Multi-Task Learning Framework for Extracting Emotion Cause Span and Entailment in Conversations

Ashwani Bhat
Indian Institute of Technology Kanpur (IIT-K)
Kanpur, India
ashubhat44@gmail.com

Ashutosh Modi
Indian Institute of Technology Kanpur (IIT-K)
Kanpur, India
ashutoshm@cse.iitk.ac.in

Abstract

Predicting emotions expressed in text is a well-studied problem in the NLP community. Recently there has been active research in extracting the cause of an emotion expressed in text. Most of the previous work has done causal emotion entailment in documents. In this work, we propose neural models to extract emotion cause span and entailment in conversations. For learning such models, we use RECCON dataset, which is annotated with cause spans at the utterance level. In particular, we propose MuTEC, an end-to-end Multi-Task learning framework for extracting emotions, emotion cause, and entailment in conversations. This is in contrast to existing baseline models that use ground truth emotions to extract the cause. MuTEC performs better than the baselines for most of the data folds provided in the dataset.

1 Introduction

Emotions are an inherent part of human behavior. The choices and actions we make/take are directly influenced by the emotions we are experiencing at any particular moment. Emotions are indicative of and influence the underlying thought process [25]. Recent developments in AI have made machines an integral part of our lives. For seamless interaction with humans, it is imperative that AI systems understand the emotion experienced by a person and what are the causes and effects of such emotions [32]. Towards this goal in the past two decades, there has been significant research and progress in the area of emotion recognition [9]. To understand what influences/causes emotions and how the emotions of a person in turn influence others, recently, there has been active interest in the task of emotion cause extraction (ECE) in documents (§2). Poria et al. [27] have extended the task of emotion cause extraction to conversations by introducing a new task that requires extraction of the cause span corresponding to a given emotion utterance in a dialogue. The authors have released the **RECCON** (Recognizing Emotion Cause in CONversations) dataset, where conversations from DailyDialog [23] and IEMOCAP [3] datasets are annotated with cause span of the emotion utterance. Fig. 1 shows a sample conversational example showing the emotion cause. The highlighted portion of text represents the cause and the directed arrow A→B, represents that B contains the cause of A and hence is the cause utterance of A. Poria et al. [27] have introduced two challenging task on RECCON: *Causal Span Extraction* (CSE) and *Causal Emotion Entailment* (CEE) (§3). The authors used gold emotion annotations during inference. However, this is not a practical assumption. To address this, in this work, we make the following contributions:

- For CSE and CEE tasks, we propose an end-to-end Multi-Task learning framework for extracting emotions, emotion cause and cause entailment in conversations (MuTEC), where emotions are predicted as auxiliary task and cause span prediction and entailment are the main tasks. We also propose an overall end-to-end model architecture to solve both the tasks using a single architecture. Incorporating emotion prediction directly into the model gives comparable, and in some cases, better performance than models that explicitly use

Workshop on Transfer Learning for Natural Language Processing, 2022, New Orleans.

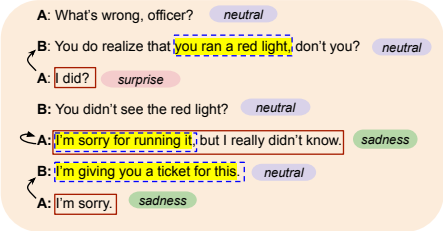


Figure 1: Example conversational dialogue from Dailydialog. Each utterance is provided with its corresponding emotion. The dashed rectangles represent the cause spans, and the arrow indicate the cause utterance for any target utterance (solid rectangle) under consideration. As shown, cause for an emotion can sometimes be within the same utterance.

gold emotion labels. We release the code for all the models and experiments: `https://github.com/Exploration-Lab/MuTEC`

- We perform a thorough analysis of the dataset and the models (§6). The original RECCON dataset is highly unbalanced with respect to the negative samples resulting in degradation in performance. We create a new version of a balanced dataset and perform experiments on it to show that reducing the negative samples helps to improve the model performance.

2 Related Work

Emotion prediction [19, 1, 33] and emotion generation [8, 15] are active areas of research. Emotion Cause Extraction (ECE) [6] is the problem of extracting cause of an emotion given emotion annotations. ECE task has attracted significant attention due to its wide applicability. ECE problem has been solved using classical machine learning-based methods [16], rule-based methods [29, 26], and deep learning methods [7, 41, 22, 21, 42, 43, 32]. However, the requirement of having gold emotion annotations at the test time has limited its usability in practical scenarios. Li et al. [21] experimented with removing the annotated emotion, but it led to significant performance drop for the ECE task. Another limitation of ECE task is that it is a two-step process. It first requires annotating the emotions and then extracting their cause, thus ignoring the mutual dependencies between the cause and the emotion.

To overcome the limitations of the ECE task, a new task was introduced by Xia and Ding [40]: Emotion Cause Pair Extraction (ECPE). This is a more challenging task aimed at extracting all cause and emotion pairs from the document. ECPE was introduced as a sentence pair classification task. ECPE task doesn't need emotion annotations to be provided at the test time. Also, since it extracts pair of emotion-cause, both the clauses are mutually indicative. To address this task, the authors proposed a two-step approach where they extracted the set of emotions and cause clauses individually in the first step and in the subsequent step, pair and filter the extracted clauses. This two-step approach suffers from 2 specific problems: (1) The errors from Step 1 are propagated to Step 2 and affect the performance of Step 2. (2) The training of the model is not directly aimed at extracting the final emotion-cause clause pair. To address the above issues, a need for an end-to-end architecture was realized. The first set of work for an end-to-end architecture was done by Ding et al. [11], where the authors used a representation scheme (in 2D) to represent emotion cause clause pairs and then integrated the cause and emotion pair interaction, prediction, and representation into a single combined framework. Song et al. [34] and Fan et al. [13] solved this problem using a graph-based approach to recognize emotions and their corresponding causes. Chen et al. [5] described this problem as a unified sequence labeling problem, where they extract emotion cause pairs using CNNs. In Ding et al. [12], the authors proposed a multi-label learning framework that extracted both the cause and emotion clauses where the windows for learning multiple labels is fixed on specific cause or emotion clause, and as the position of the clauses is moved, the window also slides. Wei et al. [39] used a ranking strategy where they ranked the emotion-cause clause pair candidates in a given document and modeled this inter-clause relationship using Graph Attention Network [36] to perform end-to-end pair extraction. Singh et al. [32] modeled the mutual interdependence between emotion clause and cause clause using neural networks and trained the entire NN in an end-to-end fashion. Recently, Sun et al. [35] argued the importance of context in order to extract emotion clause and cause clause and hence proposed a context-aware dual questioning attention network. Ding and Kejriwal [10] studied the effect of position bias on Emotion Cause Extraction. Another similar task, Emotion-Cause Span-Pair Classification and Extraction was proposed by Bi and Liu [2], in which instead of taking a definite

emotion and cause clause, they took random spans of text from the document that may span across multiple clauses.
Recently, emotion classification in conversations has been an active research area. Wang et al. [38], Shen et al. [30], Chapuis et al. [4] use transformer-based architectures to recognize emotions. Sheng et al. [31], Ghosal et al. [14], Zhong et al. [44] use graph neural networks and sequence-based networks to model the relationship between utterances and recognize the emotions. Ishiwatari et al. [18] use both contextual embeddings from transformer-based models and graph neural networks to recognize the emotions.

3 RECCON Tasks

RECCON dataset introduces two tasks for extracting emotion cause in conversations:

Task 1: Causal Span Extraction: This task involves finding the emotional cause span for a target utterance. The task has two settings. (1) In the first setting, *conversational history* is not considered (**(w/o CC)**). (2) In the other setting *conversational history* is considered (**(w/ CC)**).

Task 2: Causal Emotion Entailment: This task involves determining whether the candidate utterance causally entails the emotion utterance or not. This task is also formulated in two settings. (1) without Conversational Context **(w/o CC)**. (2) with Conversational Context **(w/ CC)**.

Three fold dataset (§5) is created using RECCON consisting of both positive and negative samples, RoBERTa [24] and SpanBERT [20] are used as the prediction models. Poria et al. [27] use gold emotion annotations during inference for both the given tasks. The authors solve **Cause Span Extraction** as a SQuAD like question answering task where target and cause utterance form the question and the answer contains the cause span. For Fig. 1, a positive sample is created as: ***Context:*** *"What's wrong, officer? You do realize that you ran a red light, don't you? I did?"* ***Question:*** *"The target utterance is I did. The evidence utterance is You do realize that you ran a red light, don't you. What is the causal span from evidence in the context that is relevant to the target utterance's emotion Surprise?"* ***Answer:*** *"you ran a red light"*. Here, the task is to predict the answer span from the context for a given question.
Causal Emotion Entailment is solved as Natural Language Inference (NLI) task. For solving this task, a binary labelled dataset is created as <*Context*> <*SEP*> <*Utterance*> <*SEP*> <*Candidate Cause utterance*> <*SEP*> <*History*>. For example, a positive sample for Fig 1 is created as: *"surprise* $< SEP >$ *I did?* $< SEP >$ *you do realize that you ran a red light, don't you?* $< SEP >$ *What's wrong, officer? You do realize that you ran a red light, don't you? I did?"*. Here, the task is to predict a binary entailment label of 0 if candidate cause utterance doesn't contain the cause of given utterance and a label of 1 if candidate cause utterance contains the cause for a given utterance.

However, we approach the problem differently. For both the tasks CSE and CEE, the input to the model is concatenation of target utterance, cause utterance and context (more details about the dataset in App. B). For example, for Fig. 1 the corresponding input to the model is: *I did?. you do realize that you ran a red light, don't you?* $< SEP >$ *What's wrong, officer? You do realize that you ran a red light, don't you? I did?"* . Given the input in this format, for CSE, the task is to predict start and end positions in the context. For CEE, the task is to predict entailment label as 1 or 0.

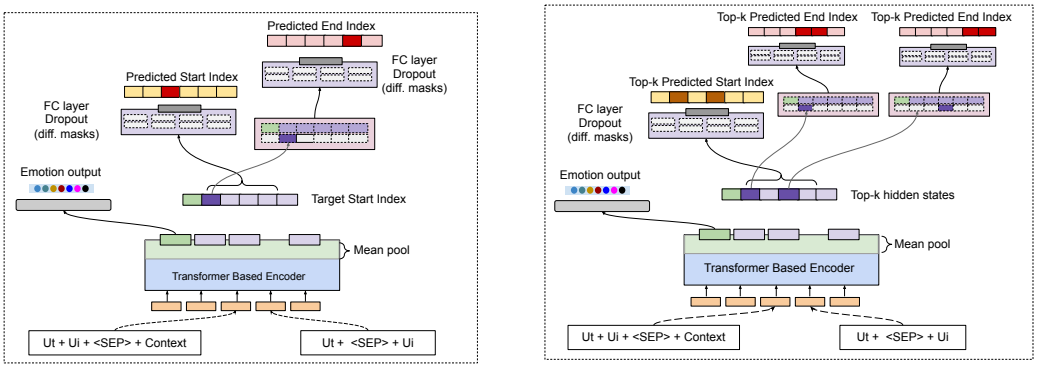
4 Proposed Models

We develop transformer-based multi-task learning models that learn both emotion and emotion cause without being provided with the gold emotion annotations during inference. We perform transfer learning via pre-trained transformer-based LMs.

4.1 Task 1: Cause Span Extraction

As an initial baseline, we solved this problem using two-step model consisting of an *Emotion Predictor (EP)* followed by the *Cause-Span Predictor (CSP)*. An advantage of the two-step model is that it is modular; separate architectures can be applied for both the emotion predictor and cause span predictor. However, there are two drawbacks: 1) The error in the first step is propagated to the next step, and 2) Such an approach assumes that emotion prediction and cause-span prediction are mutually exclusive tasks. To overcome these limitations, we propose an end-to-end architecture.

End-to-End Architecture ($\texttt{MuTEC}_{\texttt{CSE}}$**):** $\texttt{MuTEC}_{\texttt{CSE}}$ is an end-to-end multi-task framework where we perform cause span extraction as the main task and emotion prediction as an auxiliary task



(a) Training architecture for $\texttt{MuTEC}_{\texttt{CSE}}$ (b) For inference, output is $k \times k$ start-end index pairs.

Figure 2: End-to-End architecture ($\texttt{MuTEC}_{\texttt{CSE}}$) for Cause Span Extraction, for training and inference.

(Fig.2(a)). For inference, we use a slightly modified version of the training architecture (Fig.2(b)). The architecture is inspired from the idea that lets the model use the start position information to predict the end token position.

$\texttt{MuTEC}_{\texttt{CSE}}$ **Training:** The input consists of target utterance U_t, candidate cause utterance U_i and $Context$ (w/ CC setting) or no context (w/o CC setting). The input is passed into a transformer based pre-trained model, we mean pool all the 12 layers of pre-trained model to get sequence output: $\text{pool}, \text{h}_1, \text{h}_2, \ldots = E_{sb}(U_t U_i \texttt{<SEP>} Context)$ and $\text{pool}', \text{h}'_1, \text{h}'_2, \ldots = \text{meanpool}(pool, h_1, h_2, ..)_{12}$. The pooled output is used to predict the auxiliary task of emotion prediction. It is passed through a MLP layer and then through a softmax to get the predicted emotion: $emotion^{logit} = \text{MLP}(\text{pool}')$, and $emotion_{pred} = \sigma(emotion^{logit})$. During training, the given start position is used to predict the end index. The start hidden state (h_s) and the original hidden states ($h'_1, h'_2, \ldots$) are concatenated. The concatenated hidden states are passed through a multi sample dropout (MSDropout) [17] to get the predicted end logit. This end logit is then passed through a softmax layer: $start^{logit} = \text{MSDropout}([h'_1, h'_2, \ldots])$; $start_{pred} = \sigma(start^{logit})$, $end^{logit} = \text{MSDropout}([h'_1, h'_2, \ldots] \oplus h_s)$, and $end_{pred} = \sigma(end^{logit})$. Here, $\oplus$ is the concatenation operation. The training loss is a linear combination of the loss for cause-span prediction and emotion prediction: $\mathcal{L}_{total} = \mathcal{L}_{cause\_span} + \beta \mathcal{L}_{emotion}$. β is a hyperparameter, determined using the validation set.

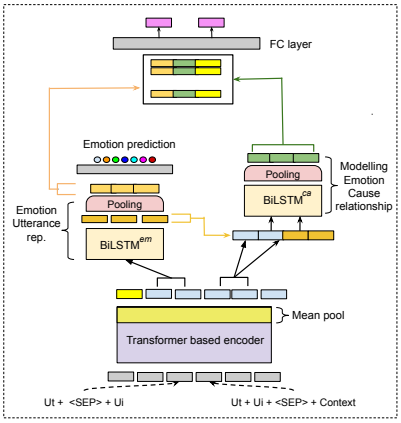
$\texttt{MuTEC}_{\texttt{CSE}}$ **Inference:** During inference, we are not provided with start positions. Hence we find top-k start indices and concatenate the hidden state of each such index to original hidden states, thus creating k different end candidate logits. For each of such k end logits, we again find top-k end indices. We refer this k as the *beam size.* This creates $k \times k$ start-end index pairs, and `argmax` over these $k \times k$ gives the predicted start and end index.

4.2 Task 2: Causal Emotion Entailment

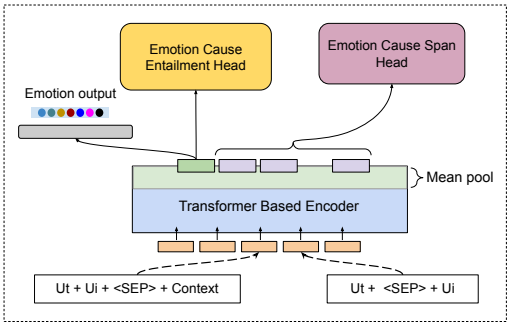
For the task of Causal Emotion Entailment, we propose a multi-task learning approach, $\texttt{MuTEC}_{\texttt{CEE}}$, that consists of three components (Fig. 3(a)). The first component learns contextual representations of the input, i.e., target utterance, candidate cause utterance, and the context. Second component models the relationship between cause and emotion utterances to obtain better representations. Finally, the third component concatenates all the representations and performs entailment (a sentence pair classification task). In order to learn better emotion representations, we include emotion prediction as an auxiliary task.

Learning Contextual Representations: Given an input: $U_t + U_i + \texttt{<SEP>} + Context$ (w/ CC) and $U_t + \texttt{<SEP>} + U_i$ (w/o CC), we use the RoBERTa model to encode the input and learn contextualized representations: $pool, h1, h2, .. = E_{rb}(U_t U_j \texttt{<SEP>} Context)$. We empirically found out that mean-pooling last 4 hidden layers gave the best results.

Modelling Emotion-Cause relationship: The representations of U_t from the first component are then passed into a first-token-level BiLSTM ($BiLSTM^{em}$), for capturing only the target utterance's context and to predict utterance emotions (by mean pooling the representations and passing it through a single layer neural network): $[h^{j'}_{t_1}, h^{j'}_{t_2}, h^{j'}_{t_3}, ...] = BiLSTM^{em}([h^{j}_{t_1}, h^{j}_{t_2}, h^{j}_{t_3}, ...])$. For the auxiliary emotion prediction, the output is mean-pooled and passed through a single neural



(a) End-to-End architecture ($\texttt{MuTEC}_{\texttt{CEE}}$) for Causal Emotion Entailment



(b) E2E model for CSE and CEE ($\texttt{MuTEC}_{\texttt{E2E}}$)

Figure 3: Proposed Models

network: $H_{jt} = \text{meanpool}([h^{j'}_{t_1}, h^{j'}_{t_2}, h^{j'}_{t_3}, ...])$ and $y'_{aux} = \sigma(MLP(H_{jt})$. The hidden representations from $BiLSTM^{em}$ i.e., $(h^{j'}_{t_1}, h^{j'}_{t_2}, h^{j'}_{t_3}, ...)$ are first concatenated with the representations of U_i and then passed into another BiLSTM ($BiLSTM^{ca}$) that uses target utterance's representations to capture the cause utterance. This is used to model the relationship between cause and emotion utterance. The output of $BiLSTM^{ca}$ gives utterance's emotion-cause representations: $[h^{j''}_1, h^{j''}_2, \ldots] = BiLSTM^{ca}([h^{j'}_1, h^{j'}_2, \ldots])$, where, $[h^{j'}_1, h^{j'}_2, \ldots] = [h^{j'}_{t_1}, h^{j'}_{t_2}, \ldots] \oplus [h^{j}_{i_1}, h^{j}_{i_2}, \ldots]$.

Entailment Prediction: The pooled representation from $BiLSTM^{em}$ and $BiLSTM^{ca}$ are concatenated along with the pool output of the pre-trained encoder ($[pool]$) and then passed through a simple MLP to perform classification: $y' = \sigma(MLP(\mathbf{X}_j))$. Here, $\mathbf{X}_j = [pool \oplus H_{jt} \oplus H_{ji}]$ and $H_{ji} = \text{meanpool}([h^{j''}_1, h^{j''}_2, h^{j''}_3, \ldots])$, $H_{jt} = \text{meanpool}([h^{j'}_{t_1}, h^{j'}_{t_2}, h^{j'}_{t_3}, \ldots])$.

Loss Function: The combined loss function is given by $\mathcal{L}_{total} = \mathcal{L}_{entail} + \beta\mathcal{L}_{emotion}$. Validation set is used to determine β. Since the dataset is highly unbalanced, we use weighted cross-entropy loss for both, where the weights are distributed as inverse of the number of class instances.

4.3 E2E Cause Span and Entailment model

In order to perform the end to end training for both the tasks using a single model, we used a similar architecture to Fig. 2(a) and Fig. 2(b) and added a cause entailment head on top (Fig. 3(b)). The model uses a transformer based encoder (RoBERTa-base) as base layer with three heads, namely, emotion head, cause span head and cause entailment head, on top. The entire model is trained end-to-end. The emotion and cause entailment head is single layered neural net. The emotion cause span head is similar to what is used in $\texttt{MuTEC}_{\texttt{CSE}}$.

Loss Function: The overall loss function for the end to end model is: $\mathcal{L}_{total} = \mathcal{L}_{causespan} + \mathcal{L}_{entail} + \mathcal{L}_{emotion}$. Since the dataset is highly unbalanced, we use weighted cross-entropy loss, where the weights are distributed as inverse of the number of class instances.

5 Experiments and Results

Dataset: The RECCON corpus annotates *DailyDialog* [23] and *IEMOCAP* [3] corpora with emotion cause information. In order to train a model, the instances which are not the cause of an utterance (non-cause utterances) were used to create negative samples. Three strategies are adopted by RECCON to create negative samples resulting in 3 data folds. *Fold 1:* negative examples are created as $(U_t, U_i)|U_i \in H(U_t) - C(U_t)$, here, $H(U_i)$ is the conversational history and $C(U_t)$ is collection of causal utterances for U_t. *Fold 2:* any non-causal utterance U_i is selected randomly along with its conversational history $H(U_i)$ from another dialogue to construct negative examples. *Fold 3:* same as Fold 2, but the only constraint here is that the utterance U_i sampled from another dialogue should have the same emotion as the target utterance U_t to create the negative example (details in App. A).

Train Fold	Test Fold	Model	w/o CC					w/ CC				
			Emotion Acc.	EM_{pos}	$F1_{pos}$	$F1_{neg}$	$F1$	Emotion Acc.	EM_{pos}	$F1_{pos}$	$F1_{neg}$	$F1$
Fold1 (DD)	Fold1 (DD)	RoBERTa-base	-	26.82	45.99	**84.55**	**73.82**	-	31.63	58.17	85.85	**75.45**
		SpanBERT	-	33.26	57.03	80.03	69.78	-	34.64	60.00	**86.02**	75.71
		Two Step	80.43	32.11	53.44	81.87	67.54	82.12	34.22	58.90	83.12	72.13
		MuTEC$_{CSE}$	82.54	**36.29**	**62.12**	61.86	53.76	83.42	**36.87**	**66.92**	73.89	62.90
		MuTEC$_{E2E}$	80.12	35.47	61.74	64.87	55.74	82.02	35.78	64.11	75.41	63.24
	Fold2 (DD)	RoBERTa-base	-	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
		Two Step	81.13	32.43	54.24	83.98	73.04	76.82	34.24	61.66	94.78	87.08
		MuTEC$_{CSE}$	81.24	**35.94**	**62.42**	64.20	54.18	69.16	**36.10**	**66.04**	**96.88**	**89.73**
		MuTEC$_{E2E}$	75.56	35.78	61.22	64.56	53.48	72.87	35.28	64.21	95.39	88.47
	Fold3 (DD)	RoBERTa-base	-	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
		Two Step	81.13	32.43	54.24	79.90	66.88	84.14	34.24	61.66	96.44	86.80
		MuTEC$_{CSE}$	80.57	**35.94**	**62.42**	62.10	51.06	86.12	**36.10**	**66.04**	**96.85**	**88.20**
		MuTEC$_{E2E}$	81.34	35.78	61.22	63.35	52.74	80.19	35.28	64.21	96.48	87.96
	Fold1 (IEMO)	RoBERTa-base	-	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
		Two Step	23.20	18.56	34.12	86.66	74.42	22.24	20.12	33.36	**93.62**	**86.72**
		MuTEC$_{CSE}$	23.66	**30.52**	**50.68**	70.21	55.64	21.22	**31.60**	**53.62**	81.78	72.56
		MuTEC$_{E2E}$	25.34	26.78	47.32	75.68	53.47	17.92	30.74	50.39	83.74	75.63
	Fold2 (IEMO)	RoBERTa-base	-	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
		Two Step	26.86	18.56	34.12	87.80	76.52	28.18	23.56	35.60	94.86	91.22
		MuTEC$_{CSE}$	25.54	**30.52**	**50.68**	71.52	57.60	27.18	**30.32**	**53.62**	**96.60**	**92.96**
		MuTEC$_{E2E}$	28.31	26.78	47.32	76.23	56.95	15.20	30.11	52.75	96.23	92.57
	Fold3 (IEMO)	RoBERTa-base	-	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
		Two Step	26.86	18.56	34.12	86.60	75.84	28.18	23.56	35.60	94.80	92.40
		MuTEC$_{CSE}$	27.30	**30.52**	**50.68**	72.63	58.16	25.30	**30.32**	**53.62**	**97.96**	**94.40**
		MuTEC$_{E2E}$	30.78	26.78	47.32	75.66	59.78	17.02	30.11	52.75	96.80	93.97

Table 1: Results for Cause Span Extraction task for Two Step, MuTEC$_{CSE}$and MuTEC$_{E2E}$on RECCON-DD and RECCON-IEMO. IEMO dataset is only used in the inference phase.

Train Fold	Test Fold	Model	w/o CC				w/ CC			
			Emotion Acc.	$F1_{pos}$	$F1_{neg}$	$macroF1$	Emotion Acc.	$F1_{pos}$	$F1_{neg}$	$macroF1$
Fold1 (DD)	Fold1 (DD)	RoBERTa-base	-	56.64	85.13	70.88	-	64.28	**88.74**	76.51
		RoBERTa-large	-	50.48	**87.35**	68.91	-	66.23	87.89	77.06
		MuTEC$_{CEE}$	83.24	**59.18**	84.20	**71.69**	84.90	**69.20**	85.90	**77.55**
		MuTEC$_{E2E}$	80.12	53.03	86.80	69.91	82.02	64.90	88.12	76.51
	Fold2 (DD)	RoBERTa-base	-	57.50	82.71	70.11	-	59.06	86.91	72.98
		RoBERTa-large	-	56.13	**88.33**	**72.23**	-	60.09	**88.00**	**74.04**
		MuTEC$_{CEE}$	80.20	**60.78**	82.96	71.87	76.20	**64.12**	81.31	72.71
		MuTEC$_{E2E}$	75.56	55.23	86.12	70.67	72.87	58.43	87.21	72.82
	Fold3 (DD)	RoBERTa-base	-	57.52	82.72	70.12	-	49.30	79.27	64.29
		RoBERTa-large	-	56.04	**88.28**	**72.16**	-	**60.63**	**88.30**	**74.46**
		MuTEC$_{CEE}$	82.40	**59.06**	82.10	70.58	84.70	49.74	56.50	53.12
		MuTEC$_{E2E}$	81.34	56.77	86.29	71.53	80.19	46.43	88.13	67.28
	Fold1 (IEMO)	RoBERTa-base	-	25.98	90.73	58.36	-	28.02	95.67	61.85
		RoBERTa-large	-	**32.34**	**95.61**	**63.97**	-	**40.83**	**95.68**	**68.26**
		MuTEC$_{CEE}$	26.12	26.40	91.50	58.95	18.02	39.64	92.51	66.07
		MuTEC$_{E2E}$	25.34	25.23	89.52	57.37	17.92	36.54	92.84	64.69
	Fold2 (IEMO)	RoBERTa-base	-	32.60	89.99	61.30	-	27.14	94.16	60.65
		RoBERTa-large	-	**36.61**	**94.60**	**65.60**	-	37.59	**94.63**	66.11
		MuTEC$_{CEE}$	30.21	32.20	90.52	61.36	15.87	**42.41**	92.40	**67.40**
		MuTEC$_{E2E}$	28.31	30.56	89.27	59.91	15.20	29.63	93.41	61.52
	Fold3 (IEMO)	RoBERTa-base	-	33.24	90.30	61.77	-	23.83	92.97	58.40
		RoBERTa-large	-	**36.55**	**94.59**	**65.57**	-	**37.87**	**94.69**	**66.28**
		MuTEC$_{CEE}$	31.50	33.54	90.26	61.90	17.89	32.56	86.40	59.48
		MuTEC$_{E2E}$	30.78	31.10	88.06	59.58	17.02	30.63	91.47	61.05

Table 2: Results for Causal Emotion Entailment. Results are provided on RECCON-DD and RECCON-IEMO where RECCON-IEMO is only used during inference.

Model Training and Inference: For evaluation, the models are trained on one fold, and other folds are used for inference (hyper-parameters in Appendix C). IEMO is the annotated IEMOCAP dataset which is only used for inference as the number of samples in the annotated IEMOCAP dataset is less for training. The experimental results are averaged across 3 runs to account for the variance in transformer based models.

Cause Span Extraction Task: SpanBERT (finetuned on SQUaD) and RoBERTa Base with a linear layer on top is used as the baseline by Poria et al. [27]. Models are evaluated using Exact match (EM), Positive F1, Negative F1, and Overall F1 (details in App. D). A positive F1 score considers only positive samples (i.e., utterances having cause spans). The negative sample has an empty span.

Dataset	Model	$F1_{pos}$	$F1_{neg}$	$macroF1$
DD	ECPE-2D	55.50	94.96	75.23
	ECPE-MLL	48.48	94.68	71.58
	Rank CP	33.00	**97.30**	65.15
	RoBERTa-base	64.28	88.74	76.51
	RoBERTa-large	66.23	87.89	77.06
	$\texttt{MuTEC}_{\texttt{CEE}}$	**69.20**	85.90	**77.55**
	$\texttt{MuTEC}_{\texttt{E2E}}$	64.90	88.12	76.51
IEMOCAP	ECPE-2D	28.67	**97.39**	63.03
	ECPE-MLL	20.23	93.55	57.65
	Rank CP	15.12	92.24	54.75
	RoBERTa-base	28.02	95.67	61.85
	RoBERTa-large	**40.83**	95.68	**68.26**
	$\texttt{MuTEC}_{\texttt{CEE}}$	39.64	92.51	66.07
	$\texttt{MuTEC}_{\texttt{E2E}}$	36.54	92.84	64.69

Table 3: Model and baseline results on Fold 1 (With CC) of Cause Entailment.

Dataset	Model	w/o CC		w/ CC	
		$\texttt{MuTEC}_{\texttt{CSE}}$	$\texttt{MuTEC}_{\texttt{CEE}}$	$\texttt{MuTEC}_{\texttt{CSE}}$	$\texttt{MuTEC}_{\texttt{CEE}}$
DD	w/ EP	62.86	58.36	65.96	68.44
	w/o EP	59.21	52.68	61.21	64.89
IEMOCAP	w/ EP	51.42	25.77	52.94	38.10
	w/o EP	50.21	24.49	51.33	32.26

Table 4: **Ablation study.** w/ EP: with emotion prediction, w/o EP: without emotion prediction. The results are shown for $F1_{pos}$ (%) for both the tasks.

Results: The results are shown in Table 1. For w/o CC and w/ CC, the EM_{pos} and F1_{pos} scores for MuTEC are significantly higher in most of the cases. In general, we also noted that the high percentage of negative samples in the dataset affected the positive sample scores as well. Also, there seems to be a tradeoff between positive sample scores and negative sample scores. In some cases, our models are not able to beat the baselines for negative sample score, though this is not the case when we are performing inference on Fold2, Fold3, and IEMO (w/ CC) when the model is trained on Fold1. For these, our model surpasses the baseline. A possible reason might be that since, in these cases, we have the context of non-cause utterance combined with the target utterance, the model is easily able to distinguish if it is a positive or a negative sample. The model gets confused for negative samples when the non-cause utterances comes from the same dialogue (Fold1, w/ CC setting). For w/o CC, it is difficult for the model to identify the negative samples resulting in a lower negative sample score. F1 score is calculated as the utterance level mean of positive and negative samples, thus resulting in lower values since the dataset is unbalanced towards negative samples. For w/o CC, since the positive samples are the same for all the folds, the EM_{pos} and $F1_{pos}$ are the same across all the folds. Only negative samples are different in these folds. We get good emotion prediction accuracy for Dailydialog. However, inference scores on IEMOCAP are low. Possible reason might be that since the training dataset (Dailydialog) and the inference dataset (IEMOCAP) are quite different, the model is not able to generalize well. Also, since IEMOCAP has some extra emotions, we clubbed similar kinds of emotions together, like happy and excited, anger and frustrated. Even with lower emotion scores, model performs better than the baselines, showing that the model is able to give better cause span results for cross-dataset settings as well. Results for training on Fold2 and Fold3 are shown in App. Table 11. Similar trends were seen in these folds as well where for positive samples we are able to get better scores and for negative samples the scores drops.

Causal Emotion Entailment: Poria et al. [27] solve this task as a natural language inference task using RoBERTa-base and RoBERTa-large with linear layer on top.

Results: The results for Causal Emotion Entailment are shown in Table 2. For the majority of the Dailydialog dataset, our models surpass the baselines. But for IEMOCAP, the results are lower than the baselines, showing that for the task of causal emotion entailment, the model is not generalizing well. RoBERTa-large gives significantly higher scores for IEMOCAP dataset showing that large-pretrained models work well in cross-dataset setting. The results for training Fold2 and Fold3 (App. Table 12) are consistent to Fold1 where RoBERTa-large shows significant improvements on IEMOCAP dataset. Attention weights of the transformer-based encoder were also visualized (App. Fig. 6), and it showed that since there are a lot of negative samples, the attention scores of the last

Train Fold	Test Fold	Model	w/o CC					w/ CC				
			Emotion Acc.	EM_{pos}	$F1_{pos}$	$F1_{neg}$	**F1**	Emotion Acc.	EM_{pos}	$F1_{pos}$	$F1_{neg}$	**F1**
Fold1 (DD)	Fold1 (DD)	RoBERTa-base	-	36.54	63.77	**70.35**	63.18	-	38.28	68.8	**83.48**	**73.98**
		SpanBERT	-	**36.96**	64.84	69.67	**65.08**	-	38.70	68.83	81.54	72.14
		MuTEC$_{CSE}$	80.23	36.50	**67.90**	65.78	60.13	81.78	**39.91**	**72.41**	72.61	65.60
		MuTEC$_{E2E}$	78.22	35.08	65.63	67.89	59.60	79.43	38.21	70.56	73.47	64.13
	Fold2 (DD)	RoBERTa-base	-	36.54	63.77	57.39	53.85	-	38.17	68.56	95.91	85.21
		SpanBERT	-	**36.96**	64.84	**66.80**	**61.88**	-	38.01	68.98	**96.24**	**85.42**
		MuTEC$_{CSE}$	79.85	36.50	**67.90**	66.14	60.32	66.43	37.30	**70.80**	95.70	84.29
		MuTEC$_{E2E}$	75.45	35.08	65.63	66.54	60.23	74.87	**38.79**	69.87	95.41	83.96
	Fold3 (DD)	RoBERTa-base	-	36.54	63.77	55.64	52.83	-	38.17	68.56	95.85	85.35
		SpanBERT	-	**36.96**	64.84	58.93	56.55	-	38.01	68.98	95.99	85.15
		MuTEC$_{CSE}$	81.40	36.50	**67.90**	64.74	**60.56**	80.74	37.30	**70.80**	**96.41**	**85.70**
		MuTEC$_{E2E}$	80.04	35.08	65.63	**65.32**	58.98	81.97	**38.79**	69.87	96.03	84.78

Table 5: Results for Cause Span Extraction task for the balanced dataset.

layer were mostly corresponding to the first token that represents the negative sample (with start and end equal to zero) for the sample example. Table 3 shows the comparison with other baselines for Cause Entailment task for Fold 1 (with CC).

6 Analysis

Ablation Study: To understand the importance of the auxiliary task of emotion prediction, we tried training the model with and without emotion prediction. Table 4 shows a performance drop when we don't train the model on the auxiliary task of emotion prediction. The study is performed on Fold1 (train and test). Ablation results across all the metrics are shown in App. E. Since the results across training in Fold2 and Fold3 are similar to Fold1, we perform the ablation using Fold1 only. For IEMO it can be seen that the difference isn't significant. That might be because the emotion prediction in itself isn't much accurate due to different emotion label distribution of both RECCON-DD and RECCON-IEMO (details in App. Table 9). To understand the effect of *beam size* (§4.1) on the SQuAD $F1_{pos}$ score for Cause Span Extraction, we experimented with different *beam size*. After beam size of 3 (refer Fig. 7 in Appendix), $F1_{pos}$ remains almost constant, thus we considered the beam size of 3 for our experiments.

Experiments with Balanced Dataset: The RECCON dataset contains lot more negative samples (data statistics in App. A). We conducted same set of experiments with balanced dataset by reducing the number of negative samples, considering only two non-cause utterances for creating negative samples for each utterance (the statistics of balanced dataset in App. Table 8). The results for balanced dataset are presented in Table 5 for CSE task. Comparing the results of before and after balancing the dataset, it is evident that reducing negative samples increases the overall score of positive samples for both the tasks. Thus having a balanced set of samples helps the model to learn better. The results of balanced dataset on CEE task is shown in Table 15 in Appendix. Our model showed similar trends as in the full dataset and gave good performance for positive samples and produced comparatively lower scores for negative samples for training and testing on similar folds.

7 Conclusion and Future Work

In this paper, we explore the task of extracting emotion cause in conversations. We experiment with the RECCON dataset. We propose a set of model architectures that do not require emotion annotations at the inference time. In particular we propose, multi-task learning approach where emotions are learned as an auxiliary task during cause span extraction (CSE) or causal emotion entailment (CEE) tasks. We also propose an overall end-to-end architecture for learning both the tasks together. As shown in experiments, the models give comparable to better results without explicit emotion annotations at inference time. For future work, including the causal reasoning along with the cause spans in the annotated dataset can help the model to understand why this particular cause was selected. Also, currently, the RECCON dataset only uses dyadic conversations. This motivates the creation of datasets that use the multi-party setting.

8 Acknowledgements

We would like to thank reviewers for their insightful comments. This research is supported by SERB India (Science and Engineering Board) Research Grant number SRG/2021/000768.

References

[1] Keshav Bansal, Harsh Agarwal, Abhinav Joshi, and Ashutosh Modi. Shapes of emotions: Multimodal emotion recognition in conversations via emotion shifts. In *Proceedings of the First Workshop on Performance and Interpretability Evaluations of Multimodal, Multipurpose, Massive-Scale Models*, pages 44–56, Virtual, October 2022. International Conference on Computational Linguistics. URL `https://aclanthology.org/2022.mmmpie-1.6`.

[2] Hongliang Bi and Pengyuan Liu. Ecsp: A new task for emotion-cause span-pair extraction and classification. *arXiv preprint arXiv:2003.03507*, 2020.

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

[4] Emile Chapuis, Pierre Colombo, Matteo Manica, Matthieu Labeau, and Chloe Clavel. Hierarchical pre-training for sequence labelling in spoken dialog. *arXiv preprint arXiv:2009.11152*, 2020.

[5] Xinhong Chen, Qing Li, and Jianping Wang. A unified sequence labeling model for emotion cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 208–218, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

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

[7] Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–19, 2017.

[8] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1374. URL `https://aclanthology.org/N19-1374`.

[9] Jiawen Deng and Fuji Ren. A survey of textual emotion recognition and its challenges. *IEEE Transactions on Affective Computing*, pages 1–1, 2021. doi: 10.1109/TAFFC.2021.3053275.

[10] Jiayuan Ding and Mayank Kejriwal. An experimental study of the effects of position bias on emotion causeextraction. *arXiv preprint arXiv:2007.15066*, 2020.

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

[12] Zixiang Ding, Rui Xia, and Jianfei Yu. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583, 2020.

[13] Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3707–3717, 2020.

[14] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*, 2019.

[15] Tushar Goswamy, Ishika Singh, Ahsan Barkati, and Ashutosh Modi. Adapting a language model for controlled affective text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2787–2801, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.251. URL `https://aclanthology.org/2020.coling-main.251`.

[16] Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1639–1649, Austin, Texas, November 2016. Association for Computational Linguistics.

[17] Hiroshi Inoue. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*, 2019.

[18] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370, 2020.

[19] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. COGMEN: COntextualized GNN based multimodal emotion recognitioN. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.306. URL `https://aclanthology.org/2022.naacl-main.306`.

[20] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[21] Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4752–4757, 2018.

[22] Xiangju Li, Shi Feng, Daling Wang, and Yifei Zhang. Context-aware emotion cause analysis with multi-attention-based neural network. *Knowledge-Based Systems*, 174:205–218, 2019.

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

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

[25] Marvin Minsky. *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind*. Simon and Schuster, 2007.

[26] Alena Neviarouskaya and Masaki Aono. Extracting causes of emotions from text. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 932–936, 2013.

[27] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Romila Ghosh, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. Recognizing emotion cause in conversations. *arXiv preprint arXiv:2012.11820*, 2020.

[28] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[29] Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. EMOCause: An easy-adaptable approach to extract emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 153–160, Portland, Oregon, June 2011. Association for Computational Linguistics.

[30] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. *arXiv preprint arXiv:2012.08695*, 2020.

[31] Dongming Sheng, Dong Wang, Ying Shen, Haitao Zheng, and Haozhuang Liu. Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4153–4163, 2020.

[32] Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. An end-to-end network for emotion-cause pair extraction. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 84–91, Online, April 2021. Association for Computational Linguistics. URL `https://aclanthology.org/2021.wassa-1.9`.

[33] Gargi Singh, Dhanajit Brahma, Piyush Rai, and Ashutosh Modi. Fine-grained emotion prediction by modeling emotion definitions. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, 2021. doi: 10.1109/ACII52823.2021.9597436.

[34] Haolin Song, Chen Zhang, Qiuchi Li, and Dawei Song. End-to-end emotion-cause pair extraction via learning to link. *arXiv preprint arXiv:2002.10710*, 2020.

[35] Qixuan Sun, Yaqi Yin, and Hong Yu. A dual-questioning attention network for emotion-cause pair extraction with context awareness. *arXiv preprint arXiv:2104.07221*, 2021.

[36] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[37] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy, July 2019. Association for Computational Linguistics.

[38] Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195, 2020.

[39] Penghui Wei, Jiahao Zhao, and Wenji Mao. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181, Online, July 2020. Association for Computational Linguistics.

[40] Rui Xia and Zixiang Ding. Emotion-cause pair extraction: A new task to emotion analysis in texts. *arXiv preprint arXiv:1906.01267*, 2019.

[41] Rui Xia, Mengran Zhang, and Zixiang Ding. Rthn: A rnn-transformer hierarchical network for emotion cause extraction. *arXiv preprint arXiv:1906.01236*, 2019.

[42] Xinglin Xiao, Penghui Wei, Wenji Mao, and Lei Wang. Context-aware multi-view attention networks for emotion cause extraction. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 128–133. IEEE, 2019.

[43] Bo Xu, Hongfei Lin, Yuan Lin, Yufeng Diao, Liang Yang, and Kan Xu. Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583, 2019.

[44] Peixiang Zhong, Di Wang, and Chunyan Miao. Knowledge-enriched transformer for emotion detection in textual conversations. *arXiv preprint arXiv:1909.10681*, 2019.

Appendix

A RECCON Statistics

RECCON dataset was build using two popular conversational datasets **DialyDialog** [23] and **IEMOCAP** [3], both already had utterance level emotions associated. The RECCON dataset used only a subset of the IEMOCAP dataset and randomly selected dialogues from the DailyDialog dataset containing a minimum of four *non-neutral* utterances. They did it because about 83% of the DailyDialog dataset has `Neutral` labels. The annotated dataset was named RECCON-IE and RECCON-DD for IEMOCAP and DailyDialog, respectively.

Table 6 shows some of the statistics of RECCON annotated dataset. From the table, it can be seen that in RECCON-IE 40.5% of utterances have a cause of the emotion in greater than three utterance distance in the conversational history, whereas in RECCON-DD only 13% of utterances have their emotion cause in greater than three distance in the conversational history. Fig. 4(a) and 4(b) shows the distribution of utterance length in RECCON-DD and RECCON-IEMO respectively.

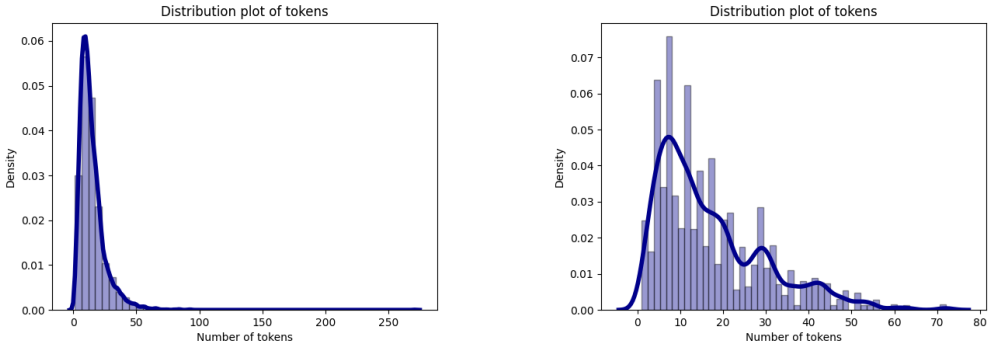


Figure 4: **Distribution plot for number of tokens in an utterance of RECCON-DD and RECCON-IEMO**

Table 7 shows the final statistics of the dataset with positive and negative samples.

B RECCON Tasks

The task for **Cause Span Extraction** was solved as Question Answering task. **Context:** Conversational history is context of a target utterance U_t. (U_t, U_i) is used for a negative examples, where $U_i \notin C(U_t)$, conversational history of U_t.
Question: The following is how the question is phrased: *"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 present in U_i if $U_i \in C(U_t)$. S is assigned an empty string for negative samples.
For **Fold2** and **Fold3** the context of U_i is considered for negative samples.

For **Causal Emotion Entailment**, the task was solved as Natural language inference task. All the folds were structured as the follows:
Input as: $< E_t >< SEP >< U_t >< SEP >< U_i >< SEP >< H(U_t) >$ and a label 1 if $U_i \in C(U_t)$ and label 0 if $U_i \notin C(U_t)$

For all folds in **Without Conversational Context (w/o CC)** setting, the context was not considered in the dataset.

We converted the data into the following:
Task 1: [id, emotion, U_t, U_i, cause_span, context]
Task 2: [id, emotion, U_t, U_i, context, labels]

Description (Number of)	RECCON-DD	RECCON-IE
Dialogues	1106	16
Utterances	11104	665
Utterances annotated with emotion cause	5861	494
Utterances cater to background cause	395	70
Utterances where cause solely lies in the same utterance	1521	80
Utterance where cause includes the same utterance along with contextual utterances	3370	243
Number of emotion Utterances		
Anger	451	89
Fear	74	-
Disgust	140	-
Frustrated	-	109
Happiness	4361	58
Sadness	351	70
Surprise	484	-
Excited	-	197
Neutral	5243	142
U_t having cause at $U_{(t-1)}$	2851	183
U_t having cause at $U_{(t-2)}$	1182	124
U_t having cause at $U_{(t-3)}$	578	94
U_t having cause at > $U_{(t-3)}$	769	200

Table 6: **Statistics of RECCON annotated dataset**. Taken from the paper[27]

Data			Train	Val	Test
Fold 1	**DD**	Positive Samples	7269	347	1894
		Negative Samples	20646	838	5330
	IEMO	Positive Samples	-	-	1080
		Negative Samples	-	-	11305
Fold 2	**DD**	Positive Samples	7269	347	1184
		Negative Samples	18428	800	4396
	IEMO	Positive Samples	-	-	1080
		Negative Samples	-	-	7410
Fold 3	**DD**	Positive Samples	7269	347	1894
		Negative Samples	18428	800	4396
	IEMO	Positive Samples	-	-	1080
		Negative Samples	-	-	7410

Table 7: **Dataset Statistics with both Positive Samples and Negative Samples.** DD refers to RECCON-DD and IEMO refers to RECCON-IEMO. Latent emotion cause is ignored.

After transforming the dataset in U_t, U_i pairs, Table 9 shows the number of emotion labels associated with Fold1.

C Model Hyperparameters

The value of hyperparameters used for both the tasks are listed in Table 10

Data			Train	Val	Test
Fold 1	**DD**	Positive Samples	7269	347	1894
		Negative Samples	7356	308	1811
Fold2	**DD**	Positive Samples	7269	347	1184
		Negative Samples	9124	400	2198
Fold 3	**DD**	Positive Samples	7269	347	1894
		Negative Samples	9124	400	2198

Table 8: **Dataset Statistics for balanced set.** Negative samples were reduced by only taking two of the non-cause utterances for each target utterance instead of all the non-cause utterances for creating negative samples.

Dataset		Happiness	Surprise	Anger	Sadness	Disgust	Fear	Excited	Frustrated
DD	Train	22095	2205	1513	1269	555	278	-	-
	Valid	785	112	139	114	10	25	-	-
	Test	4520	576	982	806	192	148	-	-
IEMO	Test	1295	-	1535	1503	-	-	5778	2274

Table 9: **Number of emotion labels in Fold1 after the dataset is transformed into U_t, U_i pairs.** Dataset is highly unbalanced and the distribution of emotion labels in DD and IEMO are not the same.

Hyperparameters	Task 1	Task 2
Number of Epochs	12	
Batch size	16	
Max. sequence length (with context)	512	
Max. sequence length (without context)	200	
Initial Learning Rate	4e-5	
Optimizer	AdamW	
Schedular	get_linear_schedule_with_warmup (warmup steps = 4)	
Max. answer length	200	-
n_hidden_states	12	4
Dropout	Multi-sample dropout (probability=0.5, layers=5)	Simple dropout (probability=0.1)
Beam_width	3	-
Bi-LSTM hidden dim	-	384
weight decay	0.001	

Table 10: **Hyperparameter values used for both the tasks.**

D Evaluation Metrics

D.1 Cause Span Extraction Metrics

For the evaluation of the models, the following metrics are used:

Exact Match (EM_{pos}): Exact Match corresponds to the percentage of exactly matched predicted spans to the gold spans.
Positive F1 ($F1_{pos}$): SQuAD F1 score [28] calculated over positive examples. This metric measures

the average overlap between the predicted spans and the ground span.

$$P_{pos} = \frac{Number\ of\ same\ tokens}{Number\ of\ predicted\ token} \tag{1}$$

$$R_{pos} = \frac{Number\ of\ same\ tokens}{Number\ of\ gold\ tokens} \tag{2}$$

$$F1_{pos} = \frac{2 * P_{pos} * R_{pos}}{(P_{pos} + R_{pos})} \tag{3}$$

Negative F1 ($F1_{neg}$): F1 score calculated over negative examples. Here, the gold spans are empty spans.

$$P_{neg} = \frac{Number\ of\ same\ empty\ spans}{Total\ number\ of\ predicted\ empty\ spans} \tag{4}$$

$$R_{neg} = \frac{Number\ of\ same\ empty\ spans}{Total\ number\ of\ gold\ empty\ spans} \tag{5}$$

$$F1_{neg} = \frac{2 * P_{neg} * R_{neg}}{(P_{neg} + R_{neg})} \tag{6}$$

F1: Overall F1 is calculated for each of the examples (positive and negative), which is followed by averaging over both of them.

D.2 Causal Emotion Entailment Metrics

Positive F1 ($F1_{pos}$): F1 Score calculated for positive examples i.e., F1 score when positive samples are considered as true class.
Negative F1 ($F1_{neg}$): F1 Score calculated for negative examples i.e., F1 score when negative samplesa are considered as true class.
Macro F1: Mean of class-wise (positive and negative) F1-scores.

The logit scores (for start and end) calculated by the model for a test sample are given in Fig. 5.

E Ablation Study

We performed ablation study on Fold1 by removing the emotion predictor for both the tasks. The details are shown in Table 13 and 14 respectively.

Train Fold	Test Fold	Model	Without cc					With cc				
			Emotion Acc.	EM_{pos}	$F1_{pos}$	$F1_{neg}$	$F1$	**Emotion Acc.**	EM_{pos}	$F1_{pos}$	$F1_{neg}$	$F1$
Fold2(DD)	Fold1 (DD)	RoBERTa-base	-	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
		Two Step	76.78	31.57	55.61	**79.63**	**68.12**	76.13	35.80	**66.38**	0.75	**17.68**
		MuTEC$_{CSE}$	79.72	**35.06**	**64.10**	60.86	50.87	78.19	32.15	61.31	**2.19**	16.89
	Fold2 (DD)	RoBERTa-base	-	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
		Two Step	79.95	**35.37**	63.13	86.17	75.81	80.09	41.29	74.31	99.95	92.23
		MuTEC$_{CSE}$	79.77	35.06	**64.10**	81.91	71.05	78.30	**42.56**	74.62	99.91	92.31
	Fold3 (DD)	RoBERTa-base	-	-	-	-	-	-	-	-	-	-
		SpanBERT	-	-	-	-	-	-	-	-	-	-
		Two Step	79.95	35.37	63.13	83.13	72.59	80.09	41.29	74.31	99.79	92.01
		MuTEC$_{CSE}$	79.85	35.06	64.10	79.23	68.25	79.15	42.56	74.62	99.75	92.09
	Fold1 (IEMO)	RoBERTa-base	-	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
		Two Step	22.43	22.69	40.35	84.01	72.69	22.55	28.43	50.30	**43.96**	**30.72**
		MuTEC$_{CSE}$	21.43	**30.28**	**50.68**	72.90	58.64	20.10	**30.28**	**58.19**	6.36	8.09
	Fold2 (IEMO)	RoBERTa-base	-	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
		Two Step	23.67	28.98	**51.5**	82.72	75.01	19.75	34.44	58.55	97.60	93.70
		MuTEC$_{CSE}$	22.02	**30.28**	50.68	80.61	68.58	20.69	**43.52**	**77.71**	98.01	94.21
	Fold3 (IEMO)	RoBERTa-base	-	-	-	-	-	-	-	-	-	-
		SpanBERT	-	-	-	-	-	-	-	-	-	-
		Two Step	23.67	28.98	51.5	81.91	74.7	20.77	34.44	58.55	97.11	92.87
		MuTEC$_{CSE}$	22.29	30.28	50.68	81.26	69.42	19.55	43.52	77.71	96.63	91.91
Fold3(DD)	Fold1 (DD)	RoBERTa-base	-	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
		Two Step	78.11	32.37	59.15	67.5	56.1	76.13	31.36	61.63	0.71	16.35
		MuTEC$_{CSE}$	80.75	**37.43**	**66.21**	53.7	45.76	84.44	34.16	64.29	**2.41**	17.75
	Fold2 (DD)	RoBERTa-base	-	-	-	-	-	-	-	-	-	-
		SpanBERT	-	-	-	-	-	-	-	-	-	-
		Two Step	79.44	32.37	58.95	87.36	77.24	79.09	40.34	74.55	99.93	92.27
		MuTEC$_{CSE}$	79.49	37.43	66.21	76.24	65.53	71.26	41.24	74.31	99.90	92.23
	Fold3 (DD)	RoBERTa-base	-	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
		Two Step	79.44	32.37	58.95	86.66	76.34	81.12	40.34	74.55	99.85	92.16
		MuTEC$_{CSE}$	80.49	**37.43**	**66.21**	74.62	63.99	92.52	41.02	75.08	99.80	92.29
	Fold1 (IEMO)	RoBERTa-base	-	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
		Two Step	23.44	26.39	44.38	82.67	70.95	22.55	26.30	44.9	**42.16**	**28.9**
		MuTEC$_{CSE}$	22.40	**36.30**	**57.54**	70.00	55.61	21.21	**32.59**	**59.47**	6.46	8.24
	Fold2 (IEMO)	RoBERTa-base	-	-	-	-	-	-	-	-	-	-
		SpanBERT	-	-	-	-	-	-	-	-	-	-
		Two Step	24.64	26.3	44.71	88.69	79.76	23.12	36.48	59.05	97.36	93.41
		MuTEC$_{CSE}$	21.01	36.3	57.74	79.38	67.32	20.18	46.20	76.64	98.31	94.47
	Fold3 (IEMO)	RoBERTa-base	-	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
		Two Step	24.64	26.30	44.71	88.27	79.84	27.12	36.48	59.05	97.64	93.61
		MuTEC$_{CSE}$	22.11	**36.30**	**57.74**	79.48	67.45	22.66	**46.20**	**76.64**	97.08	92.42

Table 11: **Comparision results for Cause Span Extraction task** for Two Step and MuTEC$_{CSE}$model architecture on RECCON-DD and RECCON-IEMO. IEMO dataset is only used in the inference phase. (Fold2 and Fold3)

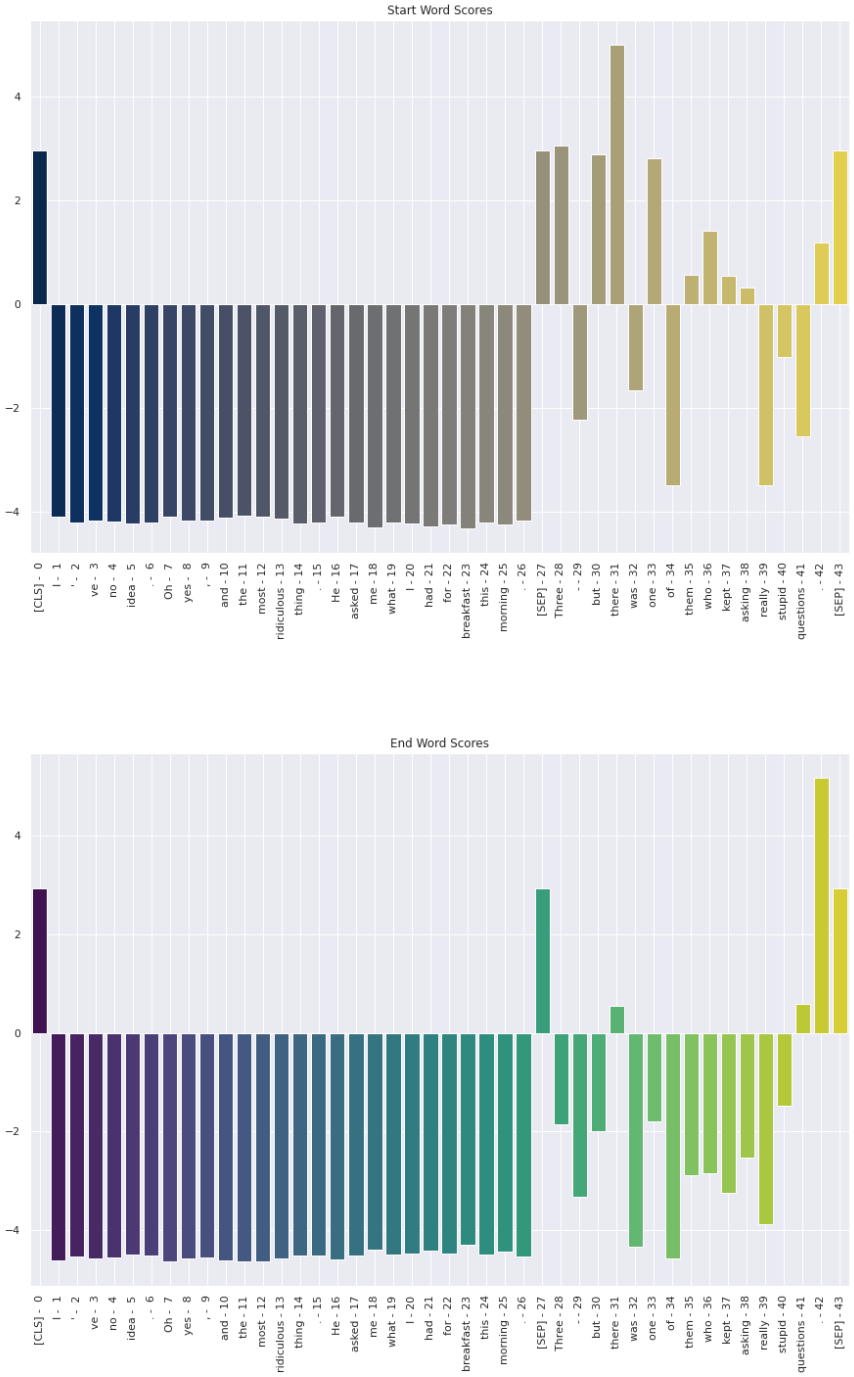


Figure 5: **Start and End Words score for an example input**. *there* is predicted as the start token and . as the end token.

Train Fold	Test Fold	Model	Without cc				With cc			
			Emotion Acc.	$F1_{pos}$	$F1_{neg}$	$macroF1$	Emotion Acc.	$F1_{pos}$	$F1_{neg}$	$macroF1$
Fold2(DD)	Fold1 (DD)	RoBERTa-base	-	52.52	75.51	64.02	-	41.86	3.25	22.55
		RoBERTa-large	-	51.57	67.58	59.57	-	43.25	19.95	31.60
		MuTEC$_{CEE}$	72.23	**53.17**	**78.23**	**65.70**	82.01	42.99	17.90	30.45
	Fold2 (DD)	RoBERTa-base	-	76.21	91.23	83.72	-	89.37	95.21	92.32
		RoBERTa-large	-	79.52	91.27	85.40	-	93.05	97.22	95.13
		MuTEC$_{CEE}$	74.14	74.53	89.91	82.22	81.55	**94.01**	**97.59**	**95.80**
	Fold3 (DD)	RoBERTa-base	-	-	-	-	-	-	-	-
		RoBERTa-large	-	-	-	-	-	-	-	-
		MuTEC$_{CEE}$	71.43	72.52	89.98	81.25	81.52	69.47	81.03	75.25
	Fold1 (IEMO)	RoBERTa-base	-	31.51	92.09	61.80	-	25.22	74.69	49.96
		RoBERTa-large	-	29.64	87.68	58.66	-	26.30	76.44	51.37
		MuTEC$_{CEE}$	18.5	30.74	90.55	60.65	15.97	**27.44**	**80.21**	**53.82**
	Fold2 (IEMO)	RoBERTa-base	-	46.12	93.80	69.96	-	65.09	95.60	80.35
		RoBERTa-large	-	48.36	92.06	70.21	-	61.12	95.59	78.35
		MuTEC$_{CEE}$	20.85	**49.02**	92.80	**70.91**	17.73	61.68	95.44	78.56
	Fold3 (IEMO)	RoBERTa-base	-	-	-	-	-	-	-	-
		RoBERTa-large	-	-	-	-	-	-	-	-
		MuTEC$_{CEE}$	18.91	61.17	95.62	78.39	20.08	46.09	92.35	69.22
Fold 3(DD)	Fold1 (DD)	RoBERTa-base	-	52.02	74.59	63.31	-	41.64	2.99	22.31
		RoBERTa-large	-	51.53	65.76	58.65	-	41.86	4.89	23.38
		MuTEC$_{CEE}$	79.98	**53.86**	**77.43**	**65.64**	90.33	**42.43**	**12.33**	**27.38**
	Fold2 (DD)	RoBERTa-base	-	-	-	-	-	-	-	-
		RoBERTa-large	-	-	-	-	-	-	-	-
		MuTEC$_{CEE}$	80.28	74.88	90.49	82.69	79.03	52.73	46.50	49.62
	Fold3 (DD)	RoBERTa-base	-	74.73	90.33	82.53	-	92.64	96.99	94.81
		RoBERTa-large	-	75.79	88.43	82.11	-	93.34	97.23	95.29
		MuTEC$_{CEE}$	80.80	74.11	90.05	82.08	95.62	**96.36**	**98.48**	**97.42**
	Fold1 (IEMO)	RoBERTa-base	-	34.74	91.46	63.10	-	19.13	54.25	36.69
		RoBERTa-large	-	27.58	84.13	55.86	-	18.33	48.01	33.17
		MuTEC$_{CEE}$	18.56	30.52	90.04	60.28	24.27	**19.90**	**56.53**	**38.21**
	Fold2 (IEMO)	RoBERTa-base	-	-	-	-	-	-	-	-
		RoBERTa-large	-	-	-	-	-	-	-	-
		MuTEC$_{CEE}$	21.41	47.76	93.08	70.42	26.86	31.16	71.29	51.73
	Fold3 (IEMO)	RoBERTa-base	-	51.23	93.70	72.46	-	63.91	94.55	79.23
		RoBERTa-large	-	43.00	88.47	65.74	-	59.03	92.21	75.62
		MuTEC$_{CEE}$	21.86	48.30	93.42	70.86	30.07	58.02	93.59	75.74

Table 12: **Comparison results for Causal Emotion Entailment.** Results are provided on RECCON-DD and RECCON-IEMO where RECCON-IEMO is only used during inference.

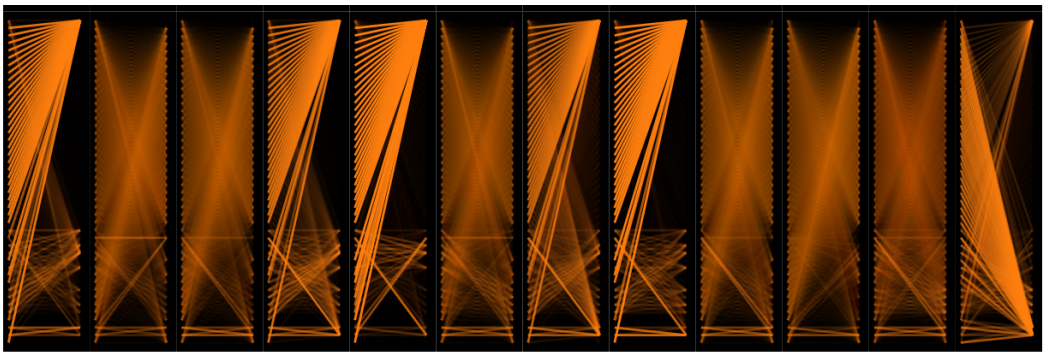


Figure 6: **Attention for Layer 12 of SpanBERT across all the attention heads for a trained model.** Bertviz [37] library was used to produce these attention visualizations on trained model (on Fold1) for task 1.

Dataset	Model Setting	Without cc				With cc			
		EM_{pos}	$F1_{pos}$	$F1_{neg}$	$F1$	EM_{pos}	$F1_{pos}$	$F1_{neg}$	$F1$
DD	With emotion prediction	36.06	62.86	62.50	52.10	36.43	65.96	75.07	63.64
	Without emotion prediction	34.75	59.21	63.58	50.33	33.32	61.21	78.09	62.87
IEMO	With emotion prediction	31.85	51.42	69.20	54.42	30.56	52.94	82.24	70.37
	Without emotion prediction	30.99	50.21	70.43	53.14	29.43	51.33	83.10	70.09

Table 13: Task 1 end-to-end model trained with and without the auxiliary emotion prediction task.

Dataset	Model Setting	Without cc			With cc		
		$F1_{pos}$	$F1_{neg}$	macro $F1$	$F1_{pos}$	$F1_{neg}$	macro $F1$
DD	With emotion prediction	58.26	84.94	71.60	68.44	86.88	77.66
	Without emotion prediction	52.68	79.69	66.19	64.89	84.37	74.63
IEMO	With emotion prediction	25.77	90.39	58.08	38.10	93.50	65.80
	Without emotion prediction	24.49	88.00	56.25	32.26	92.32	62.29

Table 14: Task 2 end-to-end model trained with and without auxiliary emotion task.

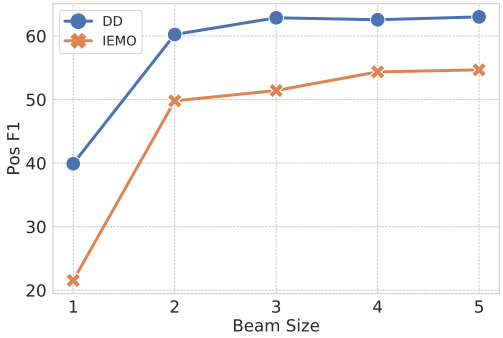


Figure 7: Effect of *Beam Size* on $F1_{pos}$ (%) for Fold1 (*w/o CC*).

Train Fold	Test Fold	Model	w/o CC				w/ CC			
			Emotion Acc.	$F1_{pos}$	$F1_{neg}$	macro F1	Emotion Acc.	$F1_{pos}$	$F1_{neg}$	macro F1
Fold1 (DD)	Fold1 (DD)	RoBERTa-base	-	75.67	69.96	72.82	-	85.12	85.14	85.13
		RoBERTa-large	-	**75.95**	69.73	72.84	-	**85.43**	84.93	**85.18**
		MuTEC$_{CEE}$	79.14	75.81	**71.21**	**73.51**	80.56	85.13	84.69	84.90
		MuTEC$_{E2E}$	78.22	73.42	70.16	71.79	79.43	82.65	**86.41**	84.53
	Fold2 (DD)	RoBERTa-base	-	66.32	55.22	60.77	-	65.09	**65.02**	65.05
		RoBERTa-large	-	67.79	**58.12**	**62.96**	-	68.48	54.20	61.34
		MuTEC$_{CEE}$	75.85	**69.55**	53.04	61.29	75.77	**69.84**	54.98	62.41
		MuTEC$_{E2E}$	75.45	65.21	56.38	60.79	74.87	68.41	62.11	**65.26**
	Fold3 (DD)	RoBERTa-base	-	66.13	54.64	60.39	-	57.14	**43.15**	**50.14**
		RoBERTa-large	-	67.76	58.04	62.90	-	59.57	13.71	36.64
		MuTEC$_{CEE}$	81.03	**69.74**	**59.44**	**64.59**	83.41	**60.54**	23.41	41.97
		MuTEC$_{E2E}$	80.04	66.28	57.62	61.95	81.97	56.23	35.21	45.72

Table 15: Results for Causal Emotion Entailment task for the balanced dataset.