Recognizing Emotion Cause in Conversations

Soujanya Poria[1], **Navonil Majumder**[1], **Devamanyu Hazarika**[2*], **Deepanway Ghosal**[1*], **Rishabh Bhardwaj**[1], **Samson Yu Bai Jian**[1], **Romila Ghosh**[4], **Niyati Chhaya**[6], **Alexander Gelbukh**[3], **Rada Mihalcea**[5]

[1] Singapore University of Technology and Design, Singapore
[2] National University of Singapore, Singapore
[3] CIC, Instituto Politécnico Nacional, Mexico
[4] Independent researcher, India
[5] University of Michigan, USA
[6] Adobe Research, India

Abstract

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 dynamic among the interlocutors. To this end, we introduce the task of *recognizing emotion cause in conversations* with an accompanying dataset named RECCON. 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 of RECCON: 1) Causal Span Extraction and 2) Causal Emotion Entailment. The dataset is available at https://github.com/declare-lab/RECCON.

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; Colneriĉ and Demsar, 2018), due to its wide range of applications in opinion mining, recommender systems, healthcare, and other areas.

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, consider the following review of a smartphone, "*I hate the touchscreen as it freezes after 2-3 touches*".

*Equal contribution. Randomly ordered.

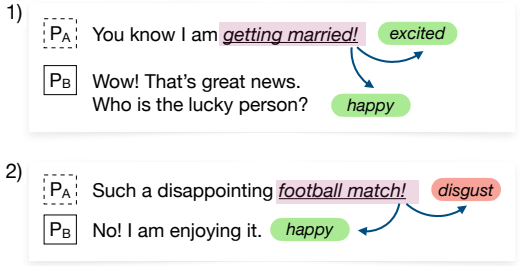


Figure 1: Emotion causes in conversations.

Understanding this text 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 is 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 (interpersonal emotional influence).

We introduce the task of **recognizing emotion cause in conversations**, 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 (conversational history). We formally define this task in §4.2.

In Fig. 1, we exemplify this task. In the first example, we are interested in knowing 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 could infer that P_B's emo-

tion 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 conversation, the cause of P_A's emotion is the event "*football match*" and a negative emotion `disgust` indicates P_A's unsatisfied experience of the match. In contrast, P_B takes pleasure of the match — sharing the same cause with P_A — with `happiness` emotion. These examples demonstrate the challenging problem of recognizing emotion causes in conversations, 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 (§5).

2. We describe a new annotated dataset for this task, RECCON[1], including both acted and real-world conversations (§4).

3. Further, we introduce two challenging sub-tasks that demand complex reasoning (§8), and provide the corresponding baselines (§6).

2 Related Work

Initial works in emotion analysis and opinion mining explored different aspects of affect beyond polarity prediction, such as identifying the opinion/emotion-feeler (or holder, source) (Das and Bandyopadhyay, 2010; Choi et al., 2005). However, the task of emotion cause extraction was studied later initially by Lee et al. (2010). Such initial works involved extracting cause events in a rule-driven manner (Chen et al., 2010). Gui et al. (2016) constructed an emotion cause extraction dataset by identifying events that trigger emotions. To avoid the latent emotions and implicit emotion causes associated with the informal text, the authors used news articles as the target corpus for cause extraction. Choosing news articles as the source data for cause extraction helped them reduce reasoning complexity for the annotators while extracting emotion causes. Ghazi et al. (2015) and Gao et al. (2017) are other notable works on Emotion Cause Extraction (ECE).

[1] pronounced as *reckon*.

Modifying 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 highlight 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.

Our work is a natural extension of these works. We propose a new dataset on conversations, which is more difficult to annotate and crack due to numerous challenges mentioned in the following sections (see §8), for example: 1) expressed emotions are not always explicit in the conversations; 2) conversations can be very informal where the phrase connecting emotion with its cause can often be implicit and thus needs to be inferred; 3) the stimuli of the elicited emotions can be located far from the target utterance in the conversation history and detecting it requires complex reasoning and co-reference often using commonsense.

3 Definiton 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 or the system. 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. Note that a text part can be both emotion evidence and cause.

We use the following terminologies throughout the paper. The **target utterance** U_t is the t^{th} utterance of a conversation, whose emotion label is known and whose emotion cause we want to identify. The **conversational history** $H(U_t)$ is the set of all utterances from the beginning of the conversation till the utterance U_t, including U_t. 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 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 §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 conversational 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 annotated with six emotions classes `happy`, `sad`, `neutral`, `anger`, `excited`, and `frustrated`. The dialogues in this dataset span across sixteen unique conversational situations. To avoid redundancy, we handpick only one dialogue from each of these situations. We denote the subset of RECCON comprising these dialogues as RECCON$_{IE}$.

DailyDialog is a natural human communication dataset covering various topics about our daily lives. All utterances are labeled with emotion categories `anger`, `disgust`, `fear`, `happy`, `neutral`, `sad`, and `surprise`. The dataset has over 83% `neutral` labels. Due to this skewness, we randomly selected dialogues which has at least four *non-neutral* utterances. We denote this subset of RECCON, comprising the dialogues from DailyDialog, as RECCON$_{DD}$. Some statistics about the annotated dataset is shown in Table 2.

Need for sampling from two different datasets. Although both IEMOCAP and DailyDialog are annotated with utterance-level emotions, they differ in many aspects. Firstly, the average number of utterances per dialogue in IEMOCAP is more than 50, whereas DailyDialog has a shorter average length of 8. Secondly, the shifts between non-neutral emotions (e.g., sad to anger, happy to excited) are more frequent in IEMOCAP compared to 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% utterances have their emotion cause in utterances at least 3 timestamps distant in the contextual history. On the contrary, this percentage is just 13% in our annotated DailyDialog dataset.

4.2 Annotation Process

Annotation guidelines. Given an utterance U_t labeled with an emotion E_t, the annotators were asked to extract the set of *causal spans* $CS(U_t)$ from the conversational history $H(U_t)$ (including utterance U_t) that sufficiently represent the causes of the emotion E_t. If the cause of E_t is latent, i.e., there is no explicit causal span in $H(U_t)$, the annotators wrote down the assumed causes that they inferred from $H(U_t)$. Each utterance was annotated by three human experts—graduate students with reasonable knowledge of the task. In fact, the annotators also flagged the utterances with explicit emotion causal spans that occur in the conversational future with respect to U_t. However, there were only seven such instances in the dataset which is too few for supervised learning. As such, we discarded those spans.

Emotional expression. An utterance can contain 1) a description of the triggers or stimuli of the expressed emotion, and / or 2) a reactionary emotional expression. In our setup, by following the discrimination among emotion evidence and cause as explained above, we instructed the annotators to look beyond emotional expressions and strive for identifying the actual emotion cause. We demonstrate one such case in Fig. 2c, where P_A explains the cause for `happiness` and the same cause evokes the emotion `excited` in P_B. Meanwhile,

the utterance 2 by P_B is merely an emotional expression. 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*".

Why span detection? Firstly, 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 towards the generated emotions (Talmy, 2000). We extend this popular formalism into a multi-span framework. Secondly, while recognizing emotion cause is driven by multiple controlling variables (see Appendix A), 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 (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, utterance-level and span-level aggregation.

Stage 1: Utterance-level aggregation. Whether an utterance is considered a causal utterance is determined by majority voting. A fourth expert annotator is brought in as the tie breaker.

Stage 2: Span-level aggregation. Per causal utterance, the union of the causal spans from the utterance-level majority annotators is taken as the final causal span if those annotator-specific causal spans share some sub-span. In case they do not share a sub-span a fourth annotator is brought in to determine the final span from the existing spans. This fourth annotator is also instructed to prefer the shorter spans over the longer ones that can sufficiently represent the cause without losing any information.

The fourth annotator could not break the tie for 34 causal utterances which we discarded from the dataset.

4.3 Dataset Statistics

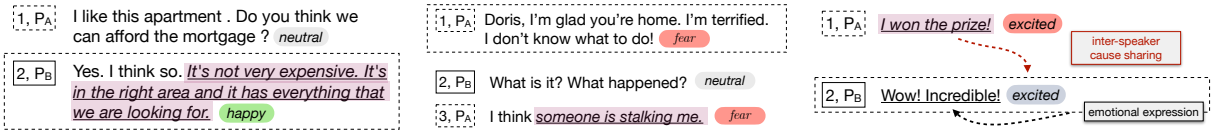
We have measured two types of inter-annotator agreement scores — i) at the utterance-level, and ii) at the span-level. Following Gui et al. (2016), we measured the inter-annotator agreement (IAA) at the utterance level, resulting in a kappa score of 0.7928. However, as pointed out by Brandsen et al. (2020), macro F1-measure is a more appropriate approach 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. This gives us a 0.8839 macro F1 score. Brandsen et al. (2020) also suggest the removal of 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, adopting this yields a lower F1 score of 0.8201. At span level, the F1 metric, as explained in Rajpurkar et al.

Dataset	Language	Source	Size	Format
Neviarouskaya and Aono (2013)	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	2105	Clauses
Gao et al. (2017)	Chinese/ English	SINA city news/ /English Novel	2619 / 2403	Clauses
RECCON (Ours)	English	IEMOCAP/ DailyDialog	1154 / 9915	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).

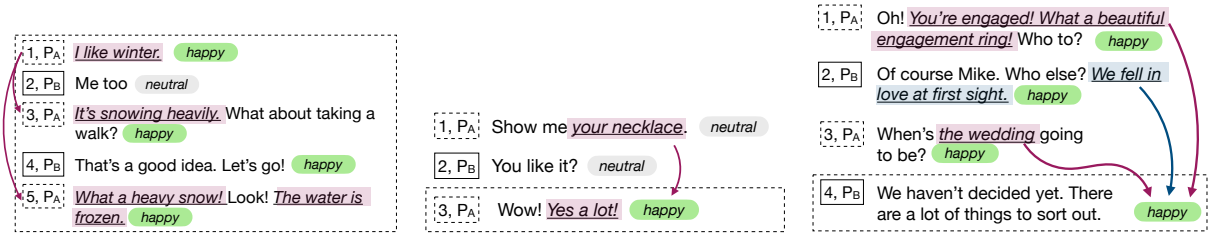
Description	RECCON$_{DD}$	RECCON$_{IE}$
# Dialogues	1106	16
# Utterances	11104	665
# Utterances Annotated with Emotion Cause	5861	494
# Utterances cater to background evidence	395	70
# Utterances where cause solely lies in the same utterance	1521	80
# Utterances where cause includes the same utterance along with contextual utterances	3370	243
# Utterance with emotion Anger	451	89
# Utterance with emotion Fear	74	-
# Utterance with emotion Disgust	140	-
# Utterance with emotion Frustration	-	109
# Utterance with emotion Happy	4361	58
# Utterance with emotion Sad	351	70
# Utterance with emotion Surprise	484	-
# Utterance with emotion Excited	-	197
# Utterance with emotion Neutral	5243	142
# UCS pairs	9915	1154
# Utterances having single cause	55%	41%
# Utterances having two causes	31%	24%
# Utterances having three causes	9%	17%
# Utterances having more than three causes	5%	18%
# Causes per utterance (Average)	1.69	2.34
# No Context	43%	35%
# Inter-Personal	32%	19%
# Self-Contagion	9%	20%
# Hybrid	11%	17%
# Latent	5%	10%
# Utterances (U_t) having cause at $U_{(t-1)}$	2851	183
# Utterances (U_t) having cause at $U_{(t-2)}$	1182	124
# Utterances (U_t) having cause at $U_{(t-3)}$	578	94
# Utterances (U_t) having cause at $> U_{(t-3)}$	769	200

Table 2: Statistics of the RECCON annotated dataset.



(a) (b) (c)

Figure 2: No context: 2a. Unmentioned Latent Cause: 2b. Distinguishing emotion cause from emotional expressions: 2c.



(a) Mood Setting (b) Generic Cause (c) Hybrid

Figure 3: *Self-contagion (3a,3b):* The cause of the emotion is primarily due to a stable mood of the speaker that was induced in the previous dialogue turns; *Hybrid (3c):* The hybrid type where the role of inter-personal emotional influence and self-contagion can be observed.

(2016), 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. 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.

5 Types of Emotion Causes

In our proposed 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 2b, 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 having felt 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. 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, we observe that the cause of the emotion is primarily due to a stable mood of the speaker that was induced in some previous dialogue turns. For example, 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 P_A 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/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 demonstrates such a case. Here, in first utterance, P_A speaks of being terrified and fearful without eliciting 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 using future spans as causal evidence.

6 Experiments

We formulate two distinct subtasks of *recognizing emotion cause in conversations*: i) Causal Span Extraction, ii) Causal Emotion Entailment.

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. Thus, we added some dialogues into them from the 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 Creating Negative Examples

The annotated dataset, RECCON (comprising 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 *recognizing emotion cause in conversations* task, we need negative examples, i.e., the instances which are not cause of the utterance. The remaining part of the paper heavily utilizes the terminologies mentioned in §3. We urge readers to refer to that section for a clearer understanding.

To this end, we adopt the following strategy 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 $\{(U_t, U_i) \mid U_i \in H(U_t) \setminus C(U_t)\}$, where $H(U_t)$ is the conversational history and $C(U_t)$ is the set of causal utterances for U_t.

We discuss the creation of Fold 2 and Fold 3 in §7. The statistic of the final dataset is shown in Table 3.

		Data	Train	Val	Test
Fold 1	DD	Positive UCS pairs	7269	347	1894
		Negative UCS pairs	20646	838	5330
	IEMO	Positive UCS pairs	-	-	1080
		Negative UCS pairs	-	-	11305
Fold 2	DD	Positive UCS pairs	7269	347	1894
		Negative UCS pairs	18428	800	4396
	IEMO	Positive UCS pairs	-	-	1080
		Negative UCS pairs	-	-	7410
Fold 3	DD	Positive UCS pairs	7269	347	1894
		Negative UCS pairs	18428	800	4396
	IEMO	Positive UCS pairs	-	-	1080
		Negative UCS pairs	-	-	7410

Table 3: The statistics of RECCON comprising both positive (valid) and negative (invalid) UCS pairs. DD → RECCON$_{DD}$; IEMO → 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). We propose two different span extraction tasks: i) With Conversational Context, and ii) Without Conversational Context.

6.2.1 Subtask Description

With Conversational Context (w/ CC) We speculate 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.

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 §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 where $U_i \in C(U_t)$, 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 the following 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

EM$_{Pos}$ (Exact Match): EM represents, with respect to the gold standard data, how many causal spans are exactly extracted by the model.

F1$_{Pos}$: This is the F1 metric introduced in (Rajpurkar et al., 2016) to evaluate predictions of extractive QA models and calculated over positive examples in the data.

F1$_{Neg}$: Negative F1 represents the F1 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 the above metrics are important for evaluation, we stress that future works should particularly consider performances for EM$_{Pos}$, F1$_{Pos}$, and F$_1$.

		Model	w/o CC				w/ CC			
			EM$_{Pos}$	F1$_{Pos}$	F1$_{Neg}$	F_1	EM	F1$_{Pos}$	F1$_{Neg}$	F_1
Fold 1	DD	RoBERTa	26.82	45.99	**84.55**	**73.82**	32.63	58.17	85.85	75.45
		SpanBERT	**33.26**	**57.03**	80.03	69.78	**34.64**	**60.00**	**86.02**	**75.71**
	IEMO	RoBERTa	9.81	18.59	**93.45**	**87.60**	10.19	26.88	**91.68**	**84.52**
		SpanBERT	**16.20**	**30.22**	87.15	77.45	**22.41**	**37.80**	90.54	82.86
Fold $\overline{1}$	DD	RoBERTa	37.76	63.87	-	-	39.02	69.13	-	-
		SpanBERT	41.96	72.01	-	-	42.24	71.91	-	-
	IEMO	RoBERTa	22.49	45.01	-	-	17.27	42.15	-	-
		SpanBERT	26.91	52.22	-	-	31.33	60.14	-	-

Table 4: Results for Causal Span Extraction task on the test sets of RECCON$_{DD}$ and RECCON$_{IE}$. All scores are in percentage and are reported at best validation F1 scores. DD → RECCON$_{DD}$; IEMO → RECCON$_{IE}$; RoBERTa → RoBERTa Base.

		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**
	IEMO	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 $\overline{1}$	DD	Base	93.12	-	-	92.64	-	-
		Large	98.87	-	-	97.78	-	-
	IEMO	Base	71.98	-	-	58.52	-	-
		Large	73.92	-	-	74.56	-	-

Table 5: Results for Causal Emotion Entailment task on the test sets of RECCON$_{DD}$ and RECCON$_{IE}$. Class wise F1 and the overall macro F1 scores are reported. All scores reported at best macro-F1 scores. DD → RECCON$_{DD}$; IEMO → RECCON$_{IE}$. All models are RoBERTa-based models.

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

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. Here the tuple (U_t, U_i, $H(U_t)$) is aimed to be classified as positive, $U_i \in C(U_t)$. For the negative example, the tuple (U_t, U_i, $H(U_t)$) should be classified as negative as $U_i \notin C(U_t)$.

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

Similar to Subtask 1, we use transformer-based models to benchmark this task. We use a <CLS> token and the emotion label $< E_t >$ 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 following models:

RoBERTa Base / Large : We use the `roberta-base/-large` models from (Liu et al., 2019) as the baselines.

6.3.3 Evaluation Metrics

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

6.4 Main Results

Table 4 reports the experimental 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 §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 $\overline{\text{Fold 1}}$) have their causal spans lie within the target utterances. In Table 4 and 5, we report the results on $\overline{\text{Fold 1}}$. 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 the subtask 1 and 2 depicts the difficulty of the tasks. As such, we see a significant room for model improvement in these two subtasks of *recognizing emotion cause in conversations*.

7 Analyses And Discussions

To further analyze the performance obtained by the models, besides Fold 1, we adopt two more strategies to create the negative examples.

1. **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.

2. **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

Model			w/o CC				w/ CC			
			EM_{Pos}	$F1_{Pos}$	$F1_{Neg}$	F_1	EM	$F1_{Pos}$	$F1_{Neg}$	F_1
Fold 1 → Fold 1	DD	RoBERTa	26.82	45.99	**84.55**	**73.82**	32.63	58.17	85.85	75.45
		SpanBERT	**33.26**	**57.03**	80.03	69.78	**34.64**	**60.00**	**86.02**	**75.71**
	IEMO	RoBERTa	9.81	18.59	**93.45**	**87.60**	10.19	26.88	**91.68**	**84.52**
		SpanBERT	**16.20**	**30.22**	87.15	77.45	**22.41**	**37.80**	90.54	82.86
Fold 1 → Fold 2	DD	RoBERTa	26.82	45.99	83.52	72.66	32.95	59.02	95.36	87.63
		SpanBERT	33.26	57.03	84.02	74.80	32.37	57.04	95.01	87.00
	IEMO	RoBERTa	9.81	18.59	92.18	85.41	10.93	28.26	95.49	90.85
		SpanBERT	16.20	30.22	88.63	79.80	24.07	40.57	96.28	92.41
Fold 1 → Fold 3	DD	RoBERTa	26.82	45.99	81.50	70.26	32.95	59.02	95.37	87.65
		SpanBERT	33.26	57.03	79.65	69.83	32.31	56.99	94.92	86.87
	IEMO	RoBERTa	9.81	18.59	91.82	84.83	10.93	28.26	95.47	90.81
		SpanBERT	16.20	30.22	86.95	77.25	24.07	40.57	96.28	92.41
Fold 2 → Fold 2	DD	RoBERTa	**33.26**	58.44	90.14	82.19	41.61	73.57	99.98	92.04
		SpanBERT	32.31	**58.61**	**90.20**	**82.29**	**41.97**	**74.85**	99.94	**92.43**
	IEMO	RoBERTa	15.93	31.74	**92.93**	**86.50**	30.28	59.14	**99.43**	94.58
		SpanBERT	**22.13**	**38.84**	90.37	82.49	**32.50**	**65.45**	98.37	**95.50**
Fold 2 → Fold 1	DD	RoBERTa	33.26	58.44	71.29	60.45	36.06	65.04	0.19	17.12
		SpanBERT	32.31	58.61	72.52	61.70	31.52	60.81	0.67	16.19
	IEMO	RoBERTa	15.93	31.74	90.70	82.91	22.96	46.87	4.66	6.35
		SpanBERT	22.13	38.84	85.03	74.34	21.85	49.18	6.36	7.40
Fold 2 → Fold 3	DD	RoBERTa	28.72	51.32	**90.06**	**82.11**	41.29	74.95	99.94	92.44
		SpanBERT	**30.62**	**54.96**	89.41	81.21	**42.61**	**75.36**	99.93	92.46
	IEMO	RoBERTa	14.54	26.51	**93.68**	**87.79**	24.35	53.46	97.84	94.08
		SpanBERT	**17.41**	**31.75**	91.85	84.86	**32.87**	**62.70**	**99.54**	**95.11**
Fold 3 → Fold 1	DD	RoBERTa	28.72	51.32	75.55	64.31	37.22	69.64	0.90	18.59
		SpanBERT	30.62	54.96	75.49	64.46	31.94	60.81	0.15	16.00
	IEMO	RoBERTa	14.54	26.51	92.33	85.61	21.20	48.34	11.42	9.76
		SpanBERT	17.41	31.75	89.41	80.94	21.48	45.49	4.01	5.84

Table 6: Results for Causal Span Extraction task on the test sets of RECCON_{DD} and RECCON_{IE}. All scores are in percentage and are reported at best validation F1 scores. RoBERTa → RoBERTa Base; DD → RECCON_{DD}; IEMO → RECCON_{IE}. Fold i → Fold j: 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 → 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**
	IEMO	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 1 → 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**
	IEMO	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 1 → 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**
	IEMO	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**
Fold 2 → Fold 2	DD	Base	76.21	91.23	83.72	89.37	95.21	92.32
		Large	**79.52**	**91.27**	**85.40**	**93.05**	**97.22**	**95.13**
	IEMO	Base	46.12	**93.80**	69.96	**65.09**	**95.60**	**80.35**
		Large	**48.36**	92.06	**70.21**	61.12	95.59	78.35
Fold 2 → Fold 1	DD	Base	**52.52**	**75.51**	**64.02**	41.86	3.25	22.55
		Large	51.57	67.58	59.57	**43.25**	**19.95**	**31.60**
	IEMO	Base	**31.51**	**92.09**	**61.80**	25.22	74.69	49.96
		Large	29.64	87.68	58.66	**26.30**	**76.44**	**51.37**
Fold 2 → Fold 3	DD	Base	74.73	**90.33**	**82.53**	92.64	96.99	94.81
		Large	**75.79**	88.43	82.11	**93.34**	**97.23**	**95.29**
	IEMO	Base	**51.23**	**93.70**	**72.46**	**63.91**	**94.55**	**79.23**
		Large	43.00	88.47	65.74	59.03	92.21	75.62
Fold 3 → Fold 1	DD	Base	**52.02**	**74.59**	**63.31**	41.64	2.99	22.31
		Large	51.53	65.76	58.65	**41.86**	**4.89**	**23.38**
	IEMO	Base	**34.74**	**91.46**	**63.10**	**19.13**	**54.25**	**36.69**
		Large	27.58	84.13	55.86	18.33	48.01	33.17

Table 7: Results for Causal Emotion Entailment task on the test sets of RECCON_{DD} and RECCON_{IE}. Class wise F1 and the overall macro F1 scores are reported. All scores reported at best macro-F1 scores. DD → RECCON_{DD}; IEMO → RECCON_{IE}. All models are RoBERTa-based models. Fold i → Fold j: Trained on Fold i, Tested on Fold j.

example in Fold 2 and 3. As a result, it helps to prevent the models from learning any trivial patterns.

The use of context (w/CC) in the baseline models improves the results (see Table 6 and 7) in Fold 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 Fold 2 and 3 as compared to Fold 1 (see Table 4 and Table 5) for both "w/o CC" and "w/ CC" constraints. However, this does not validate Fold 2 and 3 as better training datasets than Fold 1. We confirm this by training the models on Fold 2 and 3 and evaluating them on Fold 1. These two experiments are denoted with *Fold 2 → Fold 1* and *Fold 3 → Fold 1*, respectively, and the corresponding results are reported in Table 6 and 7. The outcomes of these experiments, as shown in Table 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 Fold 2, 3. Negative samples in Fold 2, 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 $F1_{Neg}$ 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 poor 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 Fold 2 and 3. This affirms the superiority of Fold 1 as a training dataset. Besides, note that Fold 1 is a more challenging and practi-

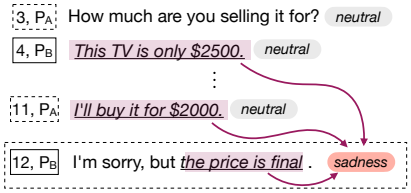


Figure 4: In this example, P_B, in utt. 12, is sad because of failing to negotiate the desired amount to sell a TV. While "*the price is final*" is a valid causal span, one also needs to identify the discussion where P_A is ready to pay only $2000, which is significantly lesser than the originally quoted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

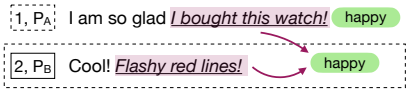


Figure 5: In this example, the emotion cause for utt. 2 may lie in phrases spoken by (and for) the counterpart (P_A) and not the target speaker (P_B) i.e., "*flashy red lines*" in P_B's utterance points to the property of the "*watch*" that P_A bought. One needs to infer such co-referential links to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

cal 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 in 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 implicitly by annotators) that include skills such as numerical reasoning (Fig. 4), amongst others.

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 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) *A (happy)*: Hello, thanks for calling 123 Tech Help, I'm Todd. How can I help you?
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).

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) *A (neutral)*: How can I help you Sir?.
B (frustrated): I just want my flip phone to work—-that's all I need.

In the above example, the cause lies in the following sentence — "*I just want my flip phone to work*", with the exact cause meaning — "*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

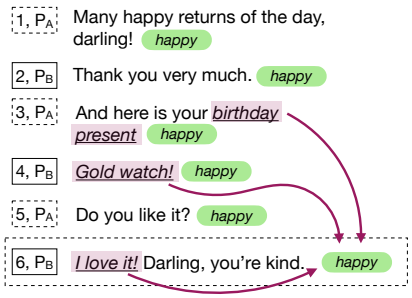


Figure 6: In this example, the cause for the happy state of P_B (utt. 6) is corroborated by three indicated spans. First, p_B gets happy over receiving a "*birthday present*" (utt. 3) which is a "*gold watch*" (utt.4). Then, the emotion evoked by the 4^{th} utterance is propagated into P_B's next utterance 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.

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.

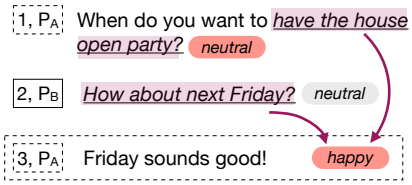


Figure 7: An example of emotional reasoning where the *happiness* in utt. 3 is caused by the triggers in utt. 1 and 2.

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.

9 Conclusion

In this work, we address the problem of **R**ecognizing **E**motion **C**ause in **CON**versations and introduce a new dataset — RECCON. It is a dialogue-level dataset containing more than $1,126$ dialogues and $10,600$ utterance causal span pairs. We identify various emotion types and key challenges that make the task of *recognizing emotion cause in conversations* extremely challenging. Further, we also propose two subtasks and formulate transformer-based strong baselines to address these tasks. Our proposed dataset only incorporates dyadic conversations. 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.

References

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a dataset for named entity recognition in the archaeology domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 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)*, pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Xinhong Chen, Qing Li, and Jianping Wang. 2020. Conditional causal relationships between emotions and causes in texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3111–3121, Online. Association for Computational Linguistics.

Ying Chen, Wenjun Hou, Xiyao Cheng, and Shoushan Li. 2018. Joint learning for emotion classification and emotion cause detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October*

31 - November 4, 2018, pages 646–651. Association for Computational Linguistics.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren 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.

Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 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*, pages 355–362. The Association for Computational Linguistics.

Niko Colneriĉ and Janez Demsar. 2018. Emotion recognition on twitter: comparative study and training a unison model. *IEEE Transactions on Affective Computing*.

Dipankar Das and Sivaji Bandyopadhyay. 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*, pages 621–628, Tohoku University, Sendai, Japan. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Phoebe C Ellsworth and Klaus R Scherer. 2003. *Appraisal processes in emotion*, pages 572–595. Oxford University Press.

Qinghong Gao, H Jiannan, X Ruifeng, Gui Lin, Yulan He, KF Wong, and Q Lu. 2017. Overview of ntcir-13 eca task. In *Proceedings of the NTCIR-13 Conference*.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II*, volume 9042 of *Lecture Notes in Computer Science*, pages 152–165. Springer.

Deepanway Ghosal, Navonil Majumder, Rada Mihalcea, and Soujanya Poria. 2020. Utterance-level dialogue understanding: An empirical study.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. 2016. Event-driven emotion cause extraction with corpus construction. In *EMNLP*, pages 1639–1649. World Scientific.

Lin Gui, Li Yuan, Ruifeng Xu, Bin Liu, Qin Lu, and Yu Zhou. 2014. Emotion cause detection with linguistic construction in chinese weibo text. In *Natural Language Processing and Chinese Computing - Third CCF Conference, NLPCC 2014, Shenzhen, China, December 5-9, 2014. Proceedings*, volume 496 of *Communications in Computer and Information Science*, pages 457–464. Springer.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans.

Bernhard Kratzwald, Suzana Ilic, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Decision support with text-based emotion recognition: Deep learning for affective computing. *arXiv preprint arXiv:1803.06397*.

Richard S Lazarus. 1982. Thoughts on the relations between emotion and cognition. *American psychologist*, 37(9):1019–1024.

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 Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Alena Neviarouskaya and Masaki Aono. 2013. Extracting causes of emotions from text. In *Sixth International Joint Conference on Natural Language Processing, IJCNLP 2013, Nagoya, Japan, October 14-18, 2013*, pages 932–936. Asian Federation of Natural Language Processing / ACL.

Robert Plutchik. 1982. A psychoevolutionary theory of emotions. *Social Science Information*, 21(4-5):529–553.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

Leonard Talmy. 2000. *Toward a cognitive semantics*, volume 2. MIT press.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1003–1012. Association for Computational Linguistics.

R. B. Zajonc. 1980. Feeling and thinking: Preferences need no inferences. *AMERICAN PSYCHOLOGIST*, pages 151–175.

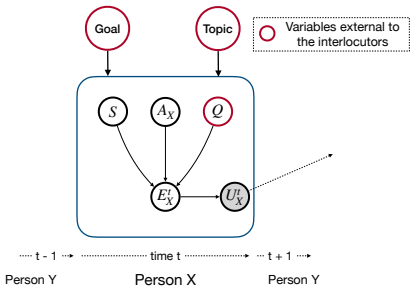


Figure 8: Emotions expressed in a dyadic conversation–between person X and Y–are governed by interactions between several variables.

A Emotion Causation Variables

Various factors can control elicited emotions in a conversation. We identify some of these factors, as stated below:

Topic: Topic is a key element that governs and drives a conversation. Without knowing the topical information, dialogue understanding can be incomplete and vague.

Goal: The goal of the speakers can directly impact the agency appraisal (Ellsworth and Scherer, 2003) of emotions, e.g., although it can be quite frustrating at times, a customer care agent generally tends to please the customers and try to be nice to them. Goals can often be very implicit in a conversation, and some conversations may not even have any goal, e.g., social casual talk. In this work, as the interlocutors' goals in a conversation are not available as prior information, we do not count on it. However, in several cases, our annotators could infer the goals from the conversational context and utilize that information to find relevant causal spans in the context.

Agency (A): Agency controls the basic behavior of a person under varied circumstances. According to the theory of appraisals (Ellsworth and Scherer, 2003) in affective computing, agency can also signify different dimensions, such as values, needs, personality, intents, and more. In our case, we ignore the role of agency in emotion causal reasoning as this information is not available in the dataset. Moreover, our dataset is already pre-annotated with emotion labels; hence the variable A does not play any role in this work.

Stimulus: Event, Situation, Experience, Statement, and Opinion (S). This variable is defined as the stimulus or trigger that evokes the emotion. The stimulus could refer to events, situations, opinions, experiences mentioned in the conversa-

tional context ($H(U_t)$) (by either of the speakers) or even be based on the counterpart's reaction towards an event cared by the speaker (inter-personal emotional influence).

Consider the following example where the first utterance by person A (P_A) is the context and the second utterance by person B (P_B) is the target. We are interested in knowing the cause of P_B's emotion (`excited` and `happy`) in this target utterance:

(3) *A (excited and happy)*: You know I am getting married!
B (excited and happy): Wow! that's great news. Who is that lucky person? When is the ceremony?

In the conversation, P_B listens to and positively reacts to the event "P_A *is getting married*". P_A also feels happy due to the same event – "*getting married*". Here, we could 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 emotions.

Another example is given below where the cause of P_A's emotion is the event "*football match*" and a negative emotion *disgust* indicates P_A's unsatisfied experience of the match. In contrast, P_B takes pleasure of the match with *happiness* emotion.

(4) *A (disgust)*: Such a disappointing football match!
B (happiness): No! I am enjoying it.

Interestingly in this example, the same event acts as trigger for both the persons and causes two contrasting emotions in them.

The conversations can be dynamic and emotions can be triggered from the situations induced in the conversations. Consider the example below:

(5) *A (neutral)*: How are you? You look a bit lost.
B (anger): Don't bother me.
A (disgust): Okay, I am going.

In this example, P_A feels disgusted in utterance 3 because of person B's angry and unexpected response. On another side, one can infer that P_B gets angry or annoyed to hear person A saying "You look a bit lost".

A stimulus can be latent too and may require the ability of commonsense inference to identify. Our annotators were instructed to identify these cases. When identified, our annotators wrote their understanding of the inferred cause in the form of natural language. Consider the example below:

(6) *A (happy)*: Hello, thanks for calling 123 Tech Help, I'm Todd. How can I help you?
B (fear): Hello ? Can you help me ? My computer ! Oh man ...

In this case, person A is happily offering help to customer B. The cause of happiness in this example is due to the event "greeting" or intention to offer help. On the other hand, person 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.

Awareness and Inclinations (Q): Q represents background knowledge, prior assumptions if any, pre-existing inter-speaker relations, speaker's knowledge and opinion about the topic, and any other background or external information that are not explicitly present in the conversational history. Such knowledge usually evolves depending on how the speaker experiences the environment and interacts with it. In the process of a conversation, certain sensory or other external events can directly initiate cognition and affect. We call these inputs as Q. These inputs can also be non-verbal cues.

Affective reactions to these sensory inputs can occur with or without any complex cognitive modeling. When the stimulus is sudden and unexpected, the affective reaction can occur before evaluating and appraising the situation through cognitive modeling. This is called *Affective Primacy* (Zajonc, 1980). For example, our immediate reaction when we encounter an unknown creature in the jungle without evaluating whether it is safe or dangerous. In our case, Q is unknown and needs to be guessed from the whole conversation.

If we refer to example 5, one can speculate that person A's opinion in utterance 1 may not be the sole reason for person B's anger. Person B may also be in a preexisting bad mood due to some prior incidents that are not captured in the course of the conversation.

Elicited emotion (E): E encodes the emotion of the speaker at time t. As proposed by the psychology theorist Lazarus in his article (Lazarus, 1982), the emotional-state can be triggered by cognition and thinking, we think in a conversation, this state can be controlled by Topic, Goal, S, P, and Q.

In this work, we identify the stimuli S that cause an expressed emotion in a conversation. We assume that these stimuli are either mentioned or can be inferred in the conversational context U.

B 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 explained in §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,

(7) *A (happy)*: Germany won the match!
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.

C Experimental Setup

C.1 Subtask 1: Causal Span Extraction

C.1.1 Subtask Description

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_c^t, and the causal span $S \in CS(U_t)$, we use the following scheme to create the context, question and answer:

- Context: The context is the utterance U_c^t where the causal span lies i.e., the causal utterance. In the case of negative example, context is a non-causal utterance U_n^t as explained in §6.1.1.

- Question: The question is 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 $< E_t >$?", where $< U_t >$ and $< E_t >$ are replaced by the plain-text utterance and emotion label.

- Answer: The causal span $S \in CS(U_t)$ appearing in U_c^t. In the case of negative example, $S \in CS(U_t)$ is assigned an empty span.

If a target utterance has multiple causal utterances and causal spans, then we create separate (Context, Question, Answer) instances for those corresponding utterances and spans.

Note that, in the w/o CC setup, we don't have any contextual information from the conversation apart from the target utterance U_t, and the causal utterance U_c^t (positive/valid examples), or non-causal utterance U_n^t (negative/invalid examples).

C.1.2 Evaluation Metrics

F1: This is the F1 metric introduced in (Rajpurkar et al., 2016) to evaluate predictions of extractive QA models. Here, both ground truth and the predicted spans are represented as bag of tokens and then the F1 measure is computed for each positive (Context, Question, Answer) instance. We then average the scores in the dataset level to obtain the final F1 score.