance. The model showed good capability to recognize all the emotions, even the underrepresented ones as showed in the confusion matrix depicted in Fig 5. We noticed that the majority of misclassified emotions were wrongly classified as neutral, as expected as it is the most represented one in the dataset. We then studied the effects of the dialogue length over the classification task and we found that small dialogue (2, 3 or 4 utterances long) were harder to classify with respect longer ones with an average of correct dialogue classification $10 - 12\%$ less than average, as depicted in figure 4. In Table 4 we reported the dialogues in which our models scores less. The emotion classification tasks for the reported cases were found hard even for us, human being. In our models we addressed the tasks as two parallel classification tasks, trying to study the effect of emotion classification combined with the detecting of trigger utterances in the context dialogue. Further works could exploit a different approach to the same task by first learn a classifier for trigger utterances and then connect it to the emotion classifier task in a two stage training pipeline. We addressed only transformer-based architectures as the major state of the art techniques, but other architectures such as RNN could be exploited for the task with good results. A subsequent work could study the effect of the different "extraction" weights from transformer based techniques.

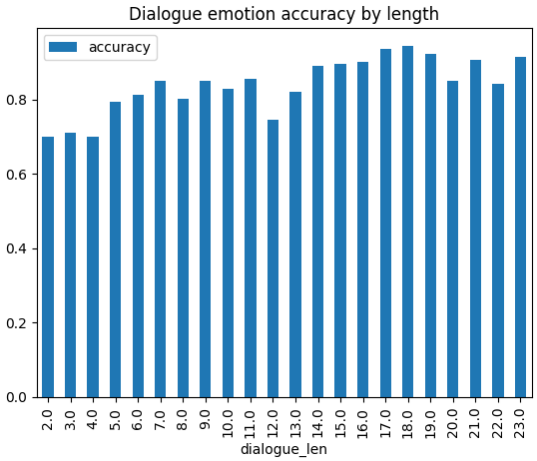

Figure 4: Average percentage of dialogue correctly classified across different dialogue length

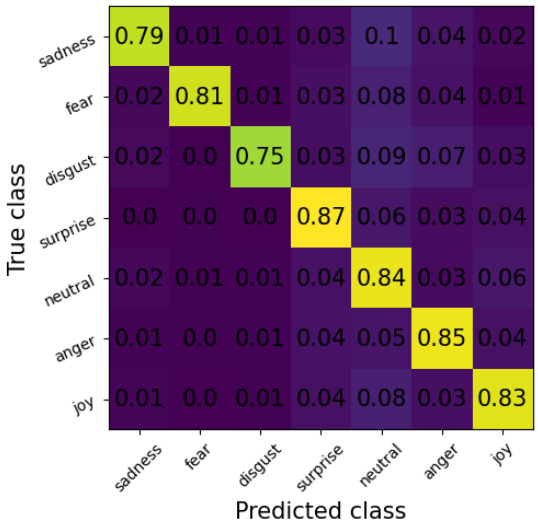

Figure 5: Confusion matrix of the ELECTRA 51 model for the test set, standardized with respect the ground truth labels

7 Conclusion

In this study, we tackled the complex task of Emotion Recognition in Conversations (ERC) and Emotion Flip Reasoning (EFR) by leveraging transformer-based models, specifically BERT and ELECTRA. Our goal was to classify the emotions of utterances within dialogues and identify the trigger utterances responsible for emotion shifts. We implemented three distinct architectures to process the outputs from BERT and ELECTRA: pooling, concatenation, and extraction. Our findings indicated that the extraction method, which uses the raw hidden states from the transformer, performed well in trigger classification but did not significantly enhance emotion recognition compared to the other methods. Interestingly, models using concatenation and no pooling configurations achieved high scores in emotion classification. Our results demonstrated that ELECTRA outperformed BERT in emotion classification, achieving an F1 macro score of 0.8182. However, the classification of trigger utterances proved to be more challenging, with the best F1 score reaching only 0.4581. Further analysis revealed no significant statistical difference in emotion classification accuracy between utterances preceded by a trigger and those that were not, suggesting that trigger utterances might not provide substantial additional information for emotion classification. The noisy and imbalanced nature of the trigger data further complicates this task. Future work could explore different architectures, such as RNNs, or a two-stage training pipeline where a trigger classifier is trained before the emotion classifier. Additionally, improving the dataset quality and addressing class imbalance could lead to better performance in trigger classification. In conclusion, while our study highlights the potential of transformer-based models like ELECTRA in EFR tasks, it also underscores the need for more sophisticated methods and better data to effectively tackle ERC challenges.

8 External sources

All the code of the project can be found at `https://github.com/PanzaResce/ediref`. The repository also contains a file named *models_std_6avg.ods* with the average and standard deviation of all the metrics over the 5 seeds for every model/architecture trained.

References

Sheng-Yeh Chen, Chao-Chun Hsu, Chuan-Chun Kuo, Lun-Wei Ku, et al. 2018. Emotionlines: An emotion corpus of multi-party conversations. *arXiv preprint arXiv:1802.08379*.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. 2014. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop*, pages 383–392. IEEE.

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. *arXiv:2103.12360*.