

LastResort at SemEval-2024 Task 3: Exploring Multimodal Emotion Cause Pair Extraction as Sequence Labelling Task

Suyash Vardhan Mathur*

IIIT Hyderabad

suyash.mathur@research.iiit.ac.in

Hardik Mittal

IIIT Hyderabad

hardik.mittal@research.iiit.ac.in

Akshett Rai Jindal*

IIIT Hyderabad

akshett.jindal@research.iiit.ac.in

Manish Shrivastava

IIIT Hyderabad

m.shrivastava@iiit.ac.in

Abstract

Conversation is the most natural form of human communication, where each utterance can range over a variety of possible emotions. While significant work has been done towards the detection of emotions in text, relatively little work has been done towards finding the cause of the said emotions, especially in multimodal settings. SemEval 2024 introduces the task of Multimodal Emotion Cause Analysis in Conversations, which aims to extract emotions reflected in individual utterances in a conversation involving multiple modalities (textual, audio, and visual modalities) along with the corresponding utterances that were the cause for the emotion. In this paper, we propose models that tackle this task as an utterance labeling and a sequence labeling problem and perform a comparative study of these models, involving baselines using different encoders, using BiLSTM for adding contextual information of the conversation, and finally adding a CRF layer to try to model the inter-dependencies between adjacent utterances more effectively. In the official leaderboard for the task, our architecture was ranked 8th, achieving an F1-score of 0.1759 on the leaderboard. We also release our code here¹.

1 Introduction

Emotion Analysis is one of the fundamental and earliest sub-fields of NLP that focus on identifying and categorizing emotions that are expressed in text. Earlier, research in this domain focused on Emotion Detection in news articles and headlines (Lei et al., 2014; Abdul-Mageed and Ungar, 2017). However, later Emotion Recognition in Conversation gained popularity due to the widespread availability of public conversation data (Gupta et al., 2017). Recently, the task of emotion cause analysis has gained traction, which tries to

identify the causes behind certain emotions (Xia and Ding, 2019a). This has widespread application such as building chatbots that can identify the emotions of the user and even identify the cause behind the emotions to perform certain actions (Pamungkas, 2019). For instance, companies can identify causes behind dissatisfaction in customer interactions and take appropriate measures (Yun and Park, 2022), AI-driven therapeutic insights can be gained using such models (DAfonso, 2020), social media content moderation can be better done (Sawhney et al., 2021), work management and team management by companies can be improved (Benke et al., 2020).

In the task Wang et al., 2024, we tackle the problem of Multimodal Emotion Cause Pair Extraction, where given a set of utterances in a conversation, we must identify the following:

1. **Emotion** of every utterance (if any). These emotions can be one of Ekman’s six basic emotions (Ekman et al., 1999).

2. **Cause** of these emotions, which is considered as the utterance that explicitly expresses an event or argument that is highly linked to the corresponding emotion.

Our proposed system tackles the task in a 3-step fashion – (a) First, we train a model to identify the emotions that are expressed in individual utterances in a conversation. (b) Next, we train a model to identify whether an utterance can be a cause of an emotion expressed in another/same utterance (candidate causes). (c) Finally, we train a model to pair emotion-utterances with their causes among the possible candidate causes. For both the (a) and (b) models we experiment with 3 basic architectures – (i) a simple Neural Network to determine the class of emotion (N-class classifier) and another Neural Network to identify whether the utterance is a candidate cause or not (binary classification). (ii) A BiLSTM (Sak et al., 2014) architecture that accounts for the surrounding context

*Equal contribution.

¹github.com/akshettrj/semEval2024_task03

of the conversation while doing the N-class and binary-classification. (iii) A BiLSTM CRF (Lafferty et al., 2001) architecture which accounts for the surrounding emotions as well while doing the N-class classification. We also experiment with different encoders for the three modalities.

2 Background

2.1 Dataset

The dataset used for this problem is **Emotion-Cause-in-Friends** prepared by Wang et al., 2023 specifically for this task. It has been prepared using conversations from the popular 1994 sitcom *Friends* as the source. This dataset contains 1,344 conversations made up of a total of 13,509 utterances, each conversation containing an average of 10 utterances. For each utterance, the dataset has an annotated transcript (covering text modality) and the corresponding video clip (covering visual and auditory modalities) from the show.

Each utterance is annotated with the emotion depicted by it, which is one of: anger, disgust, fear, joy, neutral, sadness and surprise. The dataset is highly skewed in terms of the frequency of different emotions in the dataset (see Figure 2). Further, the emotion-causes pairs for all the non-neutral utterances are provided in the dataset in a separate list.

The task MC-ECPE expects the model to take a list of such conversations and predict the emotion and emotion-cause pairs labels.

2.2 Related Work

A lot of work has been done in the field of emotion analysis in textual settings. Soon, work began on extracting not only the emotion but also the cause of that extracted emotion. People employed mainly two approaches for emotion cause analysis - 1. Extracting the potential causes given an emotion (Lee et al., 2010; Chen et al., 2010; Gui et al., 2016b) and 2. Extracting the emotion-cause pairs jointly (Xia and Ding, 2019b; Ding et al., 2020; Wei et al., 2020).

Poria et al., 2020 was the first to introduce the task of extracting emotion-cause in conversations but their focus was also only on the textual dialogues. However, in our natural way of conversation, we rely on things like facial expressions, voice intonations for determining the emotion of the speaker. We also rely on auditory and visual scenes to determine the cause of the speaker's emo-

tions. Hence, it is clear that **Emotion-Cause Pair Extraction (ECPE)** is a multimodal task requiring at least three modalities: **text**, **audio** and **video**. Busso et al., 2008; McKeown et al., 2012; Li et al., 2022a and Poria et al., 2019 worked in the field of multimodal emotion analysis in conversations but they did not consider the emotion causes.

The task of MC-ECPE was first worked on by Wang et al., 2024.

3 System Overview

3.1 Baseline I: Utterance labeling

Our baseline model treats the problem as a simple **utterance labeling task**. We use pre-trained text, audio, and image encoders to encode the individual modalities and use these to train three models that can identify the emotions in the utterances, the candidate cause utterances, and finally identify valid emotion and cause utterance(s) pairs.

- **Text Encoding:** For encoding the transcription of each utterance, we use pre-trained BERT (Devlin et al., 2018) embeddings as the baseline embeddings. Additionally, we finetune DeBERTa-Base (He et al., 2020) on the training data for our experiments. DeBERTa makes use of a disentangled attention mechanism and an enhanced masked encoder to improve upon BERT's performance in a variety of tasks. Finally, we also tried RoBERTa-Large and (Liu et al., 2019) pre-trained EmotionRoBERTa-Base² which is publicly available RoBERTa-base model finetuned on the Go Emotions dataset (Demszky et al., 2020). For every text encoder, we perform mean-pooling of the word embeddings to get the textual representation of the utterance.
- **Video Encodings:** For encoding the videos, we sampled 16 equally spaced frames from the video and mean-pooled the embeddings for the 16 frames. For encoding these 16 images, we used MVITv2-small (Li et al., 2022b) encoder, which achieves state-of-the-art performance on the Kinetics video detection task (Kay et al., 2017), which makes it an obvious choice for recognizing activities happening in the conversations relevant for emotion/cause detection.

²https://huggingface.co/SamLowe/roberta-base-go_emotions

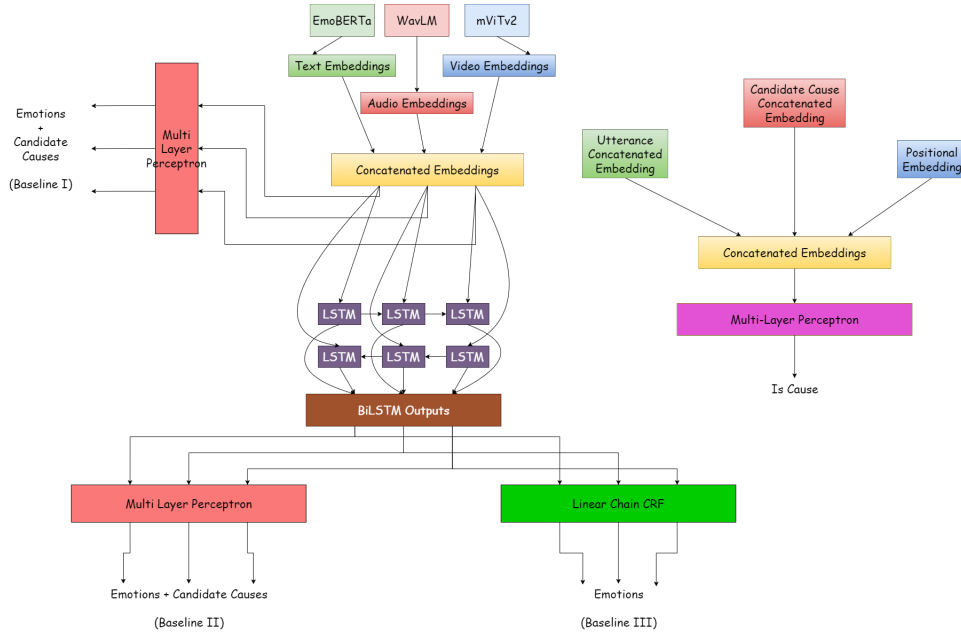


Figure 1: Model Architecture

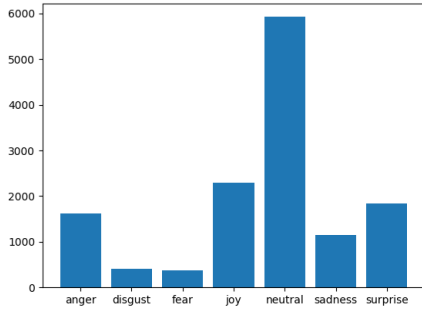


Figure 2: Emotion frequency in the dataset

- **Audio Encodings:** We used WavLM (Chen et al., 2022) for generating audio embeddings, which is trained on large audio data using masked speech representation and denoising in pre-training, making it suitable for various downstream speech tasks. We also try Wav2Vec2-Large (Baevski et al., 2020), which is trained by masking speech input in latent space and solving a contrastive task defined over a quantization of the latent representations which are jointly learned.

The model architecture is a combination of **three steps**, each of which is described below:

Step 1 – Emotion Classification

First, we concatenate the text, audio and video embeddings from the respective encoders and pass these concatenated embeddings into a dense layer,

on which a Softmax function is applied to get the probability distribution over 7 classes (6 emotions and one neutral class). Due to a skewed distribution of the emotion labels in the dataset, we make use of **weighted Cross Entropy loss** to train the model, where the weights are taken as inverse of the frequency of the labels in the training dataset.

Step 2 – Candidate Cause Identification

For identifying the candidate cause, we similarly pass concatenated embeddings through a dense layer with a Sigmoid function, which predicts the probability of whether the utterance is a candidate cause or not. Binary Cross Entropy Loss is used to train the model.

Step 3 – Emotion-Cause pairing

For pairing the emotion utterances with the candidate causes, we concatenate the representations for the emotion utterance and the cause utterance, with a distance embedding. This distance embedding is generated by giving positional embedding to each utterance, sampled from a Normal Distribution. This representation is passed through a dense layer with a Sigmoid function, which learns to predict the probability of the emotion-cause utterance pair being a valid emotion-cause pair or not for the given conversation, trained using Binary Cross Entropy Loss.

3.2 Baseline II: BiLSTM Architecture

The BiLSTM architecture is inspired by the work in Wang et al., 2024. While the Baseline I architecture treats the emotion and cause classification independently for each utterance, it is dependent on the surrounding context of the conversation too. Thus, the BiLSTM architecture models the problem as a **Sequence Labeling task**. We use the best encoders in the Baseline I architecture for generating the embeddings in this architecture.

Step 1 – Emotion Classification

Similar to the Baseline Model, we concatenate the embeddings of the three modalities, and pass them to a stacked BiLSTM. On top of the BiLSTM outputs, we apply a 7-class classifier to obtain the emotion category distribution. Similar to the Baseline I, weighted cross-entropy loss is used.

Step 2 – Candidate Cause Identification

For Candidate Cause prediction, similarly, the concatenated embeddings are passed through a BiLSTM on top of which a binary classifier is applied.

Step 3 – Emotion-Cause Pairing

The Emotion-Cause pairing model remains the same in this architecture as the Baseline I model.

In this architecture, BiLSTM provides the advantages of bidirectional and longer contexts which should help understand the emotions present in utterances better. This is because in a conversation, it is possible that the emotions are not just dependent on the current utterance, but on surrounding multimodal utterances as well.

3.3 Baseline III: BiLSTM-CRF Architecture

In the BiLSTM model, each classification decision was conditionally independent. Linear-chain CRFs are models generally used to model structured data where one output influences its neighboring outputs as it models the various transition probabilities, and have been extensively used with BiLSTMs for sequence labeling (Huang et al., 2015). This could be useful for emotion predictions because the emotion of one utterance is generally influenced by the emotions in its previous utterances. For instance, an utterance with happiness generally tends to be followed by another happiness utterance.

Step 1 – Emotion Classification

For this architecture, we add a CRF layer on top of the BiLSTM layers, and make use of the CRF-loss to train the model instead of Cross-Entropy loss as in the previous architectures. This loss models the transitions between the labels in the architecture, modelling the task as a more complex sequence labeling task. Thus, while the BiLSTM layer learns more about the language and emotions expressed through the language, the CRF layer tries to learn about the relations between the emotions.

Step 2 – Candidate Cause Identification

For Candidate Cause prediction, the architecture remains the same as in Baseline II. This is because the transitions between cause labels (being cause of an emotion in an utterance or not) does not make intuitive sense, and using BiLSTMs to capture surrounding context from other utterances is what seems more appropriate.

Step 3 – Emotion-Cause Pairing

The Emotion-Cause pairing model remains the same in this architecture as the Baseline I & II models.

4 Experimental Setup

We perform a random shuffle and use a 90-10% split for the train-validation split. The test set was provided by the authors, but its gold labels have not been made public.

The experiments involving Baseline II and III use *EmotionRoBERTa* + *WavLM* + *MViTv2* configuration. All the experiments involve applying a dropout of 0.3 on the audio, visual and textual embeddings before they are passed on to the main architectures. The BiLSTM for emotion detection consists of 4 layers while the one for candidate cause identification contains 3 layers. The dropout between the stacked layers of the BiLSTM is kept 0.3 as well. We use AdamW optimizer for all the three models, and use a linear learning rate scheduler with warmup for training the models. The Emotion Classification model is trained for 60 epochs, the Candidate Cause Identification model is trained for 40 epochs, and the Emotion-Cause Pairing Model is trained for 40 epochs as well.

In order to train the Emotion-Cause pairing model, we create positive and negative pairs during training. However, while the number of positive pairs is of the order N , the number of negative

Model Name	Emotion Detection			Candidate Cause Detection			Emotion-Cause Pairing			Emotion-Cause Pairing (Eval.)		Leaderboard		
	P	R	F1	P	R	F1	P	R	F1	wt. F1	Macro F1	wt. F1	Macro F1	
Baseline I														
BERT + WavLM + MViTv2	+	0.61	0.52	0.55	0.71	0.71	0.71	0.93	0.87	0.89	0.26	0.20	0.182	0.165
EmotionRoBERTa + WavLM + MViTv2	+	0.55	0.45	0.47	0.67	0.66	0.66	0.93	0.86	0.89	0.20	0.18	0.187	0.170
DeBERTa (finet.) + WavLM + MViTv2	+	0.44	0.36	0.38	0.60	0.60	0.60	0.92	0.85	0.87	0.10	0.10	0.094	0.094
RoBERTa-L + WavLM + MViTv2	+	0.59	0.47	0.49	0.66	0.65	0.66	0.93	0.86	0.88	0.21	0.19	0.180	0.165
EmotionRoBERTa + Wav2Vec2 + MViTv2	+	0.55	0.47	0.48	0.67	0.67	0.67	0.93	0.87	0.89	0.21	0.20	0.172	0.170
BiLSTM (Baseline II)		0.55	0.51	0.52	0.67	0.67	0.67	0.93	0.86	0.89	0.22	0.21	0.184	0.179
BiLSTM + CRF (Baseline III)	+	0.53	0.56	0.54	0.67	0.67	0.67	0.93	0.86	0.89	0.24	0.18	0.165	0.172

Table 1: Results for baselines on the ECAC dataset

pairs comes to the order of N^2 , and thus we perform a random sampling of the negative pairs to keep the positive and negative samples in the ratio 1:5. This helps us to maintain balance between the positive and negative classes.

Evaluation Metrics

We evaluate the 3 steps separately as well apart from evaluating the performance for the final Emotion-Cause pairs:

Emotion Identification: We use Weighted Precision, Recall and F1-score for the distribution between the 7 classes (6 emotions and neutral class).

Candidate Cause Identification: We again use Weighted Precision, Recall and F1-score for evaluating the prediction between the binary classes – *is_candidate_cause* and *not_candidate_cause*.

Emotion-Cause Pairing: For evaluating this, we generate positive and negative pairs and use Weighted Precision, Recall and F1-score for evaluating the classification between the positive and negative classes.

Emotion-Cause Pairs: Weighted F1-score and Macro F1-score are the official metric used for the final evaluation for the task.

5 Results and Analysis

The performance of the three Baselines can be seen in Table 1. During the Evaluation phase, our best ranked submission of Baseline II had Wt. F1 score of 0.1836 and Macro F1 score of 0.1759, ranking 8th on the leaderboard.

Baseline I

Among the encoders in Baseline I, *BERT + WavLM + MViTv2* configuration performs the best on the validation set, including the individual steps as well as the final emotion-cause pair predictions. However, on the leaderboard, *EmotionRoBERTa + WavLM + MViTv2* gives the best performance, although the difference in the leaderboard scores is marginal among the encoders. This observation might indicate that the test data is a bit different in nature from the training data.

Better performance of EmotionRoBERTa can be attributed to the fact that the model’s weights have already been finetuned towards emotion-related tasks. Further, it seems that finetuning DeBERTa on the training data caused it to overfit, leading to worse performance than vanilla BERT/RoBERTa models. RoBERTa-L performed slightly worse than BERT and EmotionRoBERTa.

Finally, WavLM being the newer architecture,

as expected performed better than Wav2Vec2. This is because WavLM is more robust than Wav2Vec2 and it is trained in a combination of supervised and self-supervised learning, making its performance much better.

Baseline II

We use the *EmotionRoBERTa + WavLM + MViTv2* configuration as encoders for the Baseline II architecture. Contrary to expectation, the wt. F1 score on the leaderboard decreased marginally, while the Macro F1 score increased marginally. This is probably because of the nature of the dataset, where the average length of a conversation is as little as 10, which causes the context of the utterance to be rather limited. In such a situation, the additional context from previous utterances doesn't prove helpful to the model, and might even prove to be noise for the model, leading to the results observed.

Baseline III

In this, we can observe a significant fall in wt. F1 and slight fall in Macro F1 score from the Baseline I and II architectures. This is in line with the observation of Baseline II that due to the nature of the dataset, sequence labeling it is not necessarily the best way to model it. Further, due to small number of utterances in the conversations, it is likely that the transition between labels needed for CRF doesn't get trained that well and leads to poorer performance.

6 Conclusion

In conclusion, we observe that the utterance labeling systems perform as good as sequence labeling systems for this specific dataset. Further, we also see that encoders which are trained on other emotion-related tasks tend to perform better on similar emotion-related tasks.

In future, it is possible to learn joint embeddings over the 3 modalities, which should provide better representations for each utterance (Girdhar et al., 2023). Further, it can be experimented to utilize the speaker information for each utterance while creating utterance representations (Liang et al., 2023).

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. [EmoNet: Fine-grained emotion detection with gated](#)

[recurrent neural networks](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

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.

Ivo Benke, Michael Thomas Knierim, and Alexander Maedche. 2020. [Chatbot-based emotion management for distributed teams: A participatory design study](#). *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–30.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. [Iemocap: interactive emotional dyadic motion capture database](#). *Language Resources and Evaluation*, 42:335–359.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Churen Huang. 2010. [Emotion cause detection with linguistic constructions](#). In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187, Beijing, China. Coling 2010 Organizing Committee.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.

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*.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. [ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170, Online. Association for Computational Linguistics.

Simon DAlfonso. 2020. [Ai in mental health](#). *Current Opinion in Psychology*, 36:112–117.

- Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.
- Rohit Girdhar, Alaeldin El-Nouby, Zhuang Liu, Man-
nat Singh, Kalyan Vasudev Alwala, Armand Joulin,
and Ishan Misra. 2023. Imagebind: One embed-
ding space to bind them all. In *Proceedings of the
IEEE/CVF Conference on Computer Vision and Pat-
tern Recognition (CVPR)*, pages 15180–15190.
- Lin Gui, Ruifeng Xu, Qin Lu, Dongyin Wu, and
Yu Zhou. 2016b. Emotion cause extraction, a chal-
lenging task with corpus construction. In *Social Me-
dia Processing*, pages 98–109, Singapore. Springer
Singapore.
- Umang Gupta, Ankush Chatterjee, Radhakrishnan
Srikanth, and Puneet Agrawal. 2017. A sentiment-
and-semantics-based approach for emotion detec-
tion in textual conversations. *arXiv preprint
arXiv:1707.06996*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and
Weizhu Chen. 2020. Deberta: Decoding-enhanced
bert with disentangled attention. *arXiv preprint
arXiv:2006.03654*.
- Zhiheng Huang, Wei Xu, and Kai Yu. 2015. [Bidirec-
tional lstm-crf models for sequence tagging](#).
- Will Kay, Joao Carreira, Karen Simonyan, Brian
Zhang, Chloe Hillier, Sudheendra Vijaya-
narasimhan, Fabio Viola, Tim Green, Trevor
Back, Paul Natsev, Mustafa Suleyman, and Andrew
Zisserman. 2017. [The kinetics human action video
dataset](#).
- John Lafferty, Andrew McCallum, and Fernando CN
Pereira. 2001. Conditional random fields: Prob-
abilistic models for segmenting and labeling se-
quence data.
- 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, Los Angeles, CA. Association for Computa-
tional Linguistics.
- Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan,
and Liu Wenyin. 2014. [Towards building a social
emotion detection system for online news](#). *Future
Generation Computer Systems*, 37:438–448. Spe-
cial Section: Innovative Methods and Algorithms for
Advanced Data-Intensive Computing Special Sec-
tion: Semantics, Intelligent processing and services
for big data Special Section: Advances in Data-
Intensive Modelling and Simulation Special Section:
Hybrid Intelligence for Growing Internet and its Ap-
plications.
- Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao
Zhu, and Erik Cambria. 2022a. [Ecpec: Emotion-
cause pair extraction in conversations](#). *IEEE Trans-
actions on Affective Computing*, 14(3):1754–1765.
- Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya
Mangalam, Bo Xiong, Jitendra Malik, and Christoph
Feichtenhofer. 2022b. Mvitv2: Improved multi-
scale vision transformers for classification and detec-
tion. In *Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition*, pages
4804–4814.
- Xingwei Liang, You Zou, and Ruifeng Xu. 2023. Si-
lstm: Speaker hybrid long-short term memory and
cross modal attention for emotion recognition in con-
versation. *arXiv preprint arXiv:2305.03506*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man-
dar Joshi, Danqi Chen, Omer Levy, Mike Lewis,
Luke Zettlemoyer, and Veselin Stoyanov. 2019. Ro-
berta: A robustly optimized bert pretraining ap-
proach. *arXiv preprint arXiv:1907.11692*.
- Gary McKeown, Michel Valstar, Roddy Cowie, Maja
Pantic, and Marc Schroder. 2012. [The semaine
database: Annotated multimodal records of emotion-
ally colored conversations between a person and a
limited agent](#). *IEEE Transactions on Affective Com-
puting*, 3(1):5–17.
- Endang Wahyu Pamungkas. 2019. [Emotionally-aware
chatbots: A survey](#).
- Soujanya Poria, Devamanyu Hazarika, Navonil Ma-
jumder, Gautam Naik, Erik Cambria, and Rada Mi-
halcea. 2019. [MELD: A multimodal multi-party
dataset for emotion recognition in conversations](#). In
*Proceedings of the 57th Annual Meeting of the As-
sociation for Computational Linguistics*, pages 527–
536, Florence, Italy. Association for Computational
Linguistics.
- Soujanya Poria, Navonil Majumder, Devamanyu Haz-
arika, Deepanway Ghosal, Rishabh Bhardwaj, Sam-
son Yu Bai Jian, Romila Ghosh, Niyati Chhaya,
Alexander F. Gelbukh, and Rada Mihalcea. 2020. [Recognizing emotion cause in conversations](#). *CoRR*,
abs/2012.11820.
- Haşim Sak, Andrew Senior, and Françoise Bea-
ufays. 2014. Long short-term memory based
recurrent neural network architectures for large
vocabulary speech recognition. *arXiv preprint
arXiv:1402.1128*.
- Ramit Sawhney, Harshit Joshi, Alicia Nobles, and Ra-
jiv Ratn Shah. 2021. Towards emotion-and time-
aware classification of tweets to assist human mod-
eration for suicide prevention. In *Proceedings of the
International AAAI Conference on Web and Social
Media*, volume 15, pages 609–620.
- Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and
Jianfei Yu. 2023. [Multimodal emotion-cause pair
extraction in conversations](#). *IEEE Transactions on
Affective Computing*, 14(3):1832–1844.
- Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and
Erik Cambria. 2024. [Semeval-2024 task 3: Multi-
modal 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.
- Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. [Effective inter-clause modeling for end-to-end emotion-cause pair extraction](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online. Association for Computational Linguistics.
- Rui Xia and Zixiang Ding. 2019a. 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.
- Rui Xia and Zixiang Ding. 2019b. [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, Florence, Italy. Association for Computational Linguistics.
- Jeewoo Yun and Jungkun Park. 2022. The effects of chatbot service recovery with emotion words on customer satisfaction, repurchase intention, and positive word-of-mouth. *Frontiers in psychology*, 13:922503.