COMMA-DEER: COmmon-sense aware Multimodal Multitask Approach for Detection of Emotion and Emotional Reasoning in Conversations

Soumitra Ghosh[1*], **Gopendra Vikram Singh**[1*], **Asif Ekbal**[1] and **Pushpak Bhattacharyya**[2]
[1]Department of Computer Science and Engineering, IIT Patna, India
[2]Department of Computer Science and Engineering, IIT Bombay, India
ghosh.soumitra2@gmail.com, {gopendra_1921cs15,asif}@iitp.ac.in, pb@cse.iitb.ac.in *

Abstract

Mental health is a critical component of the United Nations' Sustainable Development Goals (SDGs), particularly Goal 3, which aims to provide "good health and well-being". The present mental health treatment gap is exacerbated by stigma, lack of human resources, and lack of research capability for implementation and policy reform. We present and discuss a novel task of detecting *emotional reasoning (ER)* and accompanying *emotions* in conversations. In particular, we create a first-of-its-kind multimodal mental health conversational corpus that is manually annotated at the utterance level with emotional reasoning and related emotion. We develop a multimodal multitask framework with a novel multimodal feature fusion technique and a contextuality learning module to handle the two tasks. Leveraging multimodal sources of information, commonsense reasoning, and through a multitask framework, our proposed model produces strong results. We achieve performance gains of 6% accuracy and 4.62% F1 on the *emotion* detection task and 3.56% accuracy and 3.31% F1 on the *ER* detection task, when compared to the existing state-of-the-art model.

1 Introduction

Each year, mental illness costs the economy almost 12 billion working days. Mental diseases are predicted to cost the global economy $16 trillion by 2030 (Canady, 2018). Despite extensive evidence of the close association between mental health and nearly every major issue in development, people with mental illnesses are among those most at risk of being left out of development programmes. Promoting mental health and preventing mental illness should be part of larger efforts to meet the United Nations' Sustainable Development Goals.

Unlike object identification tasks, emotion perception is significantly impacted by personal bias, cultural backgrounds, and contextual information (such as the environment). Emotional information can function as an authoritative inner voice in those who suffer from anxiety, despair, guilt, or fear on a daily basis (Caprara and Cervone, 2000). Normal individuals can suppress or remove the mood influence, while clinical patients cannot stop the inner voice from being viewed as authoritative. In spite of contradictory empirical facts, someone may draw the conclusion that something is true based only on their emotional response. This process is known as *emotional reasoning* (Gangemi et al., 2013). For example,

- *I feel insecure about my wife, which means she must be cheating on me.*
- *I feel afraid, so my neighbors must have been spying on me.*

While automatic emotion recognition in conversations can facilitate developing intelligent systems, such as empathetic chatbots, customers' feedback assessment, etc., automatic identification of emotional reasoning from conversations can help to perceive the mental states of the involved persons and understand latent vulnerability factors that often raise the likelihood of self-harm (such as suicide); thus helping to determine preventive efforts.

A significant increase in tele-health usage seeking mental health services was observed during the peak of the COVID-19 pandemic (Koonin et al., 2020). Despite advancements in certain nations, persons with mental illnesses frequently face human rights abuses, discrimination, and stigma. Automated systems based on natural language processing (NLP) approaches can be incorporated into digital interventions, particularly web and smartphone apps (chatbot like WoeBot[1], virtual assistant like Ellie[2], etc.) to provide early identification, take preventive measures and provide personalized

[*] The first two authors contributed equally to this work and are jointly the first authors.

[1] https://woebothealth.com/
[2] https://youtu.be/ejczMs6b1Q4

Proceedings of the 29th International Conference on Computational Linguistics, pages 6978–6990
October 12–17, 2022.

healthcare, more so in developing countries with a high population and minimal healthcare facilities.

The lack of suitable annotated data in the public domain is a serious impediment to mental health research utilising NLP methods. Also, there are certain nuances in mental health conversations (shown in Figure 1[3]), that makes them challenging to be addressed using the existent automated systems. The first set of images in the figure displays examples of how visual signals for current textual utterances can be found in past time steps. The second series of images depicts a patient's uneven bodily motions while discussing ER experiences at various points in the discussions. Furthermore, identifying ER utterances is challenging since the automated system must have commonsense reasoning ability to distinguish factual information from others.

To this end, we introduce a novel task of detection of emotion and emotional reasoning in conversations in a multitask setting. We develop a COmmonsense aware Multimodal Multitask Approach for Detection of Emotion and Emotional Reasoning (*COMMA-DEER*) in conversations at the utterance level. We create the *DEER* corpus (Detection of Emotion and Emotional Reasoning), which is a multimodal doctor-patient conversational dataset involving various common mental illnesses annotated with emotion and *ER* at the utterance level. We compare our proposed method to various existing state-of-the-art techniques to the presented dataset. Empirical evaluation and qualitative analysis show strong performances by our approach compared to the prior works on both tasks. We intend to make the code and data available to facilitate future research in this domain.

The main contributions of this work are:

1. We introduce a novel problem of joint detection of emotional reasoning and emotion in conversations exploiting the correlatedness between the two tasks.
2. This work introduces the first multimodal mental health conversational corpus, *DEER*, manually annotated with emotion, presence of emotional reasoning, speaker information, start and end timestamps at the utterance level.
3. We propose *COMMA-DEER*, a commonsense aware multimodal multitask system for detection of emotion and emotional reasoning utterances in a conversational setting.

[3]The images used are not subjected to copyright as the source videos have been made available for 'teaching purposes' and 'medical profession and allied scientific groups'.

The rest of the paper is organised as follows. Section 2 summarises some of the prior efforts in this subject. Following that, in Section 3, we discuss the dataset preparation in detail. In Section 4, we discuss our proposed methodology for multimodal multitask experiments. We explain the experiments, results and their outcomes in Section 5. Finally, in Section 6, we conclude our work and identify the scope of future work.

2 Related Work

Arntz et al. (1995)'s study was one of the first clinical investigations to show that the affect-as-information hypothesis might be used to sustain anxiety and mood disorders. The authors in (Meeten and Davey, 2011) found that the emotional reasoning phenomenon is common among clinical populations due to its adverse negative impact. The use of ER as a source for measuring risk was investigated by Beck and Haigh (2014), who discovered that it has significant clinical relevance both as a disorder maintenance factor and as a treatment target in cognitive therapy.

The research community has focused heavily on utterance and document level emotion recognition (Mohammad et al., 2018; Bostan and Klinger, 2020). Psychology research (Pantic et al., 2005; Aviezer et al., 2012) also points to the importance of considering multimodal information to build automated systems to perceive human emotion. Given the significance of mental health and its growing influence on society, researchers (Pham et al., 2016) are currently developing methods to precisely detect human emotion in the intention of developing mental health therapeutic solutions and better understand mental health issues.

Recent works (Alswaidan and Menai, 2019; Ghosh et al., 2022) have considered external knowledge resources (ConceptNet (Speer et al., 2017), SenticNet (Cambria et al., 2018), etc.) in building systems for emotion detection in various domains (conversations, suicide notes, etc.). In mental health conversations such as doctor-patient interactions, any factual information or presence of any grounded knowledge, such as a doctor's description of a patient's mental condition or characteristics of a particular illness (examples shown in Section 3.2.1), can also be exploited to infuse additional knowledge into the context of the dialogue.

Data for multi-modal emotion detection is challenging to collect and annotate, especially for low-

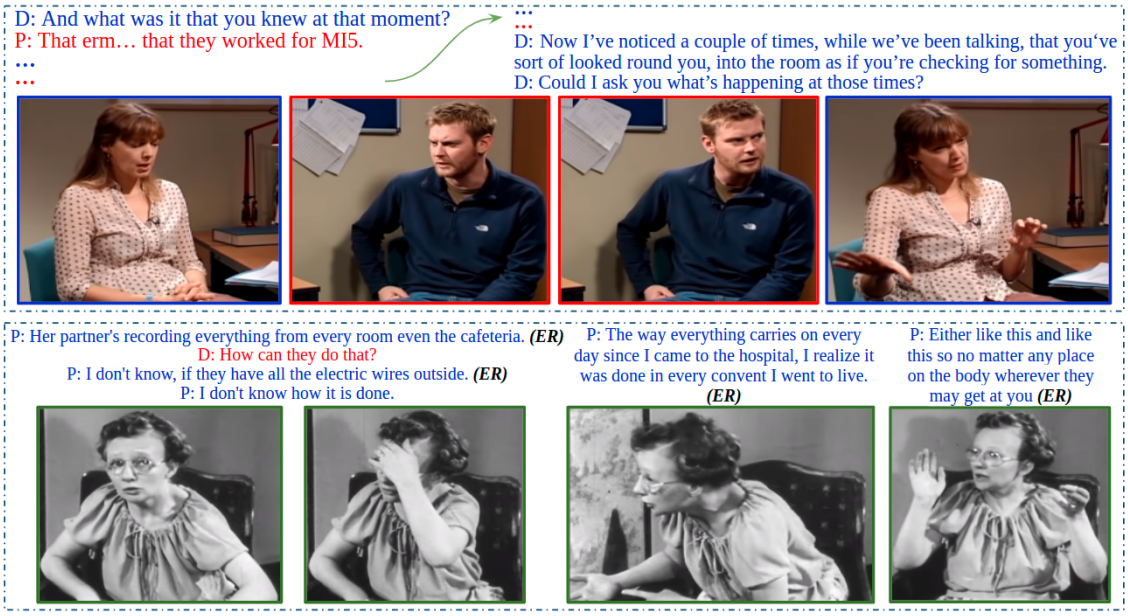

Figure 1: Sample instances from our *DEER* dataset, showing ER utterances and demonstrating how multimodality may capture various nuances in dialogues. ER: Emotional Reasoning.

resource emotions (e.g., disgust, surprise) that are encountered seldom in everyday life, which drives us to investigate this problem. Building on existing studies and the limitations of the prior works on related topics, we presented a manually annotated multimodal mental health conversational dataset and devised an automated approach leveraging commonsense knowledge to detect emotions and emotional reasoning utterances in conversations.

3 Dataset

In this section, we discuss the data collection and annotations for various attributes at the utterance level, such as speaker information, start and end timestamps, emotion, ER, and factual utterances.

3.1 Data collection and processing

We collected 30 doctor and patient interviews from YouTube[4]. Twenty of them are real interviews of different psychiatrists and various mental illness[5] patients. The rest of the ten interviews involves case studies/tutorial videos of conversations between the real psychiatrists and actors (posing as mental illness patients of various types of mental illness). Due to the sensitivity and stigma associated with mental health, readily available relevant data in the public domain is scarce. Hence, we decided to consider both the real and enacted doctor-patient conversations to create the dataset. Transcripts are manually generated for each video (wherever not available on YouTube) and marked with speaker information (*Doctor* or *Patient*) and start and end timestamps for each utterance. In a conversational video, an utterance is defined as a unit of speech bound by breaths or pauses (Hazarika et al., 2018).

3.2 Data Annotation

The annotations for the ER and emotion classes are performed at the utterance level by three annotators (one undergraduate student from the computer science discipline and two doctoral researchers from the computational linguistics discipline).

3.2.1 ER and Emotion Annotations

The annotations for ER labels were left to the annotators' discretion, based on their understanding of the phenomena of ER. Utterances were labeled as ER if they were detected as an obvious result of emotional reactions (inaccurate emotional truths) that directly contradicted any objective and/or perceptual realities. Such statements either lack counter empirical evidence or cannot be supported by common sense reasoning or common knowledge. Also, the doctor's interaction with the patient in the conversations provided essential clues and hints for ER to the annotators. When hearing a patient's emotional justification, doctors frequently express astonishment and are perplexed

[4]Some of the video ids are: *ZB28gfSmz1Y*, *8gDkFX4wprI*, *GGVYRxxsvEU*, *f744UFJSuog*

[5]*Psychosis, Schizophrenia, Paranoid Delusions, Delirium, Obsessive Compulsive Disorder, etc.*

or shocked. The annotators could easily identify strong candidate utterances for ER from the doctors' expressions, comments, vocal tone, etc. This can be understood from the below conversation snippet between a doctor and a patient, taken from our introduced DEER dataset.

Doctor: *Well you know, the other day you told me about the government being concerned in some way there's all this* [Emotion: Others]
Patient: *yes, and because the government doesn't keep the law* [Emotion: Others] –> ER
Patient: *the government is not in a harmony of the truth* [Emotion: Others] –> ER
Doctor: *but do you mean to say that the government prevents people from fulfilling the law?* [Emotion: Surprise]

Each utterance is marked with one of Ekman's (Ekman, 1992) six basic emotions: anger, disgust, fear, joy, sadness, and surprise. The necessity to add another class arose when the annotators encountered instances that have no emotion or some non-neutral emotion that do not fall into the scope of Ekman's basic emotions. We name this class as *others*. Table 1 shows the data distribution across the various emotion classes. We observe that most of the emotionally charged utterances are from patients' utterances, and their predominant emotions are anger, fear, and sadness. The majority of the doctor's utterances bear the surprise emotion as more than not; the doctors are puzzled or in disbelief on hearing the patients' emotional reasoning. We also observe that the dataset has an over-representation of the *others* class.

Figure 2 shows a snippet from the dataset, depicting ER and various associated emotions in different utterances from a doctor and patient's dyadic conversation. We show the data distribution over the emotion classes and ER categories for both doctors and patients in Table 2. We obtained 743 ER utterances from the total of 3,753 annotated utterances. An example of factual utterances from the dataset is shown below:

- 60-year-old professional woman with auditory and olfactory symptoms that feature in her persecutory delusions.

Inter-rater Agreement: We compute the Fleiss-Kappa (κ) score for the overall inter-rater agreement (Spitzer et al., 1967), as it is a popular choice when more than two raters are involved. The Emotion and ER tasks yielded scores of 0.75 and 0.83,

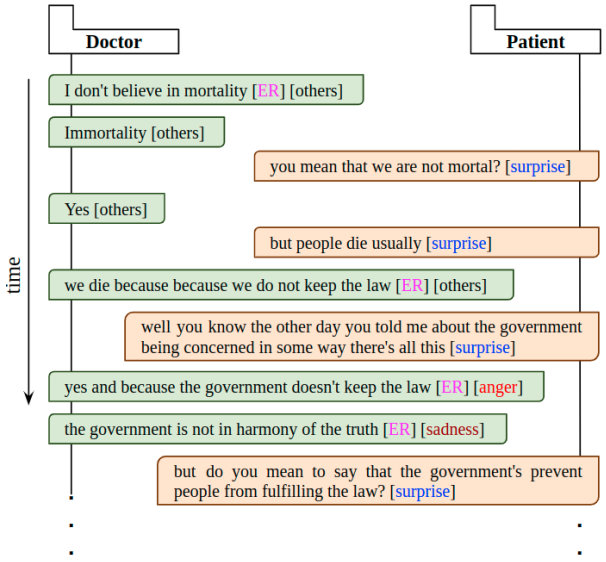

Figure 2: Sample conversation excerpt from the annotated *DEER* dataset.

Classes	Instances	κ
Anger	171	0.71
Disgust	69	0.65
Fear	163	0.74
Joy	124	0.90
Sadness	494	0.82
Surprise	174	0.58
Others	2558	0.86

Table 1: Emotion distribution.

	Doctor	*Patient*	**Total**
Emotion	217 [5]	978 [351]	1195 [356]
Others	1250 [42]	1308 [345]	2558 [387]
Total	1467 [47]	2286 [696]	3753 [743]

Table 2: Distribution of Emotion utterances and ER for doctor and patients. Values in the brackets indicate the ER counts.

respectively. According to the definition of the Fleiss' Kappa statistic (Landis and Koch, 1977), the obtained inter-rater reliability is considered to be 'almost perfect agreement' for the ER task and 'substantial agreement' for the emotion task. We also show the average per-class agreement among the annotators for each emotion in Table 1.

4 Methodology

This section formalizes our task objective and discusses our proposed *COMMA-DEER* approach. Figure 3 illustrates the general architecture of our proposed approach.

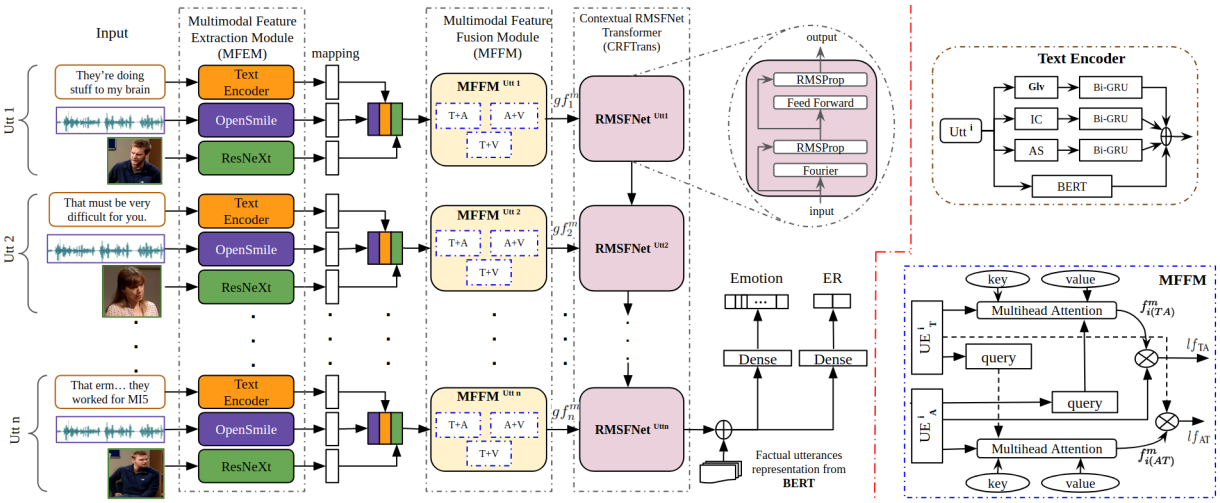

Figure 3: Architecture of the proposed *COMMA-DEER* approach. The model takes multi-modal inputs at the utterance-level and for a particular target utterance, it outputs the presence or absence of ER and the associated.

4.1 Task Definition

Our task involves a dyadic interaction between two speakers, a doctor and a patient, in which utterances are exchanged asynchronously. The objective is to identify emotion and the existence of emotional reasoning in the utterances. Let $U_t^m = (u_1^m, u_2^m, .., u_n^m)$ denote a conversation, $E_t^m = (e_1^m, e_2^m, ..., e_n^m)$ represent the corresponding emotion labels for each utterance in the conversation and $ER_n^m = (er_1^m, er_2^m, ..., er_n^m)$ depicts the presence or absence (0 or 1) of ER at the utterance level. In the m^{th} dialogue, n refers to the total number of utterances. The purpose of this task is to maximize the value of the following function:

$$\underset{\theta}{argmax}(\Pi_{i=0}^{m}\Pi_{j=0}^{n}P(e_j^i, er_j^i|u_j^i, u_{j-1}^i, ..., u_1^i; \theta))$$

where $u_{j-1}^i, ..., u_1^i$ represents the previous set of utterances ordered temporally and u_j^i is the current (target) utterance whose emotion label (e_j^i) and presence of emotional reasoning (er_j^i) is to be predicted. P is the log likelihood function and θ denotes the model parameters to be optimized.

4.2 COmmon-sense aware Multimodal Multitask Approach for Detection of Emotion and Emotional Reasoning (*COMMA-DEER*)

Our proposed approach can be considered a pipeline of 4 primary modules: a). Multimodal Feature Extraction Module (MFEM), b). Multimodal Feature Fusion Module (MFFM), c). Contextual RMS-Fourier Transformer (CRFTrans), and, d). Task-specific Layers.

4.2.1 Multimodal Feature Extraction Module (MFEM)

Text Encoder: For an automated system to be able to identify *ER* in a running conversation, it requires a significant capability to perform commonsense reasoning to comprehend situations that relate to the real world or which are non-factual-based. We generate textual features from the utterances from multiple encoding sources. First, we generate a feature vector using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019) due to its strong ability to learn context-sensitive information and its ability to generalize on various downstream tasks. Next, we generate three more feature vectors (as shown in Figure 3) for the same utterance using multiple common-sense knowledge resources: GloVe (Pennington et al., 2014), IsaCore (Cambria et al., 2014) and AffectiveSpace[6] (Cambria et al., 2015). The IsaCore and AffectiveSpace are vector spaces[7], each providing 100 dimension embeddings for the most frequent occurring words in English. This setup is loosely motivated by the work in (Ghosh et al., 2022), except the incorporation of the BERT encoder. Since we are dealing with a relatively smaller supervised dataset, BERT, which is trained on a large scale corpus enables to generate feature-rich contextual representations of the input.

A similar approach of commonsense knowledge infusion was presented by Ghosh et al. (2022). Unlike (Ghosh et al., 2022), who only exploited the

[6] https://sentic.net/downloads/

[7] In Section A.3 of the **Appendix**, we present a broad discussion of the knowledge resources.

uni-gram word vectors, we extracted the uni-gram, bi-gram, and tri-gram vectors for the tokens of our utterances. Finally, the generated resource-independent sentence vectors are linearly concatenated and passed on to the next module.

Audio Encoder: For acoustic feature extraction, we utilise openSMILE (Eyben et al., 2010). openSMILE can extract Low-Level Descriptors (LLD) and manipulate them using various filters, functionalities, and transformations.

Video Encoder: Rich emotional indications are provided by facial expressions and the visual environment. We use 3D-ResNeXt[8] (Hara et al., 2018) to capture the facial expressions and visual surroundings from the utterance video, which provides rich emotional indicators.

4.2.2 Multimodal Feature Fusion Module (MFFM)

The separate feature vectors from the three modalities are passed through separate dense layers (mapping) to make them the same length. Instead of directly concatenating the features from the multiple modalities and use it for classification, we enhance the feature vector of each modality in a pair-wise approach using the features from the other modalities. For example, to enhance the textual features (${UE^m}_{\mathrm{T}}$) with the help of the supporting acoustic information, we augment the acquired acoustic feature vectors (${UE^i}_{\mathrm{A}}$) to the text representation using the multi-head attention technique. We do the same for the acoustic feature enhancement (${UE^m}_{\mathrm{A}}$) using the textual features (${UE^m}_{\mathrm{T}}$). Likewise, we do the same for the two other paired modalities: text-video and audio-video. So, for the text-audio scenario, we obtain $f^m_{i(TA)}$ and $f^m_{i(AT)}$ as per eq. 7.

$$Key^m_i = W_{K^m_i}\gamma_x + b_{K^m_i} \quad (1)$$

$$Value^m_i = W_{V^m_i}\gamma_x + b_{V^m_i} \quad (2)$$

$$Query^m_i = W_{Q^m_i}\gamma_x + b_{Q^m_i} \quad (3)$$

$$Att(Query, Key, Value) = softmax(\frac{QK^{tra}}{\sqrt{d}})V \quad (4)$$

$$head^m_i = Att(Query^m_i, Key^m_i, Value^m_i) \quad (5)$$

$$Multihead^m_i = (head_1, head_2, ..., head_{num})^m_i \quad (6)$$

$$f^m_{i(ab)} = Multihead(UE_{\mathrm{a}}, UE_{\mathrm{b}}, UE_{\mathrm{b}})^m_i; \quad (7)$$

Here, γ is the input state vector, and $W_{K^m_i}$, $W_{V^m_i}$, and $W_{Q^m_i}$ are the corresponding weight matrices to transform γ into the Query, Key, Value vectors of the i[th] utterance in the m[th] conversation. We use Xavier-norm to initialize the parameters. Here, {a,b} ϵ {text (T), audio (A), video (V)}. *tra* denotes matrix transpose operation.

We design a gating mechanism to filter out noise[9] and as well as attend to features that are well represented in some modalities but not in the others[10]. The gating ($\bigotimes$) operation can be realized from the equations 8, 9 and 10. The equations relate to the gate on top of the MFFM module as shown in Figure 3 (component enclosed with blue dotted line). The sigmoid operation acts a forgetting mechanism for a particular modality (say T) based on the current input of another modality (say A) and vice-versa. The following dot product calculates the similarity measure between the resultant output (for T) from the previous step and the actual features of the other modality (of A). A tanh activation is applied on the outputs for non-linearity. This operation occurs for each pair-wise modality (T+A, T+V, A+V), thus producing a set of six features for each utterance which we linearly concatenate to produce an enriched feature vector (eq. 11).

$$g^m_{(TA)i} = sigmoid(W_g f^m_{(TA)i} + V_g {UE_{\mathrm{A}}}^m_i + b_g) \quad (8)$$

$$P^m_{(TA)i} = dot(g_{(TA)}, f_{(TA)})^m_i \quad (9)$$

$$lf^m_{i(TA)} = tanh(W_{lf} P^m_{(TA)} + U_{lf} {UE_{\mathrm{A}}}^m_i + b_{lf}) \quad (10)$$

$$gf^m_i = [lf_{\mathrm{TA}}; lf_{\mathrm{AT}}; lf_{\mathrm{TV}}; lf_{\mathrm{VT}}; lf_{\mathrm{AV}}; lf_{\mathrm{VA}}]^m_i \quad (11)$$

where $W_g, V_g, b_g, W_{lf}, V_{lf}, b_{lf}$ are the weight matrices that get updated during training. We use xavier-norm to initialize the parameters. [;] denotes concatenation.

4.2.3 Contextual RMS-Fourier Transformer (CRFTrans)

To model the contextual information among the utterances, we develop Contextual RMS-Fourier Transformer (*CRFTrans*). It uses the Fourier Transformer (FNet) encoder, originally introduced by Lee-Thorp et al. (2021), which is a simpler yet efficient method compared to regular transformer

[8] https://github.com/kaiqiangh/extracting-video-features-ResNeXt

[9] such as poor video/audio quality

[10] (for example, some clips are too small to generate the audio clips but text is available

Modality	*Single-Task*		*Multitask*	
	$F1^{ER}$	$F1^{Emo}$	$F1^{ER}$	$F1^{Emo}$
T	62.78	56.11	64.79	60.78
A	57.76	54.98	60.43	57.91
V	52.16	51.51	54.11	53.67
T+V	64.97	61.11	66.79	64.55
T+A	66.18	62.41	68.91	65.72
A+V	61.66	58.54	63.94	62.59
T+V+A	69.28	63.17	71.82	66.91
[T+V+A]$_{+CS}$	**72.14**	**66.33**	**73.44**	**70.14**

(a) Results of the proposed *COMMA-DEER* method on various modalities. Values in bold are the maximum scores attained. CS: common-sense.

Models	Emotion		ER	
	F1 (%)	Acc. (%)	F1 (%)	Acc. (%)
Singletask baselines				
bc-LSTM (Poria et al., 2017)	58.41	59.63	60.15	60.87
CMN (Hazarika et al., 2018)	62.64	64.52	68.17	69.31
DialogueRNN (Majumder et al., 2019)	63.92	64.16	68.67	70.13
Multitask baselines				
MTL-BERT (Peng et al., 2020)	62.33	63.51	65.61	65.91
CMSEKI (Ghosh et al., 2022)	64.14	66.31	69.88	71.31
COMMA-DEER* *(proposed)*	**70.14**	**70.93**	**73.44**	**74.62**

(b) Results from our proposed model and the various baselines. Values in bold are the maximum scores attained. Results marked with a * are statistically significant above the best performing CMSEKI baseline by the Student's t-test ($p < 0.05$).

Table 3: Results of the baselines and the proposed *COMMA-DEER* approach on the *DEER* dataset.

encoder (Vaswani et al., 2017). The self-attention sublayers in the transformer encoder are replaced with simple linear transformations that 'mix' the input tokens to create the FNet. Both the sequence dimension (seq) and the hidden dimension (hid) are transformed using 1D Fourier transforms (F): $\mathbb{R}(F_{\text{seq}}(F_{\text{hid}}(x)))$

One distinctive modification that we do in the FNet model is to use RMS-layer norm (Zhang and Sennrich, 2019) instead of the normal layer norm. The RMSNorm simplifies LayerNorm by removing the mean-centering operation or normalizing layer activations with RMS (Root Mean Square) statistic ($RMS(a) = \frac{1}{n}\sqrt{\sum_{i=1}^{n} a_i^2}$). To the best of our knowledge, we are the first to use RMSFNet as a contextual information learning module in conversations. Each utterance from the MFFM module (uf_i^m) in the input sequence is passed through the *CRFTrans* unit. Each passing utterance acts as a context to the next utterance up to the target utterance. So, $UF_i^m = [uf_1^m, uf_2^m, ..., uf_{n-1}^m] \in \mathbb{R}^{n-1,d}$ acts as the context for uf_n^m and its subsequent output from *CRFTrans* is passed for classification to the task-specific layers.

4.2.4 Task-specific layers:

We extract additional features exploiting the factual utterances (grounded knowledge) in the conversations to enhance the fused feature output of the CRFTrans module. We use a BERT encoder to create contextual features from the factual utterances associated with a dialogue to which the current [context+target] utterances belong. We linearly concatenate the BERT output with the *CRFTrans* output and feed it to two task-specific dense layers that capture emotion and ER identification task-specific characteristics. The output of the task-specific dense layers is sent to two output dense layers, which serve as output classification layers.

Calculation of loss: We adopt a principled approach to calculate the loss of our multitask approach that considers the homoscedastic uncertainty (Kendall et al., 2018) of each task while weighing multiple loss functions.

$$L = \sum_{\omega} W_\omega L_\omega \tag{12}$$

Here, ω denotes the two tasks, emotion, and emotional reasoning detection. The weights (W_ω) are updated using back-propagation for specific losses for each task. We use the categorical cross-entropy loss and binary cross-entropy loss for the emotion and ER tasks.

5 Experiments and Results

In this section, we discuss the experiments performed, and present the results and analysis.

5.1 Experimental Setup

We evaluate our proposed approach against five state-of-the-art systems: Hierarchical Attention Networks (bc-LSTM) (Poria et al., 2017), Conversational Memory Network (CMN) (Hazarika et al., 2018), DialogueRNN (Majumder et al., 2019), Mulitask BERT (MT-BERT) (Peng et al., 2020), and Cascaded Multitask System with External Knowledge Infusion (CMSEKI) (Ghosh et al., 2022). As class imbalance problem persists in the dataset, we performed 5-fold cross-validation experiments and used the macro-F1 metric to evaluate the efficacy of our method against the various baselines. We discuss the details of the baselines and the hyperparameters for our experiments in the **Appendix** (Sections A.1 and A.2).

5.2 Results and Analysis

We investigate the importance of multimodal features for our task and also investigate the role of common-sense infusion in the learning process. Table 3a presents the results for different combinations of modes used by *COMMA-DEER* on *DEER* dataset. Best performances are obtained from the trimodal setup, followed by the bimodal and unimodal networks. Textual modality performs better than the audio and visual modalities when considered alone, which may be due to the presence of lesser noise in texts compared to audio-visual sources. Our observations are consistent with the previous findings (Hazarika et al., 2018) on comparable tasks. Also, we observe that in all the scenarios, the multitask systems outperformed the single-task variants.

Comparison with Prior Works: For a comprehensive evaluation of our proposed approach, we consider various state-of-the-art systems as baselines and observe (from Table 3b) that the proposed *COMMA-DEER* system commendably outperforms all of the baseline systems on both the tasks of emotion and ER detection. Also, we observe that the multitask systems (MTL-BERT and CMSEKI) obtain better scores than the single-task setups stressing the importance of learning the two associated tasks jointly. More so, the CMSEKI system which leverages common-sense knowledge in its training process, performs overall best among the considered baselines. However, our proposed COMMA-DEER approach attains strong performance improvements of 6% accuracy (Acc.) and 4.62% F1 on the emotion detection task and 3.56% accuracy and 3.31% F1 on the ER detection task, when compared to CMSEKI model.

Ablation Experiments: To examine the importance of the modules in *COMMA-DEER*, we remove the constituent components, one at a time, and report the results in Table 4. Specifically, we conduct two ablation experiments: first, we replace RMSFnet with FNet ($[T+V+A]_{\text{-RMS}}$) in CRFTrans, and, second, we employ linear concatenation, replacing the proposed MFFM mechanism, to fuse multimodal features ($[T+V+A]_{\text{-MFFM}}$). We observe notable fall in scores when either of these modules are removed from the *COMMA-DEER* framework, especially when we remove the MFFM module.

Varying Context Length: We trained *COMMA-DEER* for the following context lengths (ψ): 0, 2, 4, 5, 6 and 8. The obtained scores are depicted in Figure 4. 0 indicates no context and the target utterance is given as input to the model only. By increasing the number of past utterances, we see a consistent increase in performance. We obtain the best results when the ψ is set as 5. On analysis of some ER cases with context length 5, we observed frequent topic drifting when context length 5, which may explain the result. However, this more exhaustive analysis is needed, preferably on a larger dataset to obtain a concrete understanding of the observation. Adding more context does not give meaningful information; thus degrading performance due to model confusion.

Setup	$F1^{Emo}$ (%)	$F1^{ER}$ (%)
COMMA-DEER	**70.14**	**73.44**
$[T+V+A]_{\text{-RMS}}$	67.11 (-3.03)	70.24 (-3.20)
$[T+V+A]_{\text{-MFFM}}$	65.78 (-4.36)	68.83 (-4.61)

Table 4: Ablation experiment's results. % fall in scores are shown in brackets.

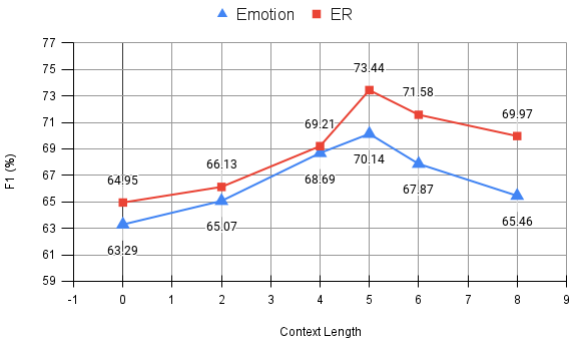

Figure 4: Graphical depiction of results of *COMMA-DEER* on varying context length.

5.2.1 Qualitative Analysis

For comprehensive evaluation of the performance of the proposed method and the considered baselines on the *DEER* dataset, we perform comparative analysis of predictions from these models on several test cases. The first example in Table 5 shows the ability of our model to correctly identify an ER utterance and also the associated emotion. On the contrary, both the CMSEKI and DialogueRNN misclassifies both the emotion and ER labels. In the second example too, we observe that our model classifies both the labels for emotion and ER correctly, unlike the CMSEKI (partially correct) and DialogueRNN (fully incorrect). Here, the target utterance is from the doctor and there is a context change in that utterance itself. Previous utterances taken as context do not entail the current utterance,

Examples	*COMMA-DEER*	CMSEKI	DialogueRNN
D: Right, ok. And have you seen much of your parents recently? *P*: Yeah. I thought about going back there but I, I don't want to drag them into this. *D*: Right. *P*: **I don't want MI5 knowing about them.**	anger ER	others Non-ER	others Non-ER
P: Yes. But in the building? No, I don't think so. As long as I pay my rent, there shouldn't be any problems. I paid my rent on Friday. *D*: **So what happened at work?**	others Non-ER	sadness Non-ER	anger ER
P: You know, jealous talk. *D*: But it was only in your apartment that you would hear these voices? *P*: Yes. *D*: Who was talking? What were they saying? *P*: **If I was doing something, lets say if I was looking for something, they'd taunt me, "she can't see it". They'd know that I was looking for something.**	fear ER	others ER	others Non-ER
P: I wasn't hungry, and I wasn't eating, because it felt tight here. *D*: You also told me about some gas... *P*: Yes, there was a smell coming from the trash chute and the air vents, always at night. *D*: What caused it? *P*: **It was some sort of gas, but I'm not sure what.**	others Non-ER	others Non-ER	others ER
P: As I said: Tuesday, Wednesday, Thursday and Friday and so he came again and he had a mask on and he'd already put something on my pillow. It's difficult to explain. A head, an animal head! To scare me! **I was laying there wondering what was on the pillow next to me.**	surprise Non-ER	others Non-ER	others Non-ER

Table 5: Sample predictions from the best performing baselines and the proposed *COMMA-DEER* approach. Labels highlighted in blue signifies correct predictions and that in red signifies mis-classifications. Target utterances are highlighted with bold font. D: Doctor; P: Patient.

which seems to cause the problem for existing systems. Similarly, in the third example, the CMSEKI and the *COMMA-DEER* predicts the ER label correctly, whereas the correct emotion label was only predicted by our proposed model. Looking at the misclassifications of the baseline systems, we observe a kind of biasness when it comes to the prediction of 'Non-ER' or 'others' class. These pair seems to occur more frequently among the different variations of the misclassifications. We also examined some of the error cases where our proposed method produced partially incorrect output (in example 4) or fully wrong labels (example 5). In example 4, although the three systems correctly predicted the emotion class as 'others', only the DialoueRNN model managed to predict the utterance as an ER. This may be attributed to its strong ability to model inter-speaker dependencies in the conversations, which is not available in *COMMA-DEER* or CMSEKI. The 5th example was misclassfied by all the three systems for both the tasks. Here, the context as well as the target utterance comes from a single speaker (patient). Our proposed method predicted 'surprise' for the utterance where the actual annotated emotion was 'fear'. We believe that the predicted emotion is not unreasonable as there is a possibility for 'surprise' too, in which case, this calls for the scope of extending this work with multilabel emotions, which are also coherent in a real-world conversational scenario.

6 Conclusion

This study presented a novel multimodal multitask system for the detection of emotion and ER in conversations. It also contributes towards mitigating the problem of scarce availability of annotated corpora by introducing a manually annotated multimodal mental health conversational corpus, *DEER*. Empirical and qualitative analysis suggests that (1). most of the existing state-of-the-art systems for conversational data perform poorly when it comes to comprehending ER in conversations; (2). the proposed system performs commendably well on both tasks and attains notable improvements from existing comparable methods; (3). performance on the ER detection task is considerably improved when we simultaneously learn the correlated task of emotion recognition.

Manual observation of the dataset shows that the doctor's responses to the patients' ER usually follow a pattern, such as responses of bewilderment and denouncement. We believe that knowing the responses of the doctor in the future time steps may significantly improve the performance for the detection of ER, which subsequently would enable detection of early warning signs of more serious mental illnesses. In future work, we want to extend this study in the above-mentioned directions as knowledge of the cognitive biases induced by ER may lead to novel therapeutic approaches.

Ethical Consideration

This work develops a resource from publicly available videos of doctor-patient interactions from YouTube. We followed the data usage restrictions and did not violate any copyright issues as the source videos have been made available for 'teaching purposes' and 'medical profession and allied scientific groups'. This study was also evaluated and approved by our Institutional Review Board (IRB). We shall make the code and data available for research purposes (on acceptance), through appropriate data agreement procedure. The findings reported here have been obtained from a small dataset of doctor-patient conversations, which may not accurately represent the phenomenon in all psychopathological disorders. We encourage further research and testing involving clinical participants. *The data is available at* `https://www.iitp.ac.in/~ai-nlp-ml/resources.html#COMMA-DEER`.

Acknowledgement

Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya PhD scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2019. KSU at semeval-2019 task 3: Hybrid features for emotion recognition in textual conversation. In *Proceedings of the 13th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2019, Minneapolis, MN, USA, June 6-7, 2019*, pages 247–250. Association for Computational Linguistics.

Arnoud Arntz, Michael Rauner, and Marcel Van den Hout. 1995. "if i feel anxious, there must be danger": Ex-consequentia reasoning in inferring danger in anxiety disorders. *Behaviour research and therapy*, 33(8):917–925.

Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, 338(6111):1225–1229.

Aaron T Beck and Emily AP Haigh. 2014. Advances in cognitive theory and therapy: The generic cognitive model. *Annual review of clinical psychology*, 10:1–24.

Laura Ana Maria Bostan and Roman Klinger. 2020. Token sequence labeling vs. clause classification for english emotion stimulus detection. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics, *SEM@COLING 2020, Barcelona, Spain (Online), December 12-13, 2020*, pages 58–70. Association for Computational Linguistics.

Erik Cambria, Jie Fu, Federica Bisio, and Soujanya Poria. 2015. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 508–514. AAAI Press.

Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1795–1802. AAAI Press.

Erik Cambria, Yangqiu Song, Haixun Wang, and Newton Howard. 2014. Semantic multidimensional scaling for open-domain sentiment analysis. *IEEE Intell. Syst.*, 29(2):44–51.

VA Canady. 2018. Mental illness will cost the world $16 trillion (usd) by 2030. *Ment. Health Wkly*, 28:7–8.

Gian Vittorio Caprara and Daniel Cervone. 2000. *Personality: Determinants, dynamics, and potentials*. Cambridge University Press.

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

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Florian Eyben, Martin Wöllmer, and Björn W. Schuller. 2010. 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*, pages 1459–1462. ACM.

Amelia Gangemi, Francesco Mancini, and Philip N Johnson-Laird. 2013. Emotion, reasoning, and psychopathology. *Emotion and reasoning*, pages 44–83.

Soumitra Ghosh, Asif Ekbal, and Pushpak Bhattacharyya. 2022. A multitask framework to detect depression, sentiment and multi-label emotion from suicide notes. *Cogn. Comput.*, 14(1):110–129.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. 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*, pages 6546–6555. Computer Vision Foundation / IEEE Computer Society.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. 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)*, pages 2122–2132. Association for Computational Linguistics.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 7482–7491. Computer Vision Foundation / IEEE Computer Society.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lisa M Koonin, Brooke Hoots, Clarisse A Tsang, Zanie Leroy, Kevin Farris, Brandon Jolly, Peter Antall, Bridget McCabe, Cynthia BR Zelis, Ian Tong, et al. 2020. Trends in the use of telehealth during the emergence of the covid-19 pandemic—united states, january–march 2020. *Morbidity and Mortality Weekly Report*, 69(43):1595.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. 2021. Fnet: Mixing tokens with fourier transforms. *arXiv preprint arXiv:2105.03824*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825. AAAI Press.

Frances Meeten and Graham CL Davey. 2011. Mood-as-input hypothesis and perseverative psychopathologies. *Clinical Psychology Review*, 31(8):1259–1275.

Saif M. Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 1–17. Association for Computational Linguistics.

Maja Pantic, Nicu Sebe, Jeffrey F. Cohn, and Thomas S. Huang. 2005. Affective multimodal human-computer interaction. In *Proceedings of the 13th ACM International Conference on Multimedia, Singapore, November 6-11, 2005*, pages 669–676. ACM.

Yifan Peng, Qingyu Chen, and Zhiyong Lu. 2020. 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*, pages 205–214. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Trang Pham, Truyen Tran, Dinh Q. Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Advances in Knowledge Discovery and Data Mining - 20th Pacific-Asia Conference, PAKDD 2016, Auckland, New Zealand, April 19-22, 2016, Proceedings, Part II*, volume 9652 of *Lecture Notes in Computer Science*, pages 30–41. Springer.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Robert L Spitzer, Jacob Cohen, Joseph L Fleiss, and Jean Endicott. 1967. Quantification of agreement in psychiatric diagnosis: A new approach. *Archives of General Psychiatry*, 17(1):83–87.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. 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*, pages 5998–6008.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Qili Zhu. 2012. Probase: a probabilistic taxonomy for text understanding. In *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2012, Scottsdale, AZ, USA, May 20-24, 2012*, pages 481–492. ACM.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 12360–12371.

A Appendix

A.1 Baselines

The following baseline methods are considered for the comprehensive evaluation of our proposed.

- bc-LSTM (Poria et al., 2017): For multimodal sentiment analysis, it is an LSTM-based network that extracts contextual information from video utterances.
- CMN (Hazarika et al., 2018): Using a gated recurrent unit on multimodal characteristics, this approach converts prior utterances of each speaker into memories. To capture inter-speaker dependencies, these memories are then combined using attention-based hops. Contextuality is discovered by combining memories using an attention-based technique.
- DialogueRNN (Majumder et al., 2019): This approach uses speaker, context, and emotion information from neighbouring utterances to represent the emotion of words in a dialogue. These variables are represented by three distinct GRU networks to keep track of the various speaker states.
- MT-BERT (Peng et al., 2020): We implement a multitask (MT) variant of BERT based on the architecture proposed by Peng *et al.* (Peng et al., 2020) for our two tasks detection of emotion and Emotional Reasoning.
- Cascaded Multitask System with External Knowledge Infusion (CMSEKI) (Ghosh et al., 2022): The CMSEKI system was introduced in the work that presented CEASE-v2.0 dataset, addressing the detection of depression, sentiment, and emotion, using commonsense knowledge. We adapted the CMSEKI system to address our *Emotion* and *ER* detection tasks.

Parameters	***COMMA-DEER***
Transformer Encoder	2 layers
Embeddings	1068
FC Layer	Dropout (Srivastava et al., 2014) = 0.3
Activations	*ReLU* for dense layers
Output	Sigmoid for Emotional Reasoning
Activations	Softmax for Emotion
Optimizer	Adam (Kingma and Ba, 2015) (lr = 0.003)
Batch	32
Epochs	30

Table 6: Hyper-parameters for our experiments.

A.2 Experimental Setting

We use PyTorch[11], a Python-based deep learning package, to develop our proposed model. We experiment with the base version of BERT imported from the huggingface transformers[12] package. We perform *grid search* to find the optimal hyper-parameters in Table 6. 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. To account for the non-determinism of TensorFlow GPU operations, we present F1 scores averaged across five 5-fold cross-validation runs. We set the sequence length as 128 and report the results with context length = 5, as we observed best scores for this setup.

A.3 External Knowledge Sources

IsaCore: IsaCore (Cambria et al., 2014) is a vector space that preserves semantic and sentiment polarity based on the relationships between instances ('birthday party' and 'china') and concepts ('special occasion' and 'country') and affective labels. It is generated by using multidimensional scaling

[11] https://pytorch.org/

[12] https://huggingface.co/docs/transformers/index

AffectNet	IsA-pet	IsA-food	Arises-joy	...
dog	0.981	0	0.789	...
cupcake	0	0.922	0.910	...
songbird	0.672	0	0.862	...
gift	0	0	0.899	...
sandwich	0	0.853	0.768	...
rotten fish	0	0.459	0	...
lottery	0	0	0.991	...
...	...	...	...	...

Table 7: A snapshot of the AffectNet matrix.

to the knowledge base that results from combining Probase (Wu et al., 2012) (the biggest current taxonomy of common knowledge) with ConceptNet (Speer et al., 2017) (natural language-based semantic network of commonsense knowledge).

AffectiveSpace 2: AffectiveSpace 2 (Cambria et al., 2015) is an unique vector space model for concept-level sentiment analysis that enables reasoning by comparison on natural language ideas even when they are represented by highly dimensional semantic characteristics. This embedding space was generated by performing principal component analysis (PCA) on the AffectNet matrix representation, which is a semantic network in which common-sense concepts are linked to semantic and affective properties. A snapshot of the AffectNet matrix is shown in Table 7.