Recognizing Emotion Cause in Conversations

Soujanya Poria · Navonil Majumder · Devamanyu Hazarika · Deepanway Ghosal · Rishabh Bhardwaj · Samson Yu Bai Jian · Pengfei Hong · Romila Ghosh · Abhinaba Roy · Niyati Chhaya · Alexander Gelbukh · Rada Mihalcea

Abstract We address the problem of recognizing emotion cause in conversations, define two novel sub-tasks of this problem, and provide a corresponding dialogue-
level dataset, along with strong Transformer-based baselines. The dataset is available at [https://github.com/declare-lab/RECCON](https://github.com/declare-lab/RECCON).

**Introduction** Recognizing the cause behind emotions in text is a fundamental yet under-explored area of research in NLP. Advances in this area hold the potential to improve interpretability and performance in affect-based models. Identifying emotion causes at the utterance level in conversations is particularly challenging due to the intermingling dynamics among the interlocutors.

**Method** We introduce the task of Recognizing Emotion Cause in CONversations with an accompanying dataset named RECCON, containing over 1,000 dialogues and 10,000 utterance cause-effect pairs. Furthermore, we define different cause types based on the source of the causes, and establish strong Transformer-based baselines to address two different sub-tasks on this dataset: causal span extraction and causal emotion entailment.

**Result** Our Transformer-based baselines, which leverage contextual pre-trained embeddings, such as RoBERTa, outperform the state-of-the-art emotion cause extraction approaches on our dataset.

**Conclusion** We introduce a new task highly relevant for (explainable) emotion-aware artificial intelligence: recognizing emotion cause in conversations, provide a new highly challenging publicly available dialogue-level dataset for this task, and give strong baseline results on this dataset.

1 Introduction

Emotions are intrinsic to humans; consequently, emotion understanding is a key part of human-like artificial intelligence (AI). Language is often indicative of one’s emotions. Hence, emotion recognition has attracted much attention in the field of natural language processing (NLP) [Kratzwald et al., 2018; Colnerič and Demsar, 2018] due to its wide range of applications in opinion mining, recommender systems, healthcare, and other areas.

In particular, emotions are an integral part of human cognition; thus understanding human emotions and reasoning about them is one the key issues in computational modeling of human cognitive processes [Izard, 1992]. Among different settings where human emotions play important cognitive role is human-human and human-computer conversations. Similarly, among different issues in automatic reasoning about human emotions is identifying the causal root of the expressed emotions in the discourse of such a conversation. During a dialog, cognitive and affective processes can be triggered by non-verbal external events or sensory input. Sometimes such affective processes can happen even before the corresponding cognitive processing by the person—a phenomenon called affective primacy [Zajonc, 1980]. On
the other hand, complex cognitive processing, which would lead to updating the computational model’s speaker state, can also happen before or after the affective reaction of the participant of the conversation.

Substantial progress has been made in the detection and classification of emotions, expressed in text or videos, according to emotion taxonomies (Ekman, 1993; Plutchik, 1982). However, further reasoning about emotions, such as understanding the cause of an emotion expressed by a speaker, has been less explored so far. For example, understanding the following review of a smartphone, “I hate the touchscreen as it freezes after 2-3 touches”, implies not only detecting the expressed negative emotion, specifically DISGUST, but also spotting its cause (Liu, 2012)—in this case, “it freezes after 2-3 touches.”

Of a wide spectrum of emotion-reasoning tasks (Ellsworth and Scherer, 2003), in this work we focus on identifying the causes (also called antecedents, triggers, or stimuli) of emotions expressed specifically in conversations. In particular, we look for events, situations, opinions, or experiences in the conversational context that are primarily responsible for an elicited emotion in the target utterance. Apart from event mentions, the cause could also be a speaker’s counterpart reacting towards an event cared for by the speaker (inter-personal emotional influence).

We introduce the task of recognizing emotion cause in conversations (REC-CON), which refers to the extraction of such stimuli behind an emotion in a conversational utterance. The cause could be present in the same or contextual utterances. We formally define this task in Section 4.2.

In Fig. 1 we exemplify this task. In the first example, we want to know the cause of person B’s ($P_B$) emotion (HAPPY). It can be seen that $P_A$ is happy due to the event “getting married” and similarly $P_B$ also reacts positively to this event. Here, we can infer that $P_B$’s emotion is caused either by the reference of the first utterance to the event of getting married or by the fact that $P_A$ is happy about getting married—both of which can be considered as stimulus for $P_B$’s emotion. In the second example, the cause of $P_A$’s emotion is the event “football match” and a negative emotion DISGUST indicates that $P_A$ is unsatisfied with the match. In contrast, $P_B$ likes the match—sharing the same cause with $P_A$—with HAPPINESS emotion. These examples demonstrate the challenging problem of recognizing emotion causes in con-

![Fig. 1: Emotion causes in conversations.](image)
versations, which to the best of our knowledge, is one of the first attempts in this area of research.

We can summarize our contributions as follows:

1. We introduce a new task, **recognizing emotion cause in conversations**, and dive into many unique characteristics of this task that is peculiar to conversations. In particular, we define the relevant types of emotion causes (Section 5).
2. Further, we describe a new annotated dataset for this task, **RECCON**\(^1\), including both acted and real-world dyadic conversations (Section 4). To the best of our knowledge, there is no other dataset for the task of emotion cause recognition in conversations.
3. Finally, we introduce two challenging sub-tasks that demand complex reasoning, and setup strong baselines to solve the sub-tasks (Section 6). These baselines surpass the performance of several newly introduced complex neural approaches, e.g., ECPE-MLL (Ding et al., 2020b), RankCP (Wei et al., 2020), and ECPE-2D (Ding et al., 2020a).

2 Related Work

Initial works on emotion analysis were applied to the opinion mining task, exploring different aspects of affect beyond polarity prediction, such as identifying the opinion / emotion feeler (holder, source) (Das and Bandyopadhyay, 2010; Choi et al., 2005). More recently, sentiment analysis research has been used in a wider context of natural language understanding and reasoning (Dragoni et al., 2021). New methods for multi-label emotion classification are developed (Ameer et al., 2020) and new corpora for emotion detection are compiled for languages other than English (Moreno Jiménez and Torres Moreno, 2020).

The task of emotion cause extraction was studied initially by Lee et al. (2010). The early works used rule-based approaches (Chen et al., 2010). Gui et al. (2016) constructed an emotion cause extraction dataset by identifying events that trigger emotions. They used news articles as their source for the corpus to avoid the latent emotions and implicit emotion causes associated with the informal text, thus reducing reasoning complexity for the annotators while extracting emotion causes. Other notable works on emotion cause extraction (ECE) are (Ghazi et al., 2015) and (Gao et al., 2017).

As a modification of the ECE task, Xia and Ding (2019) proposed emotion-cause pair extraction (ECPE) that jointly identifies both emotions and their corresponding causes (Chen et al., 2018). Further, Chen et al. (2020) recently proposed the conditional emotion cause pair (ECP) identification task, where they highlighted the causal relationship to be valid only in particular contexts. We incorporate this property in our dataset construction, as we annotate multiple spans in the conversational history that sufficiently indicate the cause. Similar to Chen et al. (2020), we also provide negative examples of context that does not contain the causal span.

\(^1\) pronounced as *reckon.*
Our work is a natural extension of those works. We propose a new dataset on dyadic conversations, which is more difficult to annotate. Additionally, the associated task of recognizing emotion cause in conversations poses a greater hitch to solve due to numerous challenges. For example, (i) expressed emotions are not always explicit in the conversations; (ii) conversations can be very informal where the phrase connecting emotion with its cause can often be implicit and thus needs to be inferred; (iii) the stimuli of the elicited emotions can be located far from the target utterance in the conversation history, so that detecting it requires complex reasoning and coreference, often using commonsense.

3 Definition of the Task

We distinguish between emotion evidence and emotion cause:

- **Emotion evidence** is a part of the text that indicates the presence of an emotion in the speaker’s emotional state. It acts in the real world between the text and the reader. Identifying and interpreting the emotion evidence is the underlying process of the well-known emotion detection task.
- **Emotion cause** is a part of the text expressing the reason for the speaker to feel the emotion given by the emotion evidence. It acts in the described world between the (described) circumstances and the (described) speaker’s emotional state. Identifying the emotion cause constitutes the task we consider in this paper.

For instance, in Fig. 1, $P_B$’s turn contains evidence of $P_B$’s emotion, while $P_A$’s turn contains its cause. The same text span can be both emotion evidence and cause, but generally this is not the case.

Defining the notion of emotion cause is, in a way, the main goal of this paper. However, short of a formal definition, we will explain this notion on numerous examples and, in computational terms, via the labeled dataset.

We use the following terminology throughout the paper. The **target utterance** $U_t$ is the $t^{th}$ utterance of a conversation, whose emotion label $E_t$ is known and whose emotion cause we want to identify. The **conversational history** $H(U)$ of the utterance $U$ is the set of all utterances from the beginning of the conversation till the utterance $U$, including $U$. A **causal span** for an utterance $U$ is a maximal sub-string, of an utterance from $H(U)$, that is a part of $U$’s emotion cause; we will denote the set of the causal spans for an utterance $U$ by $CS(U)$. A **causal utterance** is an utterance containing a causal span; we denote the set of all causal utterances for $U$ by $C(U) \subseteq H(U)$. An **utterance–causal span (UCS) pair** is a pair $(U, S)$, where $U$ is an utterance and $S \in CS(U)$.

Thus, recognizing emotion cause is the task of identifying all (correct) UCS pairs in a given text.

In the context of our training procedure, we will refer to (correct) UCS pairs as **positive examples**, whereas pairs $(U, S)$ with $S \notin CS(U)$ are **negative examples**. In Section [6.1.1] we describe the sampling strategies for negative examples.
4 Building the RECCON dataset

4.1 Emotional Dialogue Sources

We consider two popular conversation datasets IEMOCAP (Busso et al., 2008) and DailyDialog (Li et al., 2017), both equipped with utterance-level emotion labels:

IEMOCAP is a dataset of two-person conversations in English annotated with six emotion classes: ANGER, EXCITED, FRUSTRATED, HAPPY, NEUTRAL, SAD. The dialogues in this dataset span across sixteen conversational situations. To avoid redundancy, we handpicked only one dialogue from each of these situations. We denote the subset of our dataset comprising these dialogues as RECCON-IE.

DailyDialog is an English-language natural human communication dataset covering various topics on our daily lives. All utterances are labeled with emotion categories: ANGER, DISGUST, FEAR, HAPPY, NEUTRAL, SAD, SURPRISE. Since the dataset is skewed (83% NEUTRAL labels), we randomly selected dialogues with at least four non-NEUTRAL utterances. We denote the subset of RECCON comprising these dialogues from DailyDialog as RECCON-DD. Some statistics about the annotated dataset is shown in Table 2.

Thus our RECCON dataset consists of two parts, RECCON-IE and RECCON-DD. In particular, the label sets are slightly different in these two parts, as explained above.

**Why sampling from two datasets** Although both IEMOCAP and DailyDialog are annotated with utterance-level emotions, they differ in many aspects. First, IEMOCAP has more than 50 utterances per dialogue on average, whereas DailyDialog has only 8 on average. Second, the shifts between non-neutral emotions (e.g., SAD to ANGER, HAPPY to EXCITED) are more frequent in IEMOCAP than in DailyDialog; see (Ghosal et al., 2020). Consequently, both cause detection and causal reasoning in IEMOCAP are more interesting as well as difficult. Lastly, in Table 2 we can see that in our annotated IEMOCAP split, almost 40.5% of utterances have their emotion cause in utterances at least 3 timestamps distant in the contextual history. In contrast, this percentage is just 13% in our annotated DailyDialog dataset.

4.2 Annotation Process

**Annotators** The annotators were undergraduate and graduate computer science students. They had adequate knowledge about the problem of emotion cause recognition; in particular, we organized a special workshop to instruct them on the topic. Their annotations were first verified on a trial dataset, and feedback was provided to them to correct their mistakes. Once they achieved satisfactory performance on the trial dataset, they were qualified for the main dataset annotation. While the annotators were not native English speakers, they communicate in English in their daily life, and their medium of instruction in their study was English.
Annotation guidelines Given an utterance $U_i$ labeled with an emotion $E_t$, the annotators were asked to extract the set of causal spans $CS(U_t)$ that sufficiently represent the causes of the emotion $E_t$. If the cause of $E_t$ was latent, i.e., there was no explicit causal span in the dialog, the annotators wrote down the assumed causes that they inferred from the text. Each utterance was annotated by two human experts—graduate students with reasonable knowledge of the task.

In fact, the annotators were asked to look for the casual spans of $U_t$ in the whole dialog and not only in the past history $H(U_t)$. We show a case in Fig. 2b where the causal span of the emotion FEAR in utterance 1 is recognized in utterance 3: “someone is stalking me”. However, there were only seven instances of the utterances with explicit emotion causal spans in the conversational future with respect to $U_t$ in the whole dataset. So we discarded those spans and decided to consider only causal spans in $H(U_t)$; hence the definition in Section 3.

Emotional expression An utterance can contain (i) a description of the triggers or stimuli of the expressed emotion, and / or (ii) a reactionary emotional expression. In our setup, by following the discrimination among emotion evidence and cause as explained in Section 3 we instructed the annotators to look beyond just emotional expressions and identify the emotion cause. We can illustrate this with Fig. 2c where $P_A$ explains the cause for HAPPINESS; the same cause evokes the emotion EXCITED in $P_B$. Meanwhile, the utterance 2 by $P_B$ is merely an emotional expression (evidence).

Emotion cause can also corroborate in generating an emotional expression, e.g., in Fig. 2c the event “winning the prize” causes EXCITED emotion in $P_B$ which directs $P_B$ to utter the expression “Wow! Incredible”. This type of generative reasoning will be very important in our future work.

Why span detection? First, emotion-cause extraction has historically been defined as an information extraction task of identifying spans within the emotion-bearing sentences (Xia and Ding, 2019; Ghazi et al., 2015). The core assumption is that such spans are good descriptors of the underlying causes of the generated emotions (Talmy, 2000). We extend this popular formalism into a multi-span framework. Second, while recognizing emotion cause is driven by multiple controlling variables such as goal, intent, personality, we adopt this setup as these spans can often represent or allude to these controlling variables. A more elaborate setup would require explaining how the spans can be combined to form the trigger and consequently evoke the emotion (see Fig. 7); we leave such emotion causal reasoning in conversations to future work.

4.2.1 Annotation Aggregation

Following Gui et al. (2016), we aggregate the annotations in two stages: at utterance and span level.

Stage 1: Utterance-level aggregation Here, we decide whether an utterance is causal by majority voting: a third expert annotator is brought in as the tie breaker.
Fig. 2: (a) No context. (b) Unmentioned latent cause. (c) Distinguishing emotion cause from emotional expressions.

| Dataset                  | Language | Source                   | Size       | Format   |
|--------------------------|----------|--------------------------|------------|----------|
| Neviarouskaya and Aono   | English  | ABBYY Lingvo dictionary  | 532 sentences |          |
| Gui et al. [2014]        | Chinese  | Chinese Weibo            | 1333 sentences |          |
| Ghazi et al. [2015]      | English  | FrameNet                 | 1519 sentences |          |
| Gui et al. [2016]        | Chinese  | SINA city news           | 2167 clauses |          |
| Gao et al. [2017]        | Chinese / Eng. | SINA city news / English novel | 4054 / 4858 clauses |          |
| RECCON (our)             | English  | DailyDialog / IEMOCAP    | 5861 / 494 utterances | 1106 / 16 dialogues |

Table 1: Datasets for emotion cause extraction and related tasks. Datasets in [Xia and Ding 2019; Chen et al. 2020] are derived from [Gui et al. 2016].

**Stage 2: Span-level aggregation** Within each causal utterance selected at stage 1, we took the union of the candidate spans from different annotators as the final causal span only when the size of their intersection is at least 50% of the size of the shortest candidate span. Otherwise, a third annotator was brought in to determine the final span from the existing spans. This third annotator was also instructed to prefer the shorter spans over the longer ones when they can sufficiently represent the cause.
Fig. 3: (a) **Self-contagion**: The cause of the emotion is primarily due to a stable mood of the speaker that was induced in the previous dialogue turns; (b) **Generic Cause**: (c) **Hybrid**: The hybrid type with both inter-personal emotional influence and self-contagion.

without losing any information. The threshold of 50% of the shortest span was chosen empirically by examining a small subset of the dialogues. The third annotator could not break the tie for 34 causal utterances, which we discarded from the dataset.
### Table 2: Statistics of the RECCON annotated dataset. DD stands for RECCON-DD, IE for RECCON-IE.

| Number of items | DD | IE | Total |
|-----------------|----|----|-------|
| Dialogues       | 1106 | 16 | 1122 |
| Utterances      | 11104 | 665 | 11769 |
| Utterances annotated with emotion cause | 5861 | 494 | 6355 |
| Utterances that cater to background cause | 395 | 70 | 465 |
| Utterances where cause solely lies in the same utterance | 1521 | 80 | 1601 |
| Utterances where cause solely lies in the contextual utterances | 970 | 171 | 1141 |
| Utterances where cause lies both in same and context utterances | 3370 | 245 | 3613 |

| UCS pairs       | 9915 | 1154 | 11069 |
| Utterances having single cause | 55% | 41% | 54% |
| Utterances having two causes | 31% | 24% | 31% |
| Utterances having three causes | 9% | 17% | 9% |
| Utterances having more than three causes | 5% | 18% | 6% |
| Causes per utterance (average) | 1.69 | 2.34 | 1.73 |

| Utterances with | DD | IE | Total |
|-----------------|----|----|-------|
| ANGER           | 451 | 89 | 540 |
| FEAR            | 74  | –  | 74   |
| DISGUST         | 140 | –  | 140 |
| FRUSTRATION     | –   | 109| 109 |
| HAPPY           | 4361| 58 | 4419 |
| SAD             | 351 | 70 | 4419 |
| SURPRISE        | 484 | –  | 484 |
| EXCITED         | –   | 197| 197 |
| NEUTRAL         | 5243| 142| 5385 |

| Utterances $U_t$ with | DD | IE | Total |
|-----------------------|----|----|-------|
| No context            | 43%| 35%| 43% |
| Inter-personal        | 32%| 19%| 31% |
| Self-contagion        | 9% | 20%| 10% |
| Hybrid                | 11%| 17%| 11% |
| Latent                | 5% | 10%| 5%  |

#### 4.3 Dataset Statistics

In Table 1 we compare our dataset with the existing datasets in terms of size, data sources, and language. The remaining statistics of RECCON are consolidated in Table 2.

We measured inter-annotator agreement (IAA) at the level of (i) utterance and (ii) span. At the utterance level, we measured IAA following [Gui et al. (2016)](https://example.com), which gave a kappa of 0.7928. However, as pointed out by [Brandsen et al. (2020)](https://example.com), macro F1 score is more appropriate for span extraction-type tasks. Hence, at the utterance level, we also compute the pairwise macro F1 score between all possible pairs of annotators and then average them, which gives a 0.8839 macro F1 score. [Brandsen et al. (2020)](https://example.com) also suggest the removing negative examples—in our case, the utterances in the conversational history containing no causal span for the emotion of the target utterance—for macro F1 calculation, since such examples are usually very frequent, which may lead to a skewed F1 score. As expected, this yields a lower F1 score of 0.8201. At span level, the F1 score, as explained in [Rajpurkar et al. (2016)](https://example.com), is calculated for all possible pairs of annotators followed by taking their average. Overall, we obtain an F1 score of 0.8035 at span level.
5 Types of Emotion Causes

In our dataset, RECCON, we observe five predominant types of emotion causes that are based on the source of the stimuli (events / situations / acts) in the conversational context, responsible for the target emotion. The annotators were asked to flag the utterances with latent emotion cause or emotion cause of type shown in Fig. 2b (unmentioned latent cause), as explained below. The distribution of these cause types is given in Table 2.

Type 1: No Context The cause is present within the target utterance itself. The speaker feeling the emotion explicitly mentions its cause in the target utterance (see Fig. 2a).

Type 2: Inter-Personal Emotional Influence The emotion cause is present in the other speaker’s utterances (see Fig. 1). We observe two possible sub-types of such influences:

2a) Trigger Events / Situations. The emotion cause lies within an event or concept mentioned by the other speaker.
2b) Emotional Dependency. The emotion of the target speaker is induced from the emotion of the other speaker over some event / situation.

Type 3: Self-Contagion In many cases, the cause of the emotion is primarily due to a stable mood of the speaker that was induced in previous dialogue turns. E.g., in a dialogue involving cordial greetings, there is a tendency for a HAPPY mood to persist across several turns for a speaker. Fig. 3a presents an example where such self-influences can be observed. Utterance 1 establishes that Pa likes winter. This concept triggers a HAPPY mood for the future utterances, as observed in utterances 3 and 5. In Fig. 3b, similarly, the trigger of emotion EXCITED in utterance 3 is mentioned by the same speaker in his or her previous utterance.

Type 4: Hybrid Emotion causes of type 2 and 3 can jointly cause the emotion of an utterance, as illustrated by Fig. 3c.

Type 5: Unmentioned Latent Cause There are instances in the dataset where no explicit span in the target utterance or the conversational history can be identified as the emotion cause. Fig. 2b shows such a case. Here, in the first utterance, Pa speaks of being terrified and fearful without indicating the cause. We annotate such cases as latent causes. Sometimes the cause is revealed in future utterances, e.g., “someone is stalking me” as the reason of being fearful. However, as online settings would not have access to the future turns, we refrain from treating future spans as causes.

6 Experiments

We formulate two distinct subtasks of recognizing emotion cause in conversations: (i) causal span extraction and (ii) causal emotion entailment. However, note that the main purposes of this work are to present a dataset and setup the strong baselines.
6.1 Compiling Dataset Splits

RECCON-DD is the subset of our dataset that contains dialogues from DailyDialog. For this subset, we created the training, validation, and testing examples based on the original splits in [Li et al., 2017]. However, this resulted in the validation and testing sets to be quite small, so we moved some dialogues to them from the original training set.

The subset RECCON-IE consists of dialogues from the IEMOCAP dataset. This subset is quite small as it contains only sixteen unique dialogues (situations). So, we consider the entire RECCON-IE as another testing set, emulating an out-of-distribution generalization test. We report results on this dataset based on models trained on RECCON-DD. In our experiments, we ignore the utterances with only latent emotion causes.

6.1.1 Generating Negative Examples

The annotated dataset, RECCON (consisting of subsets RECCON-DD and RECCON-IE) only contains positive examples, where an emotion-containing target utterance is annotated with a causal span extracted from its conversational historical context. However, to train a model for the recognizing emotion cause in conversations task, we need negative examples, i.e., the instances which are not cause of the utterance. In the sequel, we use the terminology introduced in Section 3; the reader should refer to that section for clearer understanding.

We use three different strategies to create the negative examples. In this section, we will discuss in detail Fold 1. Then, in Section 7 to further analyze the performance of our models, besides Fold 1, we will adopt two more strategies, Fold 2 and Fold 3, to create the negative examples:

**Fold 1:** Consider a dialogue $D$ and a target utterance $U_t$ in $D$. We construct the complete set of negative examples as $\{(U_t, U_i) | U_i \in H(U_t) \setminus C(U_t)\}$, where $H(U_t)$ is the conversational history and $C(U_t)$ is the set of causal utterances for $U_t$.

**Fold 2:** In this scheme, we randomly sample the non-causal utterance $U_i$ along with the corresponding historical conversational context $H(U_i)$ from another dialogue in the dataset to create a negative example.

**Fold 3:** This is similar to Fold 2 with a constraint. In this case, a non-causal utterance $U_i$ along with its historical conversational context $H(U_i)$ from the other dialogue is only sampled when its emotion matches the emotion of the target utterance $U_t$ to construct a negative example.

Note that unlike Fold 1, a negative example in Fold 2 and 3 comprising a non-causal utterance $U_i$ and a target utterance $U_t$ belong to different dialogues. For the cases where the causal spans do not lie in the target utterance, we remove the target utterance from its historical context when creating a positive example in Fold 2 and 3. As a result, it helps to prevent the models from learning any trivial patterns. The statistics for the three folds are shown in Table 3.
Table 3: The statistics of RECCON comprising both positive (valid) and negative (invalid) UCS pairs. DD stands for RECCON-DD, IE for RECCON-IE. Utterances with only latent emotion causes are ignored in our experiments.

6.2 Subtask 1: Causal Span Extraction

Causal Span Extraction is the task of identifying the causal span (emotion cause) for a target non-neutral utterance. In our experimental setup, we formulate Causal Span Extraction as a Machine Reading Comprehension (MRC) task similar to the task in Stanford Question Answering Dataset (Rajpurkar et al., 2016). Similar MRC techniques have been used in literature for various NLP tasks such as named entity recognition (Li et al., 2020) and zero shot relation extraction (Levy et al., 2017). In this work, we propose two different span extraction settings: (i) with conversational context and (ii) without conversational context.

6.2.1 Subtask Description

With Conversational Context (w/ CC) We believe that the presence of conversational context would be key to the span extraction algorithms. To evaluate this hypothesis, we design this subtask, where the conversational history is available to the model. In this setup, for a target utterance $U_t$, the causal utterance $U_i \in C(U_t)$, and a causal span $S \in CS(U_t)$ from $U_i$, we construct the context, question, and answer as follows:

**Context:** The context of a target utterance $U_t$ is the conversational history, i.e., a concatenation of all utterances from $H(U_t)$. Similarly, for a negative example $(U_t, U_i)$, where $U_i \notin C(U_t)$, conversational history of $U_t$ is used as context.

---

2 By “causal span from evidence in the context” we mean a causal span from the conversation history $H(U_t)$. 
**Question:** The question is framed as follows: “The target utterance is \(<U_t>\). The evidence utterance is \(<U_i>\). What is the causal span from evidence in the context that is relevant to the target utterance’s emotion \(<E_t>\)?”

**Answer:** The causal span \(S \in CS(U_t)\) appearing in \(U_i\) if \(U_i \in C(U_t)\). For negative examples, \(S\) is assigned an empty string.

If a target utterance has multiple causal utterances and causal spans, then we create separate (Context, Question, Answer) instances for them. Unanswerable questions are also created from invalid (cause, utterance) pairs following the same approaches explained in Section 6.1.

**Without Conversational Context (w/o CC)** In this formulation, we intend to identify whether the Causal Span Extraction task is feasible when we only have information about the target utterance and the causal utterance. Given a target utterance \(U_t\) with emotion label \(E_t\), its causal utterance \(U_i \in C(U_t)\), and the causal span \(S \in CS(U_t)\), the question is framed as framed as follows: “The target utterance is \(<U_t>\). What is the causal span from context that is relevant to the target utterance’s emotion \(<E_t>\)?”. The task is to extract answer \(S \in CS(U_t)\) from context \(U_i\). For negative examples, \(S\) is assigned an empty string.

**6.2.2 Models**

We use two pretrained Transformer-based models to benchmark the Causal Span Extraction task.

*RoBERTa Base* We use the roberta-base model (Liu et al., 2019) and add a linear layer on top of the hidden-states output to compute span start and end logits. Scores of candidate spans are computed following Devlin et al. (2019), and the span with maximum score is selected as the answer.

*SpanBERT Fine-tuned on SQuAD* We use SpanBERT (Joshi et al., 2020) as the second baseline model. SpanBERT follows a different pre-training objective compared to RoBERTa (e.g. predicting masked contiguous spans instead of tokens) and performs better on question answering tasks. In this work we are using the SpanBERT base model fine-tuned on SQuAD 2.0 dataset.

**6.2.3 Evaluation Metrics**

We use the following evaluation metrics. **EM\(_{Pos}\)** *(Exact Match)*: EM represents, with respect to the gold standard data, how many causal spans are exactly extracted by the model. **F\(_1\)\(_{Pos}\)**: This is the F1 score introduced by Rajpurkar et al. (2015) to evaluate predictions of extractive QA models and calculated over positive examples in the data. **F\(_1\)\(_{Neg}\)**: Negative F1 represents the F1 score of detecting negative examples with respect to the gold standard data. Here, for a target utterance \(U_t\), the ground truth are empty spans. **F\(_1\)**: This metric is similar to F1\(_{Pos}\), but calculated for every positive and negative example followed by an average over them.

While all these metrics are important for evaluation, we stress that future works should particularly consider performances for EM\(_{Pos}\), F\(_1\)\(_{Pos}\), and F\(_1\).
## 6.3 Subtask 2: Causal Emotion Entailment

The **Causal Emotion Entailment** is a simpler version of the span extraction task. In this task, given a target non-neutral utterance ($U_t$), the goal is to predict which particular utterances in the conversation history $H(U_t)$ are responsible for the non-neutral emotion in the target utterance. Following the earlier setup, we formulate this task with and without historical conversational context.

### 6.3.1 Subtask Description

We consider the following two subtasks:

**With Conversational Context (w/ CC)** We consider the historical conversational context $H(U_t)$ of the target utterance $U_t$ and posit the problem as a triplet classification task: the tuple $(U_t, U_i, H(U_t))$ is aimed to be classified as positive, $U_i \in C(U_t)$. For negative examples, the tuple $(U_t, U_i, H(U_t))$ should be classified as negative for $U_i \notin C(U_t)$.

**Without Conversational Context (w/o CC)** We posit this problem as a binary sentence pair classification task, where $(U_t, U_i)$ should be classified as positive as $U_i \in C(U_t)$. For the negative example $(U_t, U_i)$ where $U_i \notin C(U_t)$, the classification output should be negative.

### 6.3.2 Models

In this paper we consider the following models.
Table 5: Results for Causal Emotion Entailment task on the test sets of RECCON-DD and RECCON-IE. Class-wise F1 score and the overall macro F1 scores are reported. All scores reported at best macro F1 scores. All models are RoBERTa-based.

The cause-pivot emotion extraction setting was used for ECPE-MLL. DD stands for RECCON-DD, IE for RECCON-IE.

| Model     | w/o CC | w/ CC |
|-----------|--------|-------|
|           | Pos. F1 | Neg. F1 | macro F1 | Pos. F1 | Neg. F1 | macro F1 |
| **Fold 1** |         |        |          |         |        |          |
| DD        | Base    | 56.64  | 85.13    | 70.88   | 64.28  | 88.74    | 76.51   |
|           | Large   | 50.48  | 87.35    | 68.91   | **66.23** | 87.89    | **77.06** |
|           | ECPE-MLL| –      | –        | –       | 48.48  | 94.68    | 71.59   |
|           | ECPE-2D | –      | –        | –       | 55.50  | 94.96    | 75.23   |
|           | RankCP  | –      | –        | –       | 33.00  | **97.30** | 65.15   |
| IE        | Base    | 25.98  | 90.73    | 58.36   | 28.02  | 95.67    | 61.85   |
|           | Large   | **32.34** | **95.61** | **63.97** | **40.83** | **95.68** | **68.26** |
|           | ECPE-MLL| –      | –        | –       | 20.23  | 93.55    | 57.65   |
|           | ECPE-2D | –      | –        | –       | 28.67  | 97.39    | 63.03   |
|           | RankCP  | –      | –        | –       | 15.12  | 92.24    | 54.75   |
| **Fold 1** |         |        |          |         |        |          |
| DD        | Base    | 93.12  | –        | –       | 92.64  | –        | –       |
|           | Large   | **98.87** | –        | –       | **97.78** | –        | –       |
|           | ECPE-MLL| –      | –        | –       | 84.50  | –        | –       |
|           | ECPE-2D | –      | –        | –       | 88.13  | –        | –       |
|           | RankCP  | –      | –        | –       | 85.67  | –        | –       |
| IE        | Base    | 71.98  | –        | –       | 58.52  | –        | –       |
|           | Large   | **73.92** | –        | –       | **74.56** | –        | –       |
|           | ECPE-MLL| –      | –        | –       | 66.45  | –        | –       |
|           | ECPE-2D | –      | –        | –       | 64.33  | –        | –       |
|           | RankCP  | –      | –        | –       | 70.21  | –        | –       |

*RoBERTa Base and Large* Similar to subtask 1, we use Transformer-based models to benchmark this task. We use a `<CLS>` token and the emotion label `<Et>` of the target utterance $U_t$ in front, and join the pair or triplet elements with `<SEP>` in between to create the input. The classification is performed from the corresponding final layer vector of the `<CLS>` token. We use the roberta-base/-large models from [Liu et al., 2019] as the baselines.

*ECPE-2D* [Ding et al., 2020a] proposed an end-to-end approach for emotion cause pair extraction. They use a 2D Transformer network to improve interaction among the utterances.

*ECPE-MLL* [Ding et al., 2020b] introduced a joint multi-label approach for emotion cause pair extraction. Specifically, the joint framework comprises two modules: (i) extraction of causal utterances for the target emotion utterance, (ii) extraction of emotion utterance for a causal utterance. Both these modules were trained using a multi-label training scheme.

*RankCP* [Wei et al., 2020] proposed an end-to-end emotion cause pair extraction where first the utterance pairs are ranked and then a one-stage neural approach is
applied for inter-utterance correlation modeling that enhances the emotion cause extraction. Specifically, they apply graph attentions to model the interrelations between the utterances in a dialogue. ECPE-2D, ECPE-MLL, and RankCP use RoBERTa-base as a sentence encoder in our implementation to facilitate a fair comparison.

### 6.3.3 Evaluation Metrics

We use F1 score for both positive and negative examples, denoted as Pos. F1 and Neg. F1 respectively. We also report the overall macro F1.

### 6.4 Results and Discussions

Table 4 shows the results of the causal span extraction task where SpanBERT obtains the best performance in both RECCON-DD and RECCON-IE. SpanBERT outperforms RoBERTa Base in EM$_{Pos}$ and F1$_{Pos}$ metrics. However, the performance of SpanBERT is worse for negative examples, which consequently results in a lower F1 score compared to RoBERTa Base model in both the datasets under “w/o CC” setting. Contrary to this, the performance of the SpanBERT in the presence of context (w/ CC) is consistently higher than RoBERTa Base with respect to all the metrics in RECCON-DD.

In Table 5, we report the performance of the Causal Emotion Entailment task. Under the “w/o CC” setting, in Fold 1, RoBERTa Base outperforms RoBERTa Large by 2% in RECCON-DD. In contrast to this, in RECCON-IE, RoBERTa Large performs better and beats RoBERTa Base by 5.5% in Fold 1. On the other hand, RoBERTa Large outperforms RoBERTa Base in both RECCON-DD and RECCON-IE under the “w/ CC” setting. The performance in RECCON-IE is consistently worse than in RECCON-DD under various settings in both subtask 1 and 2. We reckon this can be due to multiple reasons mentioned in Section 4.1, making the task harder on the IEMOCAP split.

We have also analyzed the performance of the baseline models on the utterances having one or multiple causes. The models consistently perform better for the utterances having only one causal span compared to the ones having multiple causes (+7% on an average calculated over all the settings and models). In the test data of Fold 1, approximately 38% of the UCS pairs (which we call as Fold T) have their causal spans lie within the target utterances. In Table 4 and 5, we report the results on Fold T. According to these results, the models perform significantly better on such UCS pairs under all the settings in both the subtasks. The models leverage contextual information for both the subtasks in the “w/ CC” setting which substantially improves the performance of the non-contextual (refer to the “w/o CC” setting) counterpart. In this setting, SpanBERT obtains the best performance for positive examples in both RECCON-DD, and RECCON-IE. On the other hand, in the same setting, RoBERTa Large outperforms RoBERTa Base and achieves the best performance in subtask 2.

The low scores of the models in subtasks 1 and 2 show the difficulty of the tasks. This implies significant room for model improvement in these subtasks of recognizing emotion cause in conversations. Table 5 shows that all the complex neural
baselines, i.e., ECPE-MLL, ECPE-2D, and RankCP fail to outperform the very simple RoBERTa baselines introduced in this paper. This corroborates the usefulness and importance of these strong baselines, one of the major contributions of this paper.

7 Further Analysis and Discussion

For further insights into the performance of our models, we analyzed more strategies to create the negative examples: Folds 2 and 3; see Section 6.1.1 for their description.
The use of context (w/ CC) in the baseline models improves the results (see Tables 6 and 7) in Folds 2 and 3 as it highlights the contextual discrepancy or coherence between the target utterance and context which should strongly aid in identifying randomly generated negative samples from the rest. For the positive examples, we achieve a much better score in Folds 2 and 3 as compared with Fold 1 (see Tables 4 and 5) for both “w/o CC” and “w/ CC” constraints. However, this does not validate Folds 2 and 3 as better training datasets than Fold 1. We confirm this by training the models on Folds 2 and 3 and evaluating them on Fold 1. These two experiments are denoted as Fold 2 → Fold 1 and Fold 3 → Fold 1, respectively, and the corresponding results are reported in Tables 6 and 7. The outcomes of these experiments, as shown in Tables 6 and 7, show abysmal performance by the baseline models on the negative examples in Fold 1.

This may be ascribed to the fundamental difference between Fold 1 and Folds 2 and 3. Negative samples in Folds 2 and 3 are easily identifiable, as compared to Fold 1, as all the model needs to do to judge the absence of a causal span in the context is to detect the contextual incoherence of the target utterance with the context. Models fine-tuned on BERT and SpanBERT are expected to perform well at deciding contextual incoherence. Identifying negative samples in Fold 1, however, requires more sophisticated and non-trivial approach as the target utterances are, just as the positive examples, contextually coherent with the context. As such, a model that correlates contextual incoherence with negative samples naturally performs poorly on Fold 1.

The $F^1_{Neg}$ scores for Fold 2 → Fold 1, and Fold 3 → Fold 1 modes under both “w/o CC” and “w/ CC” settings are adversely affected by the low precision of the models in both the subtasks. In other words, the baseline models in these two modes perform poorly in extracting empty spans from the ground truth negative examples in subtask 1 and also classify most of the negative examples as positive in subtask 2.

On the other hand, we do not observe any significant performance drop for either negative or positive examples when the models trained in Fold 1 are evaluated in Folds 2 and 3. This affirms the superiority of Fold 1 as a training dataset. Besides, note that Fold 1 is a more challenging and practical choice than the rest of the two folds as in real scenarios, we need to identify causes of emotions within a single dialogue by reasoning over the utterances in it.

8 Challenges of the Task

This section identifies several examples that indicate the need for complex reasoning to solve the causal span extraction task. Abilities to accurately reason will help validate if a candidate span is causally linked to the target emotion. We believe these pointers would help further research on this dataset and solving the task in general.

Amount of Spans One of the primary challenges of this task is determining the set of spans that can sufficiently be treated as the cause for a target emotion. The spans should have coverage to be able to formulate logical reasoning steps (performed im-
### Table 7: Results for Causal Emotion Entailment task on the test sets of RECCON-DD and RECCON-IE. Class wise F1 scores and the overall macro F1 scores are reported. All scores reported at best macro F1 scores. DD stands for RECCON-DD, IE for RECCON-IE. All models are RoBERTa-based models. Fold $i \rightarrow$ Fold $j$ means trained on Fold $i$, tested on Fold $j$.

| Model   | w/o CC | w/ CC |
|---------|--------|-------|
|         | Pos. F1 | Neg. F1 | macro F1 | Pos. F1 | Neg. F1 | macro F1 |
| **Fold 1** |        |        |         |        |        |         |
| **DD**  |        |        |         |        |        |         |
| Base    | 56.64  | 85.13  | 70.88   | 64.28  | 88.74  | 76.51   |
| Large   | 50.40  | 87.35  | 68.91   | 66.23  | 87.89  | 77.06   |
| **IE**  |        |        |         |        |        |         |
| Base    | 25.98  | 90.73  | 58.36   | 28.02  | 95.67  | 61.85   |
| Large   | 32.34  | 95.61  | 63.97   | 40.83  | 95.68  | 68.26   |
| **Fold 2** |        |        |         |        |        |         |
| **DD**  |        |        |         |        |        |         |
| Base    | 57.50  | 82.71  | 70.11   | 59.06  | 86.91  | 72.98   |
| Large   | 56.13  | 88.33  | 72.23   | 60.09  | 88.00  | 74.04   |
| **IE**  |        |        |         |        |        |         |
| Base    | 32.60  | 89.99  | 61.30   | 27.14  | 94.16  | 60.65   |
| Large   | 36.61  | 94.60  | 65.60   | 37.59  | 94.63  | 66.11   |
| **Fold 3** |        |        |         |        |        |         |
| **DD**  |        |        |         |        |        |         |
| Base    | 57.52  | 82.72  | 70.12   | 49.30  | 79.27  | 64.29   |
| Large   | 56.04  | 88.28  | 72.16   | 60.63  | 88.30  | 74.46   |
| **IE**  |        |        |         |        |        |         |
| Base    | 33.24  | 90.30  | 61.77   | 23.83  | 92.97  | 58.40   |
| Large   | 36.55  | 94.59  | 65.57   | 37.87  | 94.69  | 66.28   |

*Emotional Dynamics* Understanding emotional dynamics in conversations is closely tied with emotion cause identification. As shown in our previous sections, many causal phrases in the dataset depend on the inter-personal event/concept mentions, emotions, and self-influences (sharing causes). We also observe that emotion causes may be present across multiple turns, thus requiring the ability to model long-term explicitly by annotators) that include skills such as numerical reasoning (see Fig. 4), among others.
Recognizing Emotion Cause in Conversations

Fig. 4: In this example, $P_B$, in utterance 12, is sad because of failing to negotiate the desired amount to sell a TV. While “the price is final” is a valid causal span, one also needs to identify the discussion where $P_A$ is ready to pay only $2000, which is significantly lesser than the originally quoted $2500.

Fig. 5: In this example, the emotion cause for utterance 2 may lie in phrases spoken by (and for) the counterpart ($P_A$) and not the target speaker ($P_B$) i.e., “flashy red lines” in $P_B$’s utterance points to the property of the “watch” that $P_A$ bought. One needs to infer such co-referential links to extract the correct causal spans.

information. Emotions of the contextual utterances help in this modeling. In fact, without the emotional information of the contextual utterances, our annotators found it difficult to annotate emotion causes in the dataset. Understanding cordial greetings, conflicts, agreements, and empathy are some of the many scenarios where contextual emotional dynamics play a significant role.

*Commonsense Knowledge* Extracting emotion causes in conversations comprises complex reasoning steps, and commonsense knowledge is an integral part of this process. The role of commonsense reasoning in emotion cause recognition is more evident when the underlying emotion the cause is latent. Consider the example below:

(1) $P_A$ (HAPPY): Hello, thanks for calling 123 Tech Help, I’m Todd. How can I help you?

$P_B$ (FEAR): Hello ? Can you help me ? My computer ! Oh man ...

In this case, $P_A$ is happily offering help to $P_B$. The cause of happiness in this example is due to the event “greeting” or intention to offer help. On the other hand, $P_B$ is fearful because of his/her broken computer. The causes of elicited emotions by both the speakers can only be inferred using commonsense knowledge.

*Complex Co-Reference* While in narratives, co-references are accurately used and often explicit, it is not the case in dialogues (see Fig. 5).
Fig. 6: In this example, the cause for the happy state of $P_B$ (utterance 6) is corroborated by three indicated spans. First, $P_B$ gets happy over receiving a “birthday present” (utterance 3) which is a “gold watch” (utterance 4). Then, the emotion evoked by the 4th utterance is propagated into $P_B$’s next utterance where it is confirmed that $P_B$ loves the gift (“I love it!”). Performing temporal reasoning over these three spans helps understand that $P_B$ is happy because of liking a present received as a birthday gift.

Exact vs. Perceived Cause  At times, the complex and informal nature of conversations prohibits the extraction of exact causes. In such cases, our annotators extract the spans that can be perceived as the respective cause. These causal spans can be rephrased to represent the exact cause for the expressed emotion. For example,

(2)  
$P_A$ (NEUTRAL): How can I help you Sir?.
$P_B$ (FRUSTRATED): I just want my flip phone to work—that’s all I need.

In this example, the cause lies in the sentence “I just want my flip phone to work”, with the exact cause meaning of “My flip phone is not working”. Special dialogue-act labels such as goal achieved and goal not-achieved can also be adopted to describe such causes.

From Cause Extraction to Causal Reasoning  Extracting causes of utterances involve reasoning steps. In this work, we do not ask our annotators to explain the reasoning steps pertaining to the extracted causes. However, one can still sort the extracted causes of an utterance according to their temporal order of occurrence in the dialogue. The resulting sequence of causes can be treated as a participating subset of the reasoning process as shown in Fig. 6. In the future, this dataset can be extended by including reasoning procedures. However, coming up with an optimal set of instructions for the annotators to code the reasoning steps is one of the major obstacles. Fig. 7 also demonstrates the process of reasoning where utterance 1 and 2 are the triggers of HAPPY emotion in the utterance 3. However, the reasoning steps that are involved to extract these causes can be defined as: $P_A$ is happy because his/her goal to participate
in the house open party is achieved after the confirmation of $P_B$ who will organize the house open party. This reasoning includes understanding discourse (Chakrabarty et al., 2019), logic and leveraging commonsense knowledge.

![Fig. 7: An example of emotional reasoning where the happiness in utterance 3 is caused by the triggers in utterances 1 and 2.](image)

More generally, emotion causal reasoning in conversations extends the task of identifying emotion cause to determining the function and explanation of why the stimuli or triggers evoke the emotion in the target utterance. Evidence utterance ($U_{tc}$): An utterance containing a span that is the target utterance’s emotion cause. As there can be multiple evidence utterances of $U_{tc}$, we represent the set all evidence utterances as $C_{Ut} = \{U_{tc}^c \mid c \leq t\}$ and $C_{Ut} \subseteq H_{Ut}$.

9 Connection to Interpretability of the Contextual Models

One of the advantages of identifying the causes of emotions in conversations is its role in interpreting a model’s predictions. We reckon two situations where emotion cause identification can be useful to verify the interpretability of the contextual emotion recognition models that rely on attention mechanisms to count on the context:

- In conversations, utterances may not contain any explicit emotion bearing words or sound neutral on the surface but still carry emotions that can only be inferred from the context. In these cases, one can probe contextual models by dropping the causal utterances that contribute significantly to evoke emotion in the target utterance. It would be interesting to observe whether the family of deep networks that rely on attention mechanisms for context modeling e.g., transformer assign higher probability scores to causal contextual utterances in order to make correct predictions.

- As discussed in Section 5, the cause can be present in the target utterance and the model may not need to cater contextual information to predict the emotion. In such cases, it would be worth checking whether attention-based models assign high probability scores to the spans in the target utterance that contribute to the causes of its emotion.

One should also note that a model does not always need to identify the cause of emotions to make correct predictions. For example,
(3) $P_A$ (HAPPY): Germany won the match!
$P_B$ (HAPPY): That’s great!

Here, a model can predict the emotion of $P_B$ by just leveraging the cues present in the corresponding utterance. However, the utterance by $P_B$ is just an expression and the cause of the emotion is an event “Germany won the match”. Nonetheless, identifying the causes of emotions expressed in a conversation makes the model trustworthy, interpretable, and explainable.

10 Conclusion

We have addressed the problem of Recognizing Emotion Cause in CONversations and introduced a new dialogue-level dataset, RECON, containing more than 1,126 dialogues (dyadic conversations) and 10,600 utterance causal span pairs. We identified various emotion types and key challenges that make the task extremely challenging. Further, we proposed two subtasks and formulated Transformer-based strong baselines to address these subtasks.

Future work will target the analysis of emotion cause in multi-party settings. We also plan to annotate the reasoning steps involved in identifying causal spans of elicited emotions in conversations. Another direction of future work is to extend the approach to multi-modal setting, both in terms of transferring our annotation to the multi-modal data where such data are available (the part of our dataset extracted from IEMOCAP) and in terms of the benchmark algorithms.

Conflict of interest

The authors declare that they have no conflict of interest.

Compliance with Ethical Standards

– This article does not contain any studies with human participants or animals performed by any of the authors.
– All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

Ameer I, Ashraf N, Sidorov G, Adorno HG (2020) Multi-label emotion classification using content-based features in Twitter. Computación y Sistemas 24(3):1159–1164, DOI 10.13053/CyS-24-3-3476
Branden A, Verberne S, Wansleeben M, Lambers K (2020) Creating a dataset for named entity recognition in the archaeology domain. In: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, pp 4573–4577, URL https://www.aclweb.org/anthology/2020.lrec-1.562
Bullo C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS (2008) IEMOCAP: Interactive emotional dyadic motion capture database. Language Resources and Evaluation 42(4):335–359
Chakrabarty T, Hidey C, Muresan S, McKeown K, Hwang A (2019) AMPERSAND: Argument mining for PERSuAsive oNline discussions. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, pp 2933–2943, DOI 10.18653/v1/D19-1291, URL https://www.aclweb.org/anthology/D19-1291
Chen X, Li Q, Wang J (2020) Conditional causal relationships between emotions and causes in texts. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, pp 3111–3121, DOI 10.18653/v1/2020.emnlp-main.252, URL https://www.aclweb.org/anthology/2020.emnlp-main.252
Chen Y, Lee SYM, Li S, Huang CR (2010) Emotion cause detection with linguistic constructions. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), Coling 2010 Organizing Committee, Beijing, China, pp 179–187, URL https://www.aclweb.org/anthology/C10-1021
Chen Y, Hou W, Cheng X, Li S (2018) Joint learning for emotion classification and emotion cause detection. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 – November 4, 2018, Association for Computational Linguistics, pp 646–651, DOI 10.18653/v1/d18-1066, URL https://doi.org/10.18653/v1/d18-1066
Choi Y, Cardie C, Riloff E, Patwardhan S (2005) Identifying sources of opinions with conditional random fields and extraction patterns. In: HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada, The Association for Computational Linguistics, pp 355–362, URL https://www.aclweb.org/anthology/H05-10457
Colneric N, Demsar J (2018) Emotion recognition on twitter: comparative study and training a unison model. IEEE Transactions on Affective Computing
Das D, Bandyopadhyay S (2010) Finding emotion holder from Bengali blog texts—An unsupervised syntactic approach. In: Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation, Institute of Digital Enhancement of Cognitive Processing, Waseda University, Tohoku University, Sendai, Japan, pp 621–628, URL https://www.aclweb.org/anthology/Y10-1071
Devlin J, Chang M, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019. Volume 1 (Long and Short Papers), Association for Computational Linguistics, pp 4171–4186, DOI 10.18653/v1/n19-1423, URL https://doi.org/10.18653/v1/n19-1423

Ding Z, Xia R, Yu J (2020a) ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In: Jurafsky D, Chail J, Schluter N, Tetreault JR (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, pp 3161–3170, DOI 10.18653/v1/2020.acl-main.288, URL https://doi.org/10.18653/v1/2020.acl-main.288

Ding Z, Xia R, Yu J (2020b) End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In: Webber B, Cohn T, He Y, Liu Y (eds) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, pp 3574–3583, DOI 10.18653/v1/2020.emnlp-main.290, URL https://doi.org/10.18653/v1/2020.emnlp-main.290

Dragoni M, Donadello I, Cambria E (2021) OntoSenticNet 2: Enhancing reasoning within sentiment analysis. IEEE Intelligent Systems 36(5)

Ekman P (1993) Facial expression and emotion. American Psychologist 48(4):384

Ellsworth PC, Scherer KR (2003) Appraisal processes in emotion, Oxford University Press, pp 572–595

Gao Q, Jianan H, Ruifeng X, Lin G, He Y, Kong K, Lu Q (2017) Overview of ntcir-13 eca task. In: Proceedings of the NTCIR-13 Conference

Ghazi D, Inkpen D, Szpakowicz S (2015) Detecting emotion stimuli in emotion-bearing sentences. In: Gelbukh AF (ed) Computational Linguistics and Intelligent Text Processing – 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II, Springer, Lecture Notes in Computer Science, vol 9042, pp 152–165, DOI 10.1007/978-3-319-18117-2_12, URL https://doi.org/10.1007/978-3-319-18117-2_12

Ghosal D, Majumder N, Mihalcea R, Poria S (2020) Utterance-level dialogue understanding: An empirical study. 2009.13902

Gui L, Yuan L, Xu R, Liu B, Lu Q, Zhou Y (2014) Emotion cause detection with linguistic construction in Chinese Weibo text. In: Zong C, Nie J, Zhao D, Feng Y (eds) Natural Language Processing and Chinese Computing – Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings, Springer, Communications in Computer and Information Science, vol 496, pp 457–464, DOI 10.1007/978-3-662-45924-9_42, URL https://doi.org/10.1007/978-3-662-45924-9_42

Gui L, Wu D, Xu R, Lu Q, Zhou Y (2016) Event-driven emotion cause extraction with corpus construction. In: EMNLP, World Scientific, pp 1639–1649

Izard CE (1992) Basic emotions, relations among emotions, and emotion-cognition relations. Psychological Review 99(3):561–565, DOI 10.1037/0033-295X.99.3.
Joshi M, Chen D, Liu Y, Weld DS, Zettlemoyer L, Levy O (2020) SpanBERT: Improving pre-training by representing and predicting spans. [1907.10529]

Kratzwald B, Ilic S, Kraus M, Feuerriegel S, Prendinger H (2018) Decision support with text-based emotion recognition: Deep learning for affective computing. arXiv preprint arXiv:180306397

Lee SYM, Chen Y, Huang CR (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. Association for Computational Linguistics, Los Angeles, CA, pp 45–53, URL https://www.aclweb.org/anthology/W10-0206

Levy O, Seo M, Choi E, Zettlemoyer L (2017) Zero-shot relation extraction via reading comprehension. In: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Association for Computational Linguistics, Vancouver, Canada, pp 333–342, DOI 10.18653/v1/K17-1034, URL https://www.aclweb.org/anthology/K17-1034

Li X, Feng J, Meng Y, Han Q, Wu F, Li J (2020) A unified MRC framework for named entity recognition. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, pp 5849–5859, DOI 10.18653/v1/2020.acl-main.519, URL https://www.aclweb.org/anthology/2020.acl-main.519

Li Y, Su H, Shen X, Li W, Cao Z, Niu S (2017) Dailydialog: A manually labelled multi-turn dialogue dataset. In: Kondrak G, Watanabe T (eds) Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 – Volume 1: Long Papers, Asian Federation of Natural Language Processing, pp 986–995, URL https://www.aclweb.org/anthology/I17-1099/

Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, DOI 10.2200/S00416ED1V01Y201204HLT016, URL https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:190711692

Moreno Jiménez LG, Torres Moreno JM (2020) LiSSS: A new corpus of literary Spanish sentences for emotions detection. Computación y Sistemas 24(3):1139–1147, DOI 10.13053/CyS-24-3-3474

Neviarouskaya A, Aono M (2013) Extracting causes of emotions from text. In: Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013, Asian Federation of Natural Language Processing / ACL, pp 932–936, URL https://www.aclweb.org/anthology/I13-1121/

Plutchik R (1982) A psychoevolutionary theory of emotions. Social Science Information 21(4-5):529–553, DOI 10.1177/053901882021004003

Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) Squad: 100,000+ questions for machine comprehension of text. [1606.05250]
Talmy L (2000) Toward a cognitive semantics, vol 2. MIT press
Wei P, Zhao J, Mao W (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, Association for Computational Linguistics, Online, pp 3171–3181. DOI 10.18653/v1/2020.acl-main.289, URL https://www.aclweb.org/anthology/2020.acl-main.289
Xia R, Ding Z (2019) Emotion-cause pair extraction: A new task to emotion analysis in texts. In: Korhonen A, Traum DR, Márquez L (eds) Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 – August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, pp 1003–1012, DOI 10.18653/v1/p19-1096, URL https://doi.org/10.18653/v1/p19-1096
Zajonc RB (1980) Feeling and thinking: Preferences need no inferences. American Psychologist pp 151–175