UWBA at SemEval-2024 Task 3: Dialogue Representation and Multimodal Fusion for Emotion Cause Analysis

Josef Baloun **Jiří Martínek** **Ladislav Lenc**
New Technologies for The Information Society, University of West Bohemia, Pilsen
{balounj,jimar,llenc}@ntis.zcu.cz

Pavel Král **Matěj Zeman** **Lukáš Vlček**
Department of Computer Science and Engineering, University of West Bohemia, Pilsen
{pkral,zemanm98,vlcek0}@kiv.zcu.cz

Abstract

In this paper, we present an approach for solving SemEval-2024 Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations. The task includes two subtasks that focus on emotion-cause pair extraction using text, video, and audio modalities. Our approach is composed of encoding all modalities (MFCC and Wav2Vec for audio, 3D-CNN for video, and transformer-based models for text) and combining them in an utterance-level fusion module. The model is then optimized for link and emotion prediction simultaneously. Our approach achieved 6th place in both subtasks. The full leaderboard can be found at https://codalab.lisn.upsaclay.fr/competitions/16141#results.

1 Introduction

The *SemEval-2024 Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations* (Wang et al., 2024) is aimed at extracting emotion-cause pairs (ECPs) in conversations. The main data source to tackle this task is recordings from the sitcom *Friends* in the English language – Emotion-Cause-in-Friends dataset Wang et al. (2022).

The detection of emotions and a deeper analysis of what causes them is one of the interesting and important tasks that have recently been tackled within the NLP community. Previously, researchers focused their efforts on text-only emotion-cause extraction (Gui et al., 2018; Gao et al., 2017; Bostan et al., 2019). However, representing dialogues solely through the text (speech transcription) is not entirely adequate, as people use different intonations and other prosodic features. Moreover, the fact that something happens during the conversation (e.g. someone walks in or something breaks) affects the dialogue in terms of emotions and their causes. We can also mention different facial expressions for different emotions. So, the fact that a conversation in its natural form is multimodal (text, audio and video) opens a big space for research.

Our approach is targeted at both subtasks of the above-mentioned Semeval 2023 task. The main difference between them is the number of different modalities used for predicting ECPs. For Subtask 1, the prediction of ECPs is solely based on the text transcription without recordings. The goal is to provide text spans along with ECPs: i.e. the segment of an utterance primarily responsible for emotion-cause. Subtask 2, does not require extracted text spans (in some cases it is impossible because the emotion-cause is not expressed in the textual form), but it is required to use other modalities embedded in available `mp4` video files to extract emotion-cause pairs together with a target emotion.

Our approach uses all modalities and encodes them into a common utterance-level representation, which is then used for *link* and *emotion* prediction. The objective is to learn both prediction tasks with two loss functions that are combined. The main idea behind this is that emotions might positively influence the links (pairs) and vice versa.

Another point we would like to highlight is that the textual input is a whole dialogue, so the context is taken into consideration. This improves the link results significantly, as shown further. After validating our codes by the SemeEval task organizers, the final implementation will be released[1].

2 Task and Background

The goal of the emotion-cause pair extraction (ECPE) task (Xia and Ding, 2019) is to extract potential pairs of emotions and corresponding causes in a conversation/document and/or other source of dialogue. It is an extension of the emotion-cause extraction – ECE task (Lee et al., 2010), where the goal is to decide if a clause/utterance is the corresponding cause, given the annotation of emotions.

[1]https://github.com/martinekj/semeval_2024_Task3_ECPE

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 316–325
June 20-21, 2024 ©2024 Association for Computational Linguistics

This SemEval task does not require only the extraction of corresponding pairs but also the prediction of emotions. In other words, extracted pairs must be complemented by the prediction of the target utterance emotion. So, for the evaluation, we have a triplet (a source utterance, a target utterance, an emotion of the target utterance). E.g., **3_joy, 2** which means that the emotion *joy* in *utterance 3* is caused by *utterance 2*.

Our team participated in both competition subtasks. We aimed to extract not only the pairs (as illustrated in the previous example) but also text spans – the exact parts of utterances/clauses primarily responsible for the emotion-cause (e.g. **3_joy, 2_You made up!** – meaning that the emotion *joy* in *utterance 3* is caused by the text span *You made up!* in *utterance 2*)

Hence, we work with all input data (i.e., multimodal – text, sound, image sequence/video) of the dataset **Emotion-Cause-in-Friends** that serves as the competition dataset.

2.1 Related Work

Lee et al. (2010) presented a text-based approach for the ECPE task. They created a rule-based system and tested it on a Chinese dataset created from the Sinica corpus.

Chen et al. (2020) proposed an approach that takes the ECPE task as a unified sequence labeling task. Their method combines a convolutional network with two bi-directional long short-term memory networks. They show that the approach outperforms several baselines. However, the score is slightly lower than baselines including BERT.

Poria et al. (2018) created the multimodal MELD dataset as an extension of the EmotionLines corpus and performed a baseline evaluation of the emotion recognition task on this data. Another multimodal dataset is presented in (Firdaus et al., 2020).

Wang et al. (2022) presented a multimodal approach for the emotion-cause pair extraction. The authors created a dataset including text, audio and video modalities. The baseline approach obtained F1 score of 0.51.

3 System Overview

We decomposed the main objective into emotion and link prediction (the estimation of pairs) tasks. The final result then consists of source and target utterances provided by the link and emotion of the target utterance.

The architecture is depicted in Figure 1. First, we encode the different modalities at the utterance or dialogue level to incorporate more context. Next, we fuse the representations at the utterance level. Once we have representations of all individual utterances in a dialogue, we predict links and emotions (Subtask 2). Based on this output, we employ our separate model for text span prediction, which is necessary for Subtask 1.

We have followed the competition rules and used pre-trained language models (PLMs), but no additional training data.

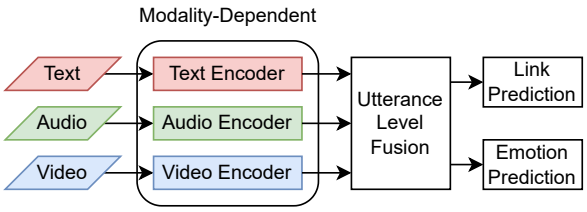

Figure 1: System architecture

3.1 Text Encoder

We employed a transformer-based encoder, such as BERT (Devlin et al., 2019), for text encoding. As depicted in Figure 2, the input consists of an entire dialogue. It commences with a CLS token and proceeds with tokens representing individual utterances. As usual, positional encoding corresponding to token position is applied. The utterances are separated by a SEP token, and the input is further extended by utterance embeddings. After encoding the tokens, we average the tokens of every single utterance to derive its representation. This way, the dialogue context is available in every single utterance.

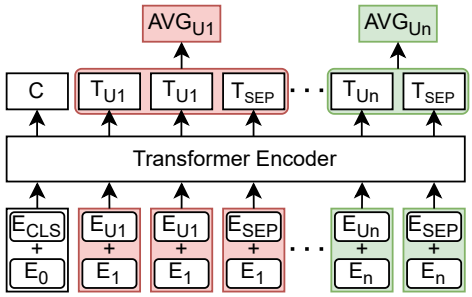

Figure 2: Text Encoder

3.2 Audio Encoder

For encoding the audio, we evaluated two methods. The first is referred to as *MFCC* feature extraction

and is based on Mel frequency cepstral coefficients (Tiwari, 2010). The second method is based on *Wav2Vec* model (Baevski et al., 2020).

3.2.1 MFCC Feature Extraction

We firstly denoise the audio files removing the background noise (background laughing track, people speaking in the café, etc.). We use the REPET-SIM (Rafii and Pardo, 2013) separation method to separate the main speaker voice line. The separated main speaker voice line audio is then used for computing the MFCC, and this audio representation is used in a long short-term memory (LSTM) model trained for the emotion recognition task. Audio feature vectors with a dimension of 2048 are acquired from the last hidden state of the model. The model comprises one bidirectional LSTM and two linear layers. The whole model is trained on the emotion recognition task.

The REPET-SIM method is a generalization of the REPET method (Rafii and Pardo, 2012). The basic idea behind the REPET method is to find repeating elements in audio, compare them with repeating models derived from them, and separate the repeating patterns through time-frequency masking. REPET-SIM specifically identifies these repeating elements by using a similarity matrix (Rafii and Pardo, 2013).

MFCC capture the shape of the power spectrum of a sound signal. They are computed by first transforming the audio into the frequency domain using the Discrete Fourier Transform and then applying the "mel" scale to approximate the human auditory perception of the sound frequency (Tiwari, 2010).

3.2.2 Wav2Vec

As an alternative, we used pre-trained version (Field, 2022) of Wav2Vec model as an audio encoder. It was fine-tuned for the emotion classification task, and subsequently, we conducted further fine-tuning on competition data. We averaged the audio sequence representations provided by the final layer resulting in a 1024-dimensional audio representation of the utterance. During the fine-tuning phase, the representation was utilized by a two-layer perceptron to predict emotion.

3.3 Video Encoder

We utilized the ResNext 3D-CNN (Hara et al., 2018) model with depth 101 pre-trained on the Kinetics (Carreira and Zisserman, 2017) dataset. For every 16 frames, this model provides an output vector with a dimension of 2048. In the case of longer videos, the final feature vector is computed with a global average pooling over the temporal dimension.

The input to the model are preprocessed image frames from the video file (using the *ffmpeg* python library). The preprocessing consists of scaling to 240x240 pixels (while preserving the aspect ratio with zero padding).

3.4 Fusion Module

The multimodal fusion relies on a transformer-based encoder. Since the dimensions of text, audio, and video representations may vary, they undergo linear projection using a linear layer with the fusion size (*fs*) as a parameter. Subsequently, they serve as tokens for the encoder input. The encoder consists of 6 layers with $fs/64$ heads and GELU activation function. The intermediate size is $4 \cdot fs$.

The fusion is done on the utterance level, so no explicit dialogue context is available, but it may be provided by encoded representations of individual modalities. For the fusion, we ignore the positional encoding.

We consider several fusion strategies. The first straightforward scenario is to use the aggregation function for each component of the encoded tokens: *AvgFusion* (averaging the representations); *MaxFusion*, *MinFusion* (taking the maximum/minimum activation across modalities).

Further, we incorporate an additional learnable fusion token (FT) that is added to the input. It is used to aggregate information for the utterance in a similar way as the CLS token is used in BERT, for example. They are labeled as: *SingleFT* (a single FT that is used as a result after encoding); *MainSpeakerFT* (a different FT for each main speaker and one FT for other speakers; *AllSpeakerFT* (incorporating FT for each speaker).

Other possible fusion strategies exist (e.g. Nagrani et al. (2021)), and incorporating some of them is our potential future work.

3.5 Emotion Prediction

Emotions are predicted directly from the fused utterance representations using a two-layer perceptron with LeakyReLU activation function. The hidden size matches the fusion size.

3.6 Link Prediction

The link prediction module is also inspired by the transformer-based encoder. The fused utterance

representations are used as the input tokens. However, our focus shifts from encoded tokens to attention matrices in this context. These matrices are processed and utilized as the adjacency matrix of the graph. Positional encoding is optional since it may be included in the utterance embedding provided by the text, audio, or video encoder.

We used six transformer-based encoder layers, followed by a specialized layer that computes the attention matrices for each head. Subsequently, these matrices are aggregated across all heads using *average*, *maximum*, or *minimum* activation. In contrast to the transformer-based encoder layer, the specialized layer also does not normalize the attention scores, as they are directly provided as logits for the links.

3.7 Text Span Analysis

This section covers the approach used for *Subtask 1*: the prediction of text spans responsible for the cause of emotion. A baseline approach with which we compare is using the entire utterance as a text span. As expected, this trivial approach results in quite a low *strict match F1* metric. The main evaluation metric for this subtask, though, is the *Proportional F1* (which considers the overlap proportion of the predicted span and the annotated one). If we uploaded the text-spans result based on this trivial approach, the resulting *Proportional F1* value is around 20% based on test data provided for the competition evaluation.

Based on this result, we can state that a significant part of training data has no specific text span that causes emotions. This might indicate that even for human annotators, it is not easy to determine a particular text span in a significant number of utterances, and he/she labeled the whole utterance.

According to the training data, we specified five text span categories and created a classifier for their prediction based solely on individual utterances with no context. Furthermore, we have defined a set of regular expressions whose goal is to automatically detect individual categories.

The most common label is `Whole Utterance`, as we declared above. Figure 3 shows all categories of text spans resulting from regular expressions. Regular expressions are used to split an utterance by the punctuation marks `(',', ';', '.', '!', '?')` and compare their results with annotated text spans.

The `First part` label corresponds to the beginning of an utterance until the first punctuation mark. In a similar way, the `Last part` category is taken, except that it is taken from the end. The `Middle part` label is the part in the middle (an utterance part without the beginning and end). It appears usually in cases where an utterance is long. For cases when all regular expressions fail (a text span is neither the whole utterance nor the first, last, or middle part), we created a category `Other`. As illustrated in Figure 3, this category is quite common in training data.

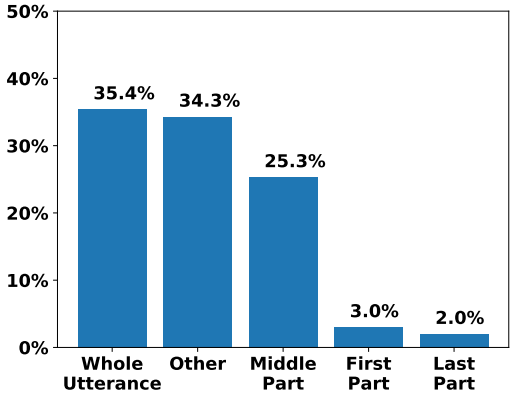

Figure 3: Distribution of the specified categories of text spans

3.7.1 Text Span Classifier

Our first intention was to use a transformer-based model and carry out the question-answering training scenario that aims at providing an "answer" (text span which causes an emotion) identified using the start span and end span generated by the model (similarly to the BERT SQuAD model Devlin et al. (2018)). All our efforts for training such a model, though, failed, due to the lack of training data. The competition rules forbid the usage of another annotated private/public data to fine-tune a model, so we decided to use another model with a much simpler learning objective.

The input text comprises tokens of the current utterance (no context is considered in this case, i.e. no information about previous/subsequent utterances in a dialogue) and is fed into a transformer model as usual with a prediction head with five output neurons.

Once the class (text span category) is predicted from the CLS token, we apply a regular expression (assigned to the predicted class) to extract a substring from the utterance and, in the sequel, the start and end index of this substring.

During the prediction phase, the class `Other` is

taken as `Whole Utterance` label since there is no regular expression associated with this class. We remind that the text span extraction task is solely text-based. The results are presented in Table 4.

4 Experimental Set-up

In the data provided, there is a mapping of video names on train/dev/test splits. According to this information, we dedicated 9,966 utterances (1,373 dialogues) for training, and the remaining dialogues/utterances have been used as development (validation) data. Our preliminary experiments, as well as the experiments resulting in the final system, have been evaluated on this development part of the dataset.

Due to the memory limitations, we fine-tuned audio and video encoders separately. After that, the whole pipeline is trained End-to-End with frozen audio/video encoders and, therefore, constant audio/video representations. We made this choice based on the preliminary experiments, where we obtained the best results with textual modality so we take audio/video features as an auxiliary input.

If not stated differently, we used the AdamW optimizer with a learning rate of 5e-5. Categorical and binary cross-entropy loss was used for emotion and link prediction, respectively. We started with 2 "warmup" epochs with frozen encoders to limit "forgetting" of the pre-trained knowledge. Further, we continued with 50+ epochs until convergence.

To increase the importance of positive links, the positive link weight is set to 5. The fusion size is set to 1024 or 1536. The batch size ranges from 2 to 24. Due to memory limitations, we adapted gradient accumulation technique. The number of samples used for weight update is 8, 12, or 24. These hyper-parameters are further studied in Appendix A. In Tables 1, 2, and 3, we report the combination of these hyper-parameters that obtained the best result among runs.

4.1 Text Encoder

We employed several Pre-trained Language Models (PLMs) and corresponding configurations (see e.g. Table 3). The learning rate for the text encoder was lowered to 1e-7 to further limit the "forgetting" of the PLM.

As described in Section 3, the input is the whole dialogue text to provide context. We set the maximum input length to 450 tokens and the maximum number of utterances to 26, according to the training part of the dataset. That condition is not met in one dialogue in the test part. In that case, the predictions are done in a sliding window manner, so it is impossible to predict the link between the first and last utterance, for example.

For the comparison, we encoded the utterances separately with no dialogue context. This scenario is depicted as ✂ in *Text* column of Tables 1, 2, and 3.

4.2 Text Spans Classifier

All our models have been trained for 30 epochs, with learning rate=1e-05, AdamW optimizer and cross-entropy loss. We picked the best model (the best epoch), based on the validation accuracy. Results for various models are presented in Table 4

4.3 Evaluation Metrics

As stated in the Semeval task information web page[2], the evaluation is based on F1 scores with the help of which we can evaluate the emotion-cause pairs of each emotion category separately and further calculate a weighted average of F1 scores (*wF1*) across the six emotion categories (*Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise*). It is the main evaluation metric for Subtask 2.

Besides the official Semeval metrics provided by the organizers, we have also employed other metrics such as *accuracy* and *macro F1* score for emotion classification task. The *jaccard index* was used for link prediction, since the adjacency matrix is sparse and, therefore, we are not very interested in true negatives.

For Subtask 1 which involves the textual cause span, two strategies are adopted to determine whether the span is extracted correctly: *strict match* (the predicted span should be exactly the same as the annotated span) and *proportional match* (considering the overlap proportion between the predicted span and the annotated one). Although at the beginning of the competition, the main evaluation metric had been *strict match*, later *proportional match* was chosen instead due to the poor results of *strict match* based on trial data published by the organizers. The main reason behind this is that it is challenging to determine the precise boundaries of cause spans.

[2] https://nustm.github.io/SemEval-2024_ECAC/

Text PLM	Text	Audio	Video	Fusion	Multi-task	Acc.	Macro F1
j-hartmann/emotion-english-roberta-large	✓	✗	✗	–	✗	0.859	0.511
bert-large-cased	✓	✗	✗	–	✗	0.859	0.509
bert-base-cased	✂	MFCC	✓	SingleFT	✓	0.852	0.504
bert-base-cased	✓	✗	✗	–	✗	0.857	0.501
bert-large-cased	✓	MFCC	✓	MainSpeakerFT	✓	0.845	0.499
j-hartmann/emotion-english-roberta-large	✓	Wav2Vec	✗	MainSpeakerFT	✗	0.856	0.498
dbmdz/bert-large-cased-finetuned-conll03-english	✓	MFCC	✓	MainSpeakerFT	✓	0.847	0.470
j-hartmann/emotion-english-distilroberta-base	✓	MFCC	✓	MainSpeakerFT	✓	0.837	0.464
–	✗	Wav2Vec	✗	–	✗	0.533	0.397
–	✗	MFCC	✓	SingleFT	✗	0.789	0.296

Table 1: Comparison of emotion prediction methods employing various modalities, fusion scenarios, and combined multi-task training. ✂ denotes separately encoded utterance text.

Text PLM	Text	Audio	Video	Fusion	Multi-task	Jaccard
bert-base-cased	✓	MFCC	✓	MinFusion	✓	0.359
bert-base-cased	✓	Wav2Vec	✗	MaxFusion	✓	0.359
bert-base-cased	✓	✗	✗	–	✓	0.346
bert-large-cased	✓	MFCC	✓	MinFusion	✓	0.342
dbmdz/bert-large-cased-finetuned-conll03-english	✓	MFCC	✓	MainSpeakerFT	✓	0.337
bert-large-cased	✓	MFCC	✓	MainSpeakerFT	✗	0.336
j-hartmann/emotion-english-roberta-large	✓	MFCC	✓	MinFusion	✓	0.331
j-hartmann/emotion-english-distilroberta-base	✓	MFCC	✓	MainSpeakerFT	✓	0.320
bert-base-cased	✂	MFCC	✓	SingleFT	✓	0.279
–	✗	MFCC	✓	SingleFT	✓	0.076

Table 2: Comparison of link prediction methods employing various modalities, fusion scenarios, and combined multi-task training. ✂ denotes separately encoded utterance text.

5 Results

In this section, we present and analyse the results from multiple perspectives.

5.1 Emotion Detection

According to the results presented in Table 1, text plays a crucial role in the emotion prediction task. The context (whole dialogue) is not essential for emotion predictions, as the results are not significantly influenced positively or negatively. We obtained very similar results regardless of whether we used utterances separately (denoted by the symbol ✂) or not.

The best model for the emotion prediction task does not include audio/video features, leading us to conclude that they are not essential for the emotion detection task. The multimodal results suggest that the most effective fusion strategy involves the utilization of the fusion token (*SingleFT* or *MainSpeakerFT*). Other fusion strategies generally yield inferior results.

To support our emotion detection results, we have created a confusion matrix for further error analysis (see Figure 4). The first column (predicted label: `neutral`) indicates that the model has tendencies to predict *neutral* label more often, probably due to the fact that this label is most common in training data. The most challenging emotions are *fear* and *disgust* (see the third and last rows).

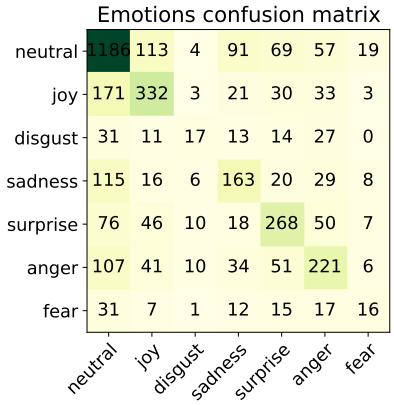

Figure 4: Confusion matrix for the emotion prediction using the submitted model – true labels are at y-axis, predicted labels are at x-axis

Since the overall score is calculated for the predicted triplet: a cause, a target utterance, and an emotion of the target utterance, we estimate the effect of emotion detection accuracy for the ECPE task by comparing the final results with the results of emotions loaded from ground truth (GT). The weighted F1 score on the dev dataset increases from

Text PLM	Text	Audio	Video	Fusion	wF1
bert-base-cased	✓	✗	✗	–	0.320
bert-base-cased	✓	MFCC	✓	MaxFusion	0.318
bert-base-cased	✓	Wav2Vec	✗	MinFusion	0.313
bert-base-cased	✓	MFCC	✓	MinFusion	0.311
bert-large-cased	✓	MFCC	✓	SingleFT	0.310
j-hartmann/emotion-english-roberta-large	✓	MFCC	✓	MinFusion	0.294
dbmdz/bert-large-cased-finetuned-conll03-english	✓	MFCC	✓	MainSpeakerFT	0.289
j-hartmann/emotion-english-distilroberta-base	✓	MFCC	✓	MainSpeakerFT	0.278
bert-base-cased	✂	MFCC	✓	SingleFT	0.262
–	✗	MFCC	✓	SingleFT	0.028

Table 3: Comparison of models trained in multi-task scenario for SemEval task. ✂ stands for separately encoded utterance text.

0.331 to 0.602 if the emotions are correct. Such results suggest that improving the emotion detection model (and increasing the precision and recall) will cause a significantly better overall ECPE score.

5.2 Link Prediction

Our findings demonstrate that injecting emotion information through multi-task training is advantageous for link prediction. The top-performing Jaccard index of 0.359 was attained with multi-task training, while without it, the highest results reached 0.336, marking an improvement of 2.3%.

Results from Table 2 clearly show that the text modality is crucial, as well as the context of the whole dialogue. Using other modalities is helpful in this case.

We have an interesting observation that is contrary to the emotion detection task. In the case of link prediction, it seems beneficial to use fusion strategies based on the aggregation function instead of the fusion token (FT).

5.3 Emotion2Emotion Link Analysis

Our next analysis should shed light on how the model behaves when linking source and target emotions regardless of the utterance texts. We created two matrices (see Figure 5) as follows. We gradually loop through all ground truth and predicted emotion-cause pairs and calculated emotion2emotion pair counts.

Both matrices are very similar (except for the first column, which is automatically zeroed as a part of postprocessing, because ECPE containing *neutral* emotion is irrelevant[3]). It shows that the information about emotions is important for the link prediction task. We supported this observation experimentally, incorporating emotion injection in multi-task training.

The high values in both matrices appear on the diagonal. These represent cases where the source emotion (cause) matches the emotion of the target utterance. The model has evidently learned this behavior, reflecting the human annotations.

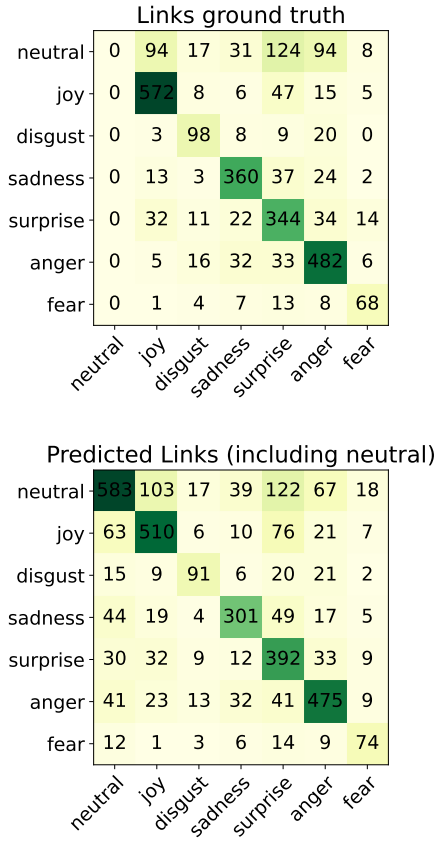

Figure 5: Links in ground truth (top) and predicted links (bottom), emotions loaded from the ground truth – dev dataset

[3]The model creates emotion-cause pairs in *neutral* emotions that should not be possible (no such pairs are present in the training dataset). We remove such pairs before we create the final prediction `json` file.

5.4 Text Span Classification

Table 4 shows the results of various models for the task of text span classification. All models have obtained *accuracy* around 74% – 77% and *Macro-averaged F1 score* above 60%.

Model	Acc.	Macro F1
bert-base-cased	74.1	62.3
bert-large-cased	76.0	60.8
conll03-bert-large-cased[4]	76.6	63.1
roberta-base	**76.8**	61.2
roberta-large	76.7	**67.6**
emotion-roberta-base[5]	76.5	62.4

Table 4: Text span classification results (in %)

For the final submission, we used the bert-large-cased model. However, after the end of the competition, we conducted further experiments with Roberta-like models (see the bottom part of Table 4), and we achieved significantly better scores, particularly in F1 macro.

5.5 Overall Results

For Subtask 2, the main evaluation metric is the weighted F1 score (wF1), as indicated in Table 3 for various models. However, our final submission for the competition is a combination of two models: the best one from emotion detection experiments (Table 1) and the best one for link prediction (Table 2). Such a combination resulted in 0.331 wF1 on dev dataset and 6th place overall in the competition[6]. This is significantly better than the best model from Table 3. All qualitative and error analyses presented above were made based on this setup since test labels are not available at the time being.

For subtask 1, our best model also achieved 6th place in the competition. Our *weighted-avg. proportional F1 score* on test data is 0.208.

Our key findings during the result analysis are as follows:

1. Basic processing of audio/video modalities has brought us only a small positive impact in the case of the link prediction task.

2. The context of the whole dialogue (processing multiple utterances) is crucial for link prediction.

3. We encountered conflicts in fusion strategies (whether to use the aggregation or the fusion token); our best model for emotion prediction is text-only with no fusion mechanism, while the best model for linking benefits from the aggregation fusion strategy.

4. We have obtained better overall results with two separate models.

5. The information about emotion is important for the link prediction task and significantly improves the results.

6 Conclusion

We have participated in two tasks in the *SemEval-2024 Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations* and obtained 6th place overall.

Our model incorporates all modalities (text, audio and video features). The main information lies within the text since the models based solely on textual modality are consistently among the best ones (see Tables 1 – 3). We proposed and implemented several strategies for the fusion of modalities at the utterance level.

To benefit more from video features, we mean that better preprocessing might be helpful (e.g., detecting a main speaker and focusing on her/his face). A possible bottleneck is in the fixed representation of the audio/video features. Optimizing them during the learning process might improve their positive impact and, subsequently, the overall task success rate. Moreover, we can benefit from the usage of a bigger model.

Our experiments have shown that the one multi-task model may not be ideal since optimal hyper-parameters differ for link prediction and emotion detection tasks. Therefore, our final submission is the composition of two models. We have provided a good starting point and a set of analyses for further research.

As a future work, one of our ideas is to use a single learning objective. In such a model, it would not be necessary to have an emotion module since everything would be managed by the link module with a multi-head self-attention matrix where each head would represent a link of one emotion. The training objective should be simpler since it uses single-label classification across components of attention matrices of individual heads. In this way, we can prevent the prediction of neutral links.

[4] dbmdz/bert-large-cased-finetuned-conll03-english

[5] j-hartmann/emotion-english-distilroberta-base

[6] The Semeval official evaluation resulted in 0.251 wF1 on test data

Acknowledgements

This work has been partly supported by the OP JAC project DigiTech no. CZ.02.01.01/00/23_021/0008402 and by the Grant No. SGS-2022-016 Advanced methods of data processing and analysis. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Laura Bostan, Evgeny Kim, and Roman Klinger. 2019. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *arXiv preprint arXiv:1912.03184*.

Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308.

Xinhong Chen, Qing Li, and Jianping Wang. 2020. A unified sequence labeling model for emotion cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 208–218.

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*.

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.

Rob Field. 2022. Speech emotion recognition by fine-tuning wav2vec 2.0. https://huggingface.co/r-f/wav2vec-english-speech-emotion-recognition.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th international conference on computational linguistics*, pages 4441–4453.

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Lin Gui, Yulan He, Kam-Fai Wong, and Qin Lu. 2017. Overview of ntcir-13 eca task. In *NTCIR*.

Lin Gui, Ruifeng Xu, Dongyin Wu, Qin Lu, and Yu Zhou. 2018. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 145–160. World Scientific.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6546–6555.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 45–53.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Zafar Rafii and Bryan Pardo. 2012. Repeating pattern extraction technique (repet): A simple method for music/voice separation. *IEEE transactions on audio, speech, and language processing*, 21(1):73–84.

Zafar Rafii and Bryan Pardo. 2013. Online repet-sim for real-time speech enhancement. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 848–852.

Vibha Tiwari. 2010. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2022. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, pages 1–12.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

A Additional Results

Additional results to support the choice of hyper-parameters such as fusion size, positive link weight, and update size, are presented in Tables 5, 6, and 7, respectively.

Fusion size is the hyper-parameter of the Fusion Module (see Section 3). There is a possible pattern in Table 5, that a larger fusion size is beneficial for emotion classification in terms of macro F1 score models. On the other hand, it depends also on the model size. Larger models such as "bert-large-cased" may benefit from larger fusion size, but it seems the opposite for smaller ones such as "bert-base-cased".

Fusion Size	PLM	Jaccard	Acc.	Macro F1	wF1
768	all	0.339	0.845	0.467	–
1024	all	0.338	0.848	0.469	0.295
1536	all	0.336	0.843	0.482	0.293
768	bert-base-cased	0.340	0.847	0.464	–
1024	bert-base-cased	0.339	0.849	0.465	0.294
1536	bert-base-cased	0.325	0.838	0.472	0.277
768	bert-large-cased	0.335	0.844	0.469	–
1024	bert-large-cased	0.338	0.846	0.485	0.305
1536	bert-large-cased	0.342	0.844	0.494	0.302

Table 5: Average results for fusion size hyperparameter across PLMs: Jaccard index is used for link prediction, Accuracy and Macro F1 for emotion prediction, and wF1 for ECPE task

Positive link weight is used in the loss function to increase the importance of positive links in a sparse adjacency matrix. According to Table 6, the hyper-parameter importance is not so significant and the differences are more likely due to other settings such as fusion scenario or PLM. Generally, it worked well with a weight set to 5, which is also the best in terms of macro pair F1 score (*mpF1*).

Weight	Jaccard	Acc.	mpF1
1	0.311	0.836	0.331
5	0.325	0.833	0.345
10	0.319	0.829	0.336
20	0.329	0.834	0.339
50	0.319	0.838	0.335

Table 6: Average results for different weights of positive link: Jaccard index is used for link prediction, Accuracy for emotion prediction, and mpF1 for ECPE task

Update size represents the number of samples used for one weight update using gradient accumulation technique. There is a drop in performance with larger update size in Table 7.

Update Size	Jaccard	Acc.	Macro F1	wF1
8	0.338	0.846	0.484	0.310
12	0.336	0.845	0.480	0.297
24	0.345	0.849	0.469	0.301
60	0.340	0.852	0.462	0.302
120	0.290	0.841	0.414	0.207

Table 7: Average results for different update size (batch · gradient accumulation steps): Jaccard index is used for link prediction, Accuracy and Macro F1 for emotion prediction, and wF1 for ECPE task