# DeCoDE: Detection of Cognitive Distortion and Emotion cause extraction in clinical conversations

Gopendra Vikram Singh[1*][0000−0003−1104−5856], Soumitra Ghosh[1*][0000−0003−1910−4320], Asif Ekbal[1][0000−0003−3612−8834], and Pushpak Bhattacharyya[2][0000−0001−5319−5508]

[1] Indian Institute of Technology Patna, India
[2] Indian Institute of Technology Bombay, India
{gopendra_1921cs15,asif}@iitp.ac.in, ghosh.soumitra2@gmail.com, pb@cse.iitb.ac.in

**Abstract.** Despite significant evidence linking mental health to almost every major development issue, individuals with mental disorders are among those most at risk of being excluded from development programs. We outline a novel task of detection of *Cognitive Distortion* and *Emotion Cause* extraction of associated *emotions* in conversations. Cognitive distortions are inaccurate thought patterns, beliefs, or perceptions that contribute to negative thinking, which subsequently elevates the chances of several mental illnesses. This work introduces a novel multi-modal mental health conversational corpus manually annotated with *emotion*, *emotion causes*, and the presence of *cognitive distortion* at the utterance level. We propose a multitasking framework that uses multi-modal information as inputs and uses both external commonsense knowledge and factual knowledge from the dataset to learn both tasks at the same time. This is because commonsense knowledge is a key part of understanding how and why emotions are implied. We achieve commendable performance gains on the *cognitive distortion* detection task (+3.91 F1 %) and the *emotion cause* extraction task (+3 ROS points) when compared to the existing state-of-the-art model.

**Keywords:** cognitive distortion, emotion cause, mental health, angular momentum, multi-modal, multi-task, attention, conversations

## 1 Introduction

The World Health Organization (WHO) estimated a cost of $1 trillion per year in lost productivity due to depression and anxiety disorders[1]. The COVID-19 pandemic's trauma has also exacerbated the world's mental health crises. Negatively biased errors in thinking, also known as *Cognitive Distortion* [8] is a major

[*] These authors contributed equally to this work and are joint first authors.
[1] https://www.who.int/teams/mental-health-and-substance-use/promotion-prevention/mental-health-in-the-workplace

contributor to the development of many different mental illnesses. Short-term use may help with stress and boost confidence, but chronic use can lead to mental decline and the onset of feelings of depression and anxiety [33]. Cognitive distortion manifests itself in a variety of ways, and the study in [4] found ten main manifestations of the same. Mindreading, catastrophizing, all-or-nothing thinking, emotive reasoning, labeling, mental filtering, overgeneralization, personalization, should statements, and diminishing or rejecting the positive are examples of these. Given the relevance of the interpersonal context in the start and progression of several mental health conditions [8, 16], early detection of cognitive distortions among individuals may play an important role in the symptomatology of such illnesses.

Humans often exercise some restraint while interacting with one another. But people were more likely to talk about their thoughts and feelings with a virtual therapist than with a real one. In order to create efficient and low-cost interactive systems (like chatbots), which are often quick to install and may be utilized in combination with a human therapist, understanding human emotional states is vital. In order to identify how to best avoid acts of self-harm (such as suicide), it is important to recognize not just the emotional states but also the cause(s) of those feelings. This will allow for a deeper knowledge of the mental health of those involved. In such a situation, the Emotion cause extraction (ECE) task, which seeks to identify the possible causes behind a certain emotion expression in the text, might be useful.

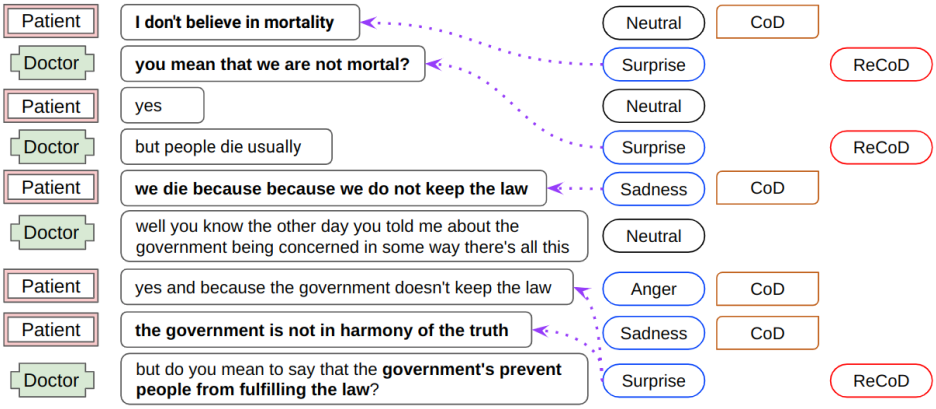


Fig. 1: Sample snapshot of our CoDeC Dataset. CoD: Cognitive Distortion; ReCoD: Response to CoD. The font highlighted in bold is the causal span.

Over the years, several conversational datasets have been introduced on various domains such as TV shows, social media, news, etc., but no such dataset exists, to our knowledge, related to mental health. Also, there is a big shortage of multi-modal datasets that can be used in clinical conversation settings.

The task of emotion-cause extraction in conversations is very nascent, and the existing studies [28, 12] on this topic provide baseline systems that are majorly fine-tuned language models on emotion-cause annotated datasets. Certain features distinguish mental health discussions from other conversational datasets. We see a general trend in the flow of a typical doctor-patient interview. The doctor inquires about numerous elements and situations concerning the patient's well-being, and the patient relates the scenario they have been experiencing. At the end of the session, the doctor provides his diagnosis or remedy for the patient's ailment. In order to identify cognitively warped phenomena in any patient statement, information from future time steps may be required in addition to the context history. Figure 1 illustrates a conversation snippet describing the phenomenon of cognitive distortion, its response to it, and the various types of association of causes for emotions.

We take this opportunity to introduce a high-quality multi-modal clinical conversation dataset and a task-specific framework, especially for the task of emotion cause extraction. The current study analyzes emotion, emotion causes, and cognitive distortion in videos of dyadic conversations between psychiatrists and mental illness patients. The dataset and code is open-sourced to aid research[2].

The main contributions of this work are summarized below:

1. We propose the novel task of Detection of Cognitive Distortion and Emotion cause extraction in clinical conversations.
2. We introduce the first Cognitive Distortion and Emotion Cause (*CoDEC*) annotated multi-modal clinical conversation dataset comprising doctor-patient interactions on the premise of mental health interviews. We also provide manual annotations for Cognitive Distortion and Emotion at the utterance level.
3. We develop an emotion-aware multi-modal multi-task framework for the Detection of Cognitive Distortion and Emotion cause extraction (*DeCoDE*) in Clinical Conversations.
4. We also hypothesize that the performance of the above two tasks can be enhanced by the incorporation of information from future time steps.

The remainder of the paper is structured as follows. Section 2 summarises some previous works in this area. Following that, we go into the dataset preparation in depth in Section 3. We address our suggested methodology for multimodal multitask experiments in Section 4. In Section 5, we discuss the experiments, their results, and their outcomes. Finally, in Section 6, we bring our effort to a close and define the scope of future work.

## 2 Related Work

A few studies using computational methods have focused on the detection of various mental health issues, but none of them have concentrated on identifying

[2] `https://www.iitp.ac.in/~ai-nlp-ml/resources.html#DeCoDE-CoDEC`

the core cause of such difficulties, which is a cognitive distortion in general. The emotion-cause extraction (ECE) problem has also received a lot of attention in several studies, but none of them have focused on the domain of clinical conversations or mental health, leaving a major gap that calls for more research in this area. In this section, we go through some of the earlier studies on mental health as well as emotion-cause extraction techniques.

### 2.1 Mental Health Studies

Mental health illnesses in general, being a major public health issue [25], have gained attention in past research, including computational studies. Although depression has received the most attention, other mental illnesses such as anxiety disorder, schizophrenia, post-traumatic stress disorder, suicide risk, and self-harm have also been studied [31]. Furthermore, psychological research supports the use of multimodal data for developing automated systems to recognize human emotion [1, 26]. The impact of online content for doctor-patient interactions on patient satisfaction was investigated in [2]. The authors in [7] developed a deep learning method to categorize a variety of detrimental mental-health emotions, including addiction, anxiety, despair, stress, etc. While there is growing interest in the subject of the explainability of machine learning models in NLP [13], there is less such research for mental health condition identification. ECE tasks may be the initial step in making any automated system that can aid with mental health condition identification intelligible.

### 2.2 Emotion Cause Extraction

Due to the inherent long-term dependencies present in the utterances, determining the causes of emotions in a conversational situation is a challenging task. First proposed by Lee et al. [20] as a word-level sequence labeling problem, the ECE task was re-formalized in [14] as a clause-level extraction problem. End-to-end networks, such as the one shown in [29], have been proven to provide additional advantages over multi-stage approaches by leveraging the interdependence between the extracted emotion words and cause clauses. Li et al. [22] developed a context-aware co-attention model for the extraction of emotion cause pair. The authors in [3] suggested a strategy that narrows the search field and enhances productivity by matching emotions and causes concurrently utilizing the local search. Emotion-cause recognition in a conversation scenario was initially introduced in [28], which provided an emotion-cause annotated conversation corpus and evaluated it using a pair of deep learning-based systems. In a similar way, Ghosh et al. [12] introduced an emotion-cause annotated suicide note corpus and solved the emotion-cause identification and extraction independently. In this study, we consider the works in [28, 12] as baselines to evaluate our suggested approach and mental health conversation dataset.

## 3 Dataset

In the following subsections, we discuss the various aspects of the developed *CoDEC* dataset.

### 3.1 Data Collection

Mental health-related content is scarcely available in the public domain, mainly due to its sensitive nature of it and also the associated stigma in sharing such content. YouTube is one of the most popular social media sites for sharing videos. It has a wide range of content about mental health, most of which is meant to promote and support educational needs. Applying a certain combination of keywords and phrases[3], we collected 30 doctor-patient conversation sessions[4] where the patients suffer from some form of cognitive distortions (such as polarized thinking, catastrophizing, over-generalization, etc.). Among the collected videos, 13 are from female patients and 17 are from male patients. Twenty of them are genuine interviews with psychiatrists and patients. The remaining 10 interviews are case studies/tutorial films with actual psychiatrists and actors conversing (posing as mental illness patients of various types of mental illness). Because mental health is sensitive and stigmatized, easily available relevant data in the public domain is limited. As a result, we chose to generate the dataset by considering both actual and enacted doctor-patient exchanges. The average number of utterances in the conversations is 125.1 and the average sentence length is 11.41 words.

### 3.2 Data Annotation

Each utterance of the conversations is marked with a start and end time stamp, which is essential to extract video extracts per utterance during multi-modal training. Also, each utterance is marked with speaker information (doctor or patient), the presence of factual information (fact), and a response to cognitive distortion (ReCoD). The annotations for emotion, emotion causes, and cognitive distortion are performed by three annotators[5], and final labels are obtained by performing a majority vote among the individual annotations. Text transcripts of some videos were already available from the uploading source. For the rest, we first collected the auto-generated transcripts in English and manually validated them to correct any inherent errors and produce good-quality transcripts for each utterance of the conversations.

[3] mental health, psychiatric interview, psychotic, paranoia, hallucination, etc.

[4] Links to some sample videos: `https://www.youtube.com/watch?v=P7qMfG-yNfA` `https://www.youtube.com/watch?v=Ii2FHbtVJzc`

[5] 2 Ph.D. linguistics degree holders and 1 undergraduate student from the computer science discipline

**Annotating Cognitive Distortion** Identifying the utterances with cognitive distortion is a challenging task. With a sound understanding of the phenomenon of cognitive distortion and its various forms, the annotators identified utterances as cognitively distorted if they presented biasedness and/or depicted irrational ways of perceiving real-world situations. Doctor responses to patient utterances at various junctures of the conversations presented vital clues to anticipate CoD utterances. We compute the Fleiss-Kappa ($\kappa$) score for the overall inter-rater agreement [30], as it is a popular choice when more than two raters are involved. The cognitive distortion task yielded a score of 0.83 which is considered to be 'almost perfect agreement'.

**Annotating Emotion** Each utterance is marked with one of Ekman's [9] six basic emotions: anger, disgust, fear, joy, sadness, and surprise. We add a *neutral* class to accommodate all other utterances that are out of scope of the Ekman's emotions. Table 1b shows the distribution of utterances over the various emotion classes. We also observe that the dataset has an over-representation of the *others* class. The average Fleiss-Kappa [30] score obtained for the Emotion task is 0.77 which signifies 'substantial agreement' among the annotators.

Table 1: Dataset Details

(a) Frequency of utterances over various attributes. CoD: Cognitive Distortion; ReCoD: Response to CoD

| **Attribute** | **Count** |
|---|---|
| CoD | 743 |
| ReCoD | 410 |
| One Cause | 410 |
| Two Causes | 179 |
| Three Causes | 36 |

(b) Emotion and Cause distribution.

| **Class** | **Count** | **# Causes** |
|---|---|---|
| Anger | 184 | One: 101; Two: 42; Three: 10 |
| Disgust | 77 | One: 49; Two: 22; Three: 2 |
| Fear | 169 | One: 96; Two: 32; Three: 6 |
| Joy | 128 | One: 28; Two: 7; Three: 2 |
| Sadness | 503 | One: 198; Two: 80; Three: 10 |
| Surprise | 176 | One: 78; Two: 24; Three: 2 |
| Neutral | 2516 | No causal spans exists |

**Annotating Emotion Cause** Following the work in [28, 12], we marked the causal spans (cs) for an emotion of each utterance in the dataset. We mark at most 3 causal spans for each utterance as we observed most utterances have single causes and few of them have two or more causes. The final causal span for an utterance $U_t$ is marked using the span-level aggregation approach detailed in [15]. For a target utterance $U_t$, $C_t$ denotes the set of causal spans ($C_t$ = {$cs_1$,$cs_2$,$cs_3$}) for $U_t$. The causal spans are marked from *v+1* utterances, where *v* denotes the number of context utterances of $U_t$ and $v+1^{th}$ utterance is the target utterance $U_t$ itself. We quantify the inter-rater agreement using the macro-F1 metric based on earlier work on span extraction [28, 12], and we get an F1-score of 0.81, indicating that the annotations are of very high quality.

Table 1 shows the various details of the *CoDEC* dataset. The distribution of utterances over the various emotion classes and the number of causal spans per emotion class is shown in Table 1b.

## 4 Methodology

In this section, the *CoDEC* framework is illustrated for Cognitive Distortion detection and Emotion cause extraction from conversations. The system leverages the utterance-level emotion information for which the causes are to be extracted. The overall architecture of the proposed method is shown in Figure 2.

### 4.1 Problem Definition

Given a document $D = [u_1, \cdots u_i \cdots, u_p]$ composed of a sequence of utterances (u), and each utterance can be further decomposed into a sequence of words, represented as $u_i = [word_{i,1}, \cdots term_{i,j} \cdots, term_{i,q}]$, where p indicates the number of utterances in the document, and q denotes the length of the word sequence contained in the utterance. Let $E = [e_1, \cdots e_i \cdots, e_p]$, denotes the utterance-level emotions in the document D. For a target utterance $u_t$, the task objective is to detect whether the utterance is cogntive distortion or not (0 or 1) and extract all possible causal spans for the given emotion $e_t$.

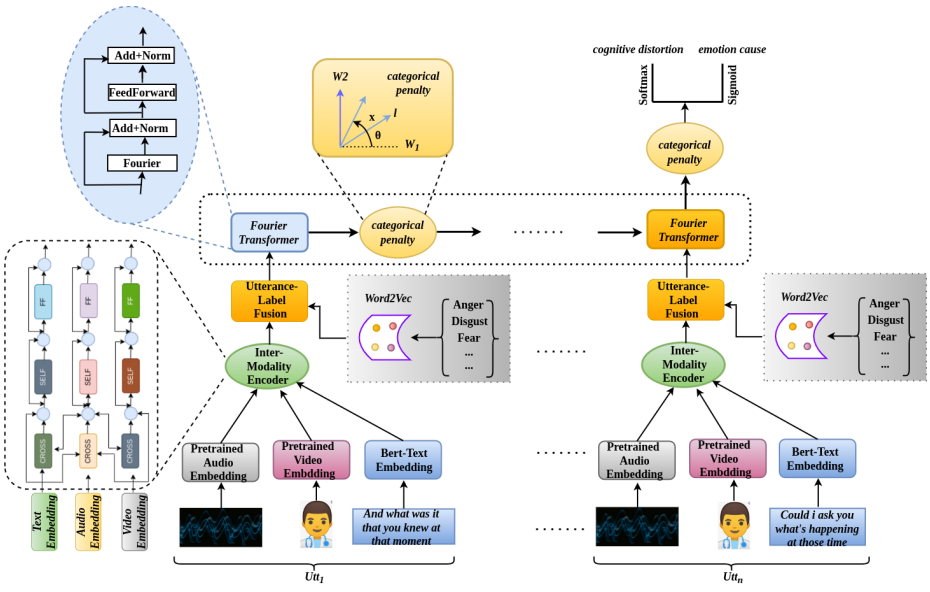


Fig. 2: Architectural diagram of the proposed framework

### 4.2 Detection of Cognitive Distortion and Emotion cause extraction (*DeCoDE*)

We illustrate the various components of the *DeCoDE* method below.

**Input Feature Representation** We generate textual features for the utterances in the conversations using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) [6] due to its strong ability to learn context-sensitive information and its ability to generalize on various downstream tasks. We utilize openSMILE[6][10] tonal low-level features group to extract the acoustic features. We employ 3D-ResNeXt-1014[7] [17] to extract visual features from the video snippets at the utterance level. We fetch the word vectors for the input emotion labels from the Word2Vec [24] pre-trained word embeddings.

**Inter-Modality Encoder (IME)** The inter-modality encoder comprises primarily three self-attention sub-layers, two bi-directional cross-attention sub-layers, and three feed-forward sub-layers. The output of the $r^{th}$ layer is used as the input to the $(r-1)^{th}$ layer, therefore $N$ * inter-modality layers are stacked together in the encoder. Here, we capture the inter-relatedness of the audio and visual counterparts with respect to the textual representation as it is obvious that textual utterance is more contributing than the other modalities. Subsequently, in the $r^{th}$ layer, the bi-directional cross-attention sub-layer is designed to interchange the knowledge and align the features between the said modalities in order to learn the inter-modality representations. In addition, to build internal connections the self-attention sub-layers are then applied to the output of the cross-attention sub-layer. Finally, the $r^{th}$ layer output $\hat{h}_i^k$ and $\hat{a}_i^k$ are produced by feed-forward layers. Residual connections and layer normalization are added after each sub-layer, in a similar manner as the single-modality encoders.

**Utterance-Label Fusion (ULF)** We combine the utterance level semantic features from BERT and the emotion features from Word2Vec through self-attention [32]. The separate feature vectors are passed through independent dense layers to make them the same length. Unlike linear concatenation of emotion labels with corresponding utterances [28], our strategy enables to learn the importance of the emotion for a particular utterance and dynamically update its initial weights during training.

**Contextual Fourier Transformer (CFT)** To model the contextual information among the utterances, we develop Contextual Fourier Transformer. The Fourier Transformer encoder [21] is an efficient method to the regular transformer encoder [32] where the self-attention sublayers in the transformer encoder are replaced with simple linear transformations that 'mix' the input tokens to create

[6] `https://github.com/audeering/opensmile`
[7] `https://github.com/kaiqiangh/extracting-video-features-ResNeXt`

the FNet. Each utterance from the ULF module in the input sequence is passed through the CFT module. Each passing utterance acts as a context for the next utterance up to the target utterance.

**Categorical Penalty** To aid the model in understanding how an emotion label and its related utterance are linked, we add a Categorical Penalty word to the intermediate CFT units. This will enable better prediction ability of the start and end tokens. For this purpose, we first represent softmax and sigmoid by the equations below.

$$\mathcal{L} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\log\frac{\exp^{\mathcal{W}l_i+b_i}}{\sum_{j=1}^{N}\exp^{\mathcal{W}l_j+b_j}} \qquad \mathcal{L} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\frac{1}{\exp^{\mathcal{W}l_i+b_i}} \tag{1}$$

Where $l_i \in \mathbb{R}^d$ is the feature of $i^{th}$ sample. $b_s$ is batch size. $b_i$ and $b_j$ denote the bias. $\mathcal{W} \in \mathbb{R}^{d*n}$ denotes the weight matrix. It is known for information extraction tasks, that finding the decision boundary for the start and end markers of a span is challenging, and a simple softmax/sigmoid classifier will not be able to handle this distinction effectively. Because of this, some samples can fall into the wrong region due to the ambiguity of the classification boundary. This can call for a higher convergence rate. To handle this we use the strategy used in Insightface loss [5] which normalizes the feature $l_i$ and the weight matrics $\mathcal{W}$ to measure the similarity of feature by the difference of angle by which it maps the vector more closely. It adds a penalty value $x$ into the angle to force the feature to converge.

$$\mathcal{L}_{u1} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\log\frac{\exp^{a(cos(\theta+x))}}{\exp^{a(cos(\theta+x))}+\sum_{j=1}^{N}\exp^{a(cos(\theta))}} \tag{2}$$

$$\mathcal{L}_{u2} = -\frac{1}{b_s}\sum_{i=1}^{b_s}\frac{1}{\exp^{a(cos(\theta+x))}+\exp^{a(cos(\theta))}} \tag{3}$$

Where $\mathcal{L}_{u1}$ and $\mathcal{L}_{u2}$ is updated loss function for softmax and sigmoid respectively, $\theta$ denotes the angle between weight $\mathcal{W}$ and feature $l$ and a denotes the amplifier function.

**Task-specific layers:** The output from the last CFT unit which corresponds to the target utterance $U_t$ is passed to two task-specific dense layers and the following output layers for the CoD and ECE tasks. The output layer for the ECE task is a linear layer to calculate span start and end logits which employ sigmoid activation in which the threshold value is set at 0.4. This results in the output of the probability of three first tokens and three last tokens, which signifies the capability to output three causal spans at most.

**Calculation of loss** The model is trained using a unified loss function as shown in equation 4. We employ categorical cross-entropy loss and binary cross-entropy loss for the CoD and ECE tasks, respectively.

$$L = \sum_{\omega} W_\omega L_\omega \tag{4}$$

Here, $\omega$ denotes the two tasks, CoD and ECE. The weights ($W_\omega$) are updated using back-propagation for specific losses for each task.

## 5 Experiments and Results

This section discusses the experiments performed, the results, and the analysis.

### 5.1 Experimental Setup

Since the *CoDEC dataset* is having skewed class proportion, we report both the accuracy and macro-F1 scores for the CoD task. Following the work in [12], we report the full match (FM), partial match (PM), Hamming Distance (HD), Jaccard Similarity (JS) and Ratcliffe-Obershelp Similarity (ROS) measures to evaluate the ECE task. We use PyTorch[8], a Python-based deep learning package, to develop our proposed model. We experiment with the base version of BERT imported from the huggingface transformers[9] package. To determine the optimal value of the additive angle $x$, which influences performance, we tested five values ranging from 0.1 to 0.5. The default value of $x$ is set at 0.3. We set amplification value $a$ as 64. For openSMILE, voice normalization and voice intensity threshold are used to discriminate between samples with and without speech. Z-standardization is used for voice normalizing. ResNext has been pre-trained on Kinetics at 1.5 features per second and a resolution of 112. All experiments are carried out on an NVIDIA GeForce RTX 2080 Ti GPU. We perform 80-20 split of the *DeCoDE* dataset for training and testing purposes. The best model is saved on the performance of the validation set. We run our experiments for 200 epochs and report the averaged scores after 5 runs of the experiments to account for the non-determinism of Tensorflow GPU operations.

**Baselines:** For the comprehensive evaluation of our proposed *DeCoDE* method and the introduced *CoDEC* dataset, we consider the following systems as baselines in this study: RoBERTa [23], SpanBERT [19], MT-BERT [27] and Cascaded Multitask System with External Knowledge Infusion (CMSEKI) [11]. Similar to the *DeCoDE* method, to adapt the baselines to our multi-task scenario, we add a linear layer on top of the hidden-states output in the output layer of the ECE task to calculate span start and end logits. The output layer for the ECE task employs sigmoid activation in which the threshold value is set as 0.4.

### 5.2 Results and Discussion

We investigate the contribution of multi-modal aspects to the tasks at hand. The results of our *DeCoDE* method on the *CoDEC* dataset are shown in Table 2. The trimodal configuration yields the best results, followed by the bimodal and the unimodal networks. This may be due to the fact that texts have less

[8] `https://pytorch.org/`
[9] `https://huggingface.co/docs/transformers/index`

background noise than audio-visual sources, yet when the three are compared, textual modality outperforms the others. For similar jobs, our results are consistent with prior research [18]. We performed experiments using the *DeCoDE* method varying the context length on the *CoDEC* dataset and observe that the best results are obtained when the context length is set as 5. Figure 3 illustrates the detailed results of the *DeCoDE* method on various context length sizes.

Table 2: Experimental results of *DeCoDE* on various modalities

| | ***Cognitive Distortion*** | | ***Emotion Cause*** | | | | |
|---|---|---|---|---|---|---|---|
| ***Modality*** | **F1%** | **Acc. %** | **FM** | **PM** | **HD** | **JF** | **ROS** |
| *T* | 66.68 | 68.71 | 21.98 | 28.31 | 0.45 | 0.58 | 0.69 |
| *A* | 62.69 | 64.11 | 20.74 | 24.46 | 0.41 | 0.53 | 0.68 |
| *V* | 55.96 | 52.13 | 18.29 | 19.31 | 0.37 | 0.48 | 0.61 |
| *T+V* | 68.31 | 69.59 | 25.19 | 29.78 | 0.49 | 0.61 | 0.71 |
| *T+A* | 69.74 | 71.11 | 27.31 | 31.91 | 0.51 | 0.63 | 0.72 |
| *A+V* | 66.22 | 67.63 | 24.33 | 27.58 | 0.47 | 0.59 | 0.70 |
| *T+V+A* | **73.48** | **75.91** | **29.43** | **33.24** | **0.53** | **0.65** | **0.74** |

Table 3: Results from our proposed model and the various baselines. Values in bold are the maximum scores attained.

| | **Cognitive Distortion** | | **Emotion Cause** | | | | |
|---|---|---|---|---|---|---|---|
| **Models** | **F1 (%)** | **Acc. (%)** | **FM** | **PM** | **HD** | **JF** | **ROS** |
| | | | *Baselines* | | | | |
| *RoBERTa* [19] | 67.16 | 69.24 | 25.73 | 25.51 | 0.46 | 0.59 | 0.69 |
| *SpanBERTa* [23] | 65.79 | 66.83 | 23.58 | 21.12 | 0.44 | 0.57 | 0.67 |
| *MTL-BERT* [27] | 66.93 | 69.79 | 25.11 | 23.67 | 0.47 | 0.58 | 0.69 |
| *CMSEKI* [11] | 70.31 | 71.47 | 27.11 | 28.59 | 0.50 | 0.62 | 0.71 |
| | | | *Proposed* | | | | |
| *DeCoDE* | 73.48 | 75.91 | 29.43 | 33.24 | 0.53 | 0.65 | 0.74 |
| $DeCoDE_{-[CP]}$ | 71.25 | 72.35 | 27.22 | 30.89 | 0.50 | 0.62 | 0.72 |
| $DeCoDE_{-[EMO]}$ | 71.76 | 73.17 | 28.18 | 31.17 | 0.51 | 0.63 | 0.71 |
| $DeCoDE_{-[EMO+CP]}$ | 69.47 | 70.85 | 25.91 | 29.33 | 0.49 | 0.60 | 0.69 |
| $\boldsymbol{DeCoDE_{+[ReCoD]}}$ | **74.21** | **76.31** | **30.15** | **34.31** | **0.54** | **0.66** | **0.74** |

**Comparison with Prior Works:** Table 3 shows that CMSEKI is the best-performing baseline, which is not surprising given that it uses commonsense knowledge from external knowledge sources to grasp the input information. However, the proposed *DeCoDE* method outperforms the performances of CMSEKI for all metrics, specifically by 3.17% F1 for CoD task and 3 ROS

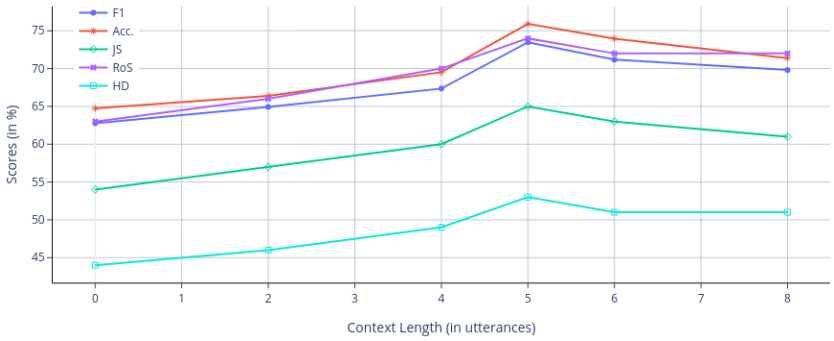


Fig. 3: Results on varying context length

points. Low performances by RoBERTa [23] and SpanBERT [19] shows the difficulty of powerful language models in perceiving critical tasks as emotion cause extraction, more so, especially in clinical situations where training data may be insufficient. We also observe that harnessing the information from future utterances (doctor's responses) enhances the performances of the *DeCoDE* method for both the tasks (as shown by $DeCoDE_{+[ReCoD]}$). This shows the relevance of the information available at future time steps, particularly in the case of clinical conversations, in comprehending mental health-related discussion.

**Ablation Experiments:** As shown in Table 3, we performed an ablation study on the *DeCoDE* dataset to analyze the performance of the different modules in our proposed strategy. The values of all the metrics over both the CoD and ECE tasks are shown to decrease when either the categorical penalty factor ($DeCoDE_{-[CP]}$) or the emotion task ($DeCoDE_{-[EMO]}$) is removed. The decrease is more profound when we remove both the penalty factor and the emotion task from the *DeCoDE* method ($DeCoDE_{-[EMO+CP]}$). This confirms that the inclusion of the categorical penalty factor and the emotion information of the utterances is an integral contributor to the performances of the cognitive distortion and emotion-cause extraction tasks.

**Qualitative Analysis:** We performed an extensive analysis of the predictions from the various systems. It is observed that the proposed DeCoDE performs comparatively better than MTL-BERT and CMSEKI systems in generating correct predictions for both the CoD and ECE tasks. Some sample instances are shown in Table 4. In the first two instances, we can see that the DeCoDE method correctly predicts both the causal spans and the CoD label. The MTL-BERT system extracts an incomplete span for the first example whereas it is unable to extract any part of the cause in the second case. In the third example, we see that both the baselines incorrectly classified the utterance as CoD, however, the DeCoDE method correctly categorized it as non-CoD. Lengthier utterances seem to cause difficulty for all systems, as can be seen from the last example. Although CMSEKI and the proposed DeCoDE are able to predict the CoD label correctly, it manages to extract a part of the causal span fully.

Table 4: Sample predictions from the various systems. Color Coding: Blue- Correct, Red: Incorrect; Teal: Incomplete. [Y] and [N] indicate Yes and No predictions for the CoE task, respectively.

| DeCoDE | CMSEKI | MTL-BERT |
|---|---|---|
| **Actual:** *she might be reporting back to them [Y]* | | |
| she might be reporting back to them [Y] | might be reporting back to [Y] | she might be reporting back to them [Y] |
| **Actual:** *i am a lord god jehovah [Y]* | | |
| i am a lord god jehovah [Y] | No Cause [N] | a lord god jehovah [Y] |
| **Actual:** *you started carrying guns [N]* | | |
| you started carrying guns [N] | you started carrying guns [Y] | you started carrying guns [Y] |
| **Actual:** *try to ensure somehow that they are being raised properly, from a distance [N]* | | |
| to ensure somehow that they are being raised [N] properly | try to ensure somehow that they are being raised properly [Y] | that they are being raised properly, from a distance [N] |

## 6 Conclusion

In this work, we present the first multi-modal, emotion-cause annotated clinical conversation dataset, consisting of conversations between doctors and patients in the context of mental health interviews. Additionally, we present sentence-level manual annotations for cognitive distortion and emotion. In order to extract the emotional causes of cognitive distortions in clinical conversations, we develop, *DeCoDE*, a multi-modal, multi-task framework that takes into account the inherent speaker's emotions present in utterances of any conversations. To the best of our knowledge, the *DeCoDE* framework is the first task-specific system to address the emotion-cause extraction task in conversations. We demonstrate the efficacy of our technique by comparing it to different state-of-the-art baselines.

Even if a negative emotional context has little to do with how the patient feels about other people or things, the patient's behaviours and judgments may be negatively impacted. Future research would concentrate on creating techniques to educate people about the cognitive biases brought on by cognitive distortions in order to provide fresh treatment approaches. It is also important to pay attention to how to properly capture the implicit aspects of complex causation.

## Ethical Consideration

This study has been evaluated and approved by our Institutional Review Board (IRB). The videos used to create the dataset for this study do not have any copyright clauses attached to them. Furthermore, the videos are shared via various channels for the main purpose of facilitating research and educational purposes.

## References


1. Aviezer, H., Trope, Y., Todorov, A.: Body cues, not facial expressions, discriminate between intense positive and negative emotions. Science **338**(6111), 1225–1229 (2012)
2. Chen, S., Guo, X., Wu, T., Ju, X.: Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis. Information Processing & Management **57**(5), 102253 (2020)
3. Cheng, Z., Jiang, Z., Yin, Y., Yu, H., Gu, Q.: A symmetric local search network for emotion-cause pair extraction. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 139–149. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.12, https://aclanthology.org/2020.coling-main.12
4. David, B., Burns, M.: Feeling good-the new mood therapy. NY: Signet Books. Chin, Richard.(1995) p. 3 (1980)
5. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: 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). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/n19-1423, https://doi.org/10.18653/v1/n19-1423
7. Dheeraj, K., Ramakrishnudu, T.: Negative emotions detection on online mental-health related patients texts using the deep learning with mha-bcnn model. Expert Systems with Applications **182**, 115265 (2021)
8. Dozois, D.J., Beck, A.T.: Cognitive schemas, beliefs and assumptions. Risk factors in depression pp. 119–143 (2008)
9. Ekman, P.: An argument for basic emotions. Cognition & emotion **6**(3-4), 169–200 (1992)
10. Eyben, F., Wöllmer, M., Schuller, B.W.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th International Conference on Multimedia 2010, Firenze, Italy, October 25-29, 2010. pp. 1459–1462. ACM (2010). https://doi.org/10.1145/1873951.1874246, https://doi.org/10.1145/1873951.1874246
11. Ghosh, S., Ekbal, A., Bhattacharyya, P.: A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. Cogn. Comput. **14**(1), 110–129 (2022). https://doi.org/10.1007/s12559-021-09828-7, https://doi.org/10.1007/s12559-021-09828-7
12. Ghosh, S., Roy, S., Ekbal, A., Bhattacharyya, P.: Cares: Cause recognition for emotion in suicide notes. In: European Conference on Information Retrieval. pp. 128–136. Springer (2022)
13. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). pp. 80–89. IEEE (2018)

14. Gui, L., Wu, D., Xu, R., Lu, Q., Zhou, Y.: Event-driven emotion cause extraction with corpus construction. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1639–1649 (2016)
15. Gui, L., Wu, D., Xu, R., Lu, Q., Zhou, Y.: Event-driven emotion cause extraction with corpus construction. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. pp. 1639–1649. Association for Computational Linguistics, Austin, Texas (Nov 2016). https://doi.org/10.18653/v1/D16-1170, https://aclanthology.org/D16-1170
16. Hammen, C.L., Shih, J.: Depression and interpersonal processes. (2014)
17. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 6546–6555. Computer Vision Foundation / IEEE Computer Society (2018). https://doi.org/10.1109/CVPR.2018.00685, http://openaccess.thecvf.com/content_cvpr_2018/html/Hara_Can_Spatiotemporal_3D_CVPR_2018_paper.html
18. Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L., Zimmermann, R.: Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers). pp. 2122–2132. Association for Computational Linguistics (2018). https://doi.org/10.18653/v1/n18-1193, https://doi.org/10.18653/v1/n18-1193
19. Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2020). https://doi.org/10.1162/tacl_a_00300, https://aclanthology.org/2020.tacl-1.5
20. Lee, S.Y.M., Chen, Y., Huang, C.R.: 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. pp. 45–53 (2010)
21. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms. arXiv preprint arXiv:2105.03824 (2021)
22. Li, X., Song, K., Feng, S., Wang, D., Zhang, Y.: 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. pp. 4752–4757. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1506, https://aclanthology.org/D18-1506
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019), http://arxiv.org/abs/1907.11692
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)
25. Organization, W.H., et al.: Depression: A global crisis. world mental health day, october 10 2012. World Federation for Mental Health, Occoquan, Va, USA (2012)
26. Pantic, M., Sebe, N., Cohn, J.F., Huang, T.S.: Affective multimodal human-computer interaction. In: Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005. pp. 669–676.

ACM (2005). https://doi.org/10.1145/1101149.1101299, https://doi.org/10.1145/1101149.1101299
27. Peng, Y., Chen, Q., Lu, Z.: An empirical study of multi-task learning on BERT for biomedical text mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020. pp. 205–214. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.bionlp-1.22, https://doi.org/10.18653/v1/2020.bionlp-1.22
28. Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S.Y.B., Hong, P., Ghosh, R., Roy, A., Chhaya, N., et al.: Recognizing emotion cause in conversations. Cognitive Computation **13**(5), 1317–1332 (2021)
29. Singh, A., Hingane, S., Wani, S., Modi, A.: 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. pp. 84–91 (2021)
30. Spitzer, R.L., Cohen, J., Fleiss, J.L., Endicott, J.: Quantification of agreement in psychiatric diagnosis: A new approach. Archives of General Psychiatry **17**(1), 83–87 (1967)
31. Uban, A.S., Chulvi, B., Rosso, P.: Understanding patterns of anorexia manifestations in social media data with deep learning. In: Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access. pp. 224–236. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.clpsych-1.24, https://aclanthology.org/2021.clpsych-1.24
32. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
33. Yüksel, A., Bahadir-Yilmaz, E.: Relationship between depression, anxiety, cognitive distortions, and psychological well-being among nursing students. Perspectives in psychiatric care **55**(4), 690–696 (2019)