ENHANCING EMOTION RECOGNITION IN TEXTUAL CONVERSATION BY LEVERAGING COMMON-SENSE

A Thesis

by

RIZU JAIN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Ruihong Huang
Co-Chair of Committee,	Theodora Chaspari
Committee Member,	Vinayak Krishnamurthy
Head of Department,	Scott Schaefer

May 2021

Major Subject: Computer Science

ABSTRACT

A core aspect of human-like artificial intelligence is the awareness of emotions. Recognizing emotions in conversations (ERC) is especially difficult to solve because of several challenges like contextual modeling, interlocutor profiling, recognizing emotion shifts, and multiparty conversations. Analyzing the results from the state-of-the-art techniques reveal that ERC models could be improved by using common-sense knowledge. Once incorporated correctly, the advancements in general common-sense knowledge models could be leveraged directly in ERC models. Furthermore, insights from these experiments will be applicable in other language tasks as well.

In this work, I propose two approaches for incorporating common-sense knowledge in ERC models: first, implicit incorporation by fine-tuning language models using the common-sense inferences for the given data, and second, explicit incorporation by applying cross-attention on common-sense knowledge for each utterance. I demonstrate that the proposed methods perform well on IEMOCAP, a widely used dyadic conversation dataset with human-annotated emotions.

DEDICATION

To my mother, father, and brother for being my constant source of happiness.

ACKNOWLEDGMENTS

I would like to express my earnest appreciation and gratitude to Dr. Ruihong Huang for her guidance and mentorship in the research work outlined in this thesis. I want to thank my fellow researchers of the NLP lab at Texas A&M University for their inspiration and suggestions. I extend special thanks to my committee for reviewing this thesis document and helping me to improve it. I would also like to acknowledge the contribution of my colleague, Shubham Sanghavi, who supported me with certain facets of experimentation.

I also thank my friends and family who were always at the giving end of encouragements for all my endeavours.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a thesis committee consisting of Dr. Ruihong Huang and Dr. Theodora Chaspari of the Department of Computer Science and Engineering and Dr. Vinayak Krishnamurthy of the Department of Mechanical Engineering.

The text corpus used in the thesis was published by Speech Analysis and Interpretation Laboratory, USC. The baseline ERC model was developed and released by Deep Cognition and Language Research (DeCLaRe) Lab, SUTD. All other work conducted for the thesis was completed by the student independently.

Funding Sources

Graduate study was supported by a Graduate Teaching Assistantship from the Department of Computer Science and Engineering at Texas A&M University.

NOMENCLATURE

BERT	Bidirectional Encoder Representations from Transformers
BPE	Byte Pair Encoding
CNN	Convolutional Neural Networks
CSK	Common-sense knowledge
ERC	Emotion Recognition in Conversations
GCN	Graph convolutional networks
GPT	Generative Pre-Training
GRU	Gated recurrent units
IEMOCAP	Interactive Emotional Dyadic Motion Capture database
LSTM	Long short-term memory
MLM	Masked Language Modelling
NLP	Natural Language Processing
NSP	Next Sentence Prediction
RNN	Recurrent Neural Networks
SUTD	Singapore University of Technology and Design
USC	University of Southern California

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

Emotion is fundamental to humans and therefore, the awareness of emotions in human conversations is an important aspect of human-like artificial intelligence. Conversational data is now being increasingly available due to the advent of social platforms. Utilizing this data for emotion recognition in conversations (ERC) can help analyze the fine-grained sentiments in multifarious scenarios. With more affective dialogue systems, applications like online health programs, court trials, interviews, social networking websites, recommendation systems, etc. shall all benefit as well.

Emotion recognition in computational linguistics is the method of recognizing distinct emotions reflected in the text. This discipline has not yet achieved the popularity and pervasiveness of the broader field of sentiment analysis - and a way to tackle this problem would be to address its research challenges as outlined in further sections.

Current state-of-the-art approaches use contextual modelling and feature representations from language models to achieve high accuracy on the ERC task. Additional information about the emotion of an utterance can be obtained from automatic common-sense knowledge (CSK) generators. Like for the utterance 8 in Fig. 1.1, common-sense knowledge would suggest that Chandler is regretful and help in inferring his emotion to be sad. However, the current state-of-the-art models seem to be lacking common-sense reasoning in some cases. In this work, I explore two approaches for leveraging automatically generated common-sense knowledge for recognizing emotion in textual conversation. The first approach is implicit incorporation by fine-tuning language models using the common-sense knowledge for the given utterance, and the second approach is explicit incorporation by applying cross-attention on common-sense knowledge representation vector for each utterance. The results for different combinations of these approaches are compiled and analyzed to draw material insights for incorporating common-sense knowledge. The implicit method of CSK incorporation improves performance over the baseline models and is better at predicting emotion shifts, a key challenge in the ERC task.

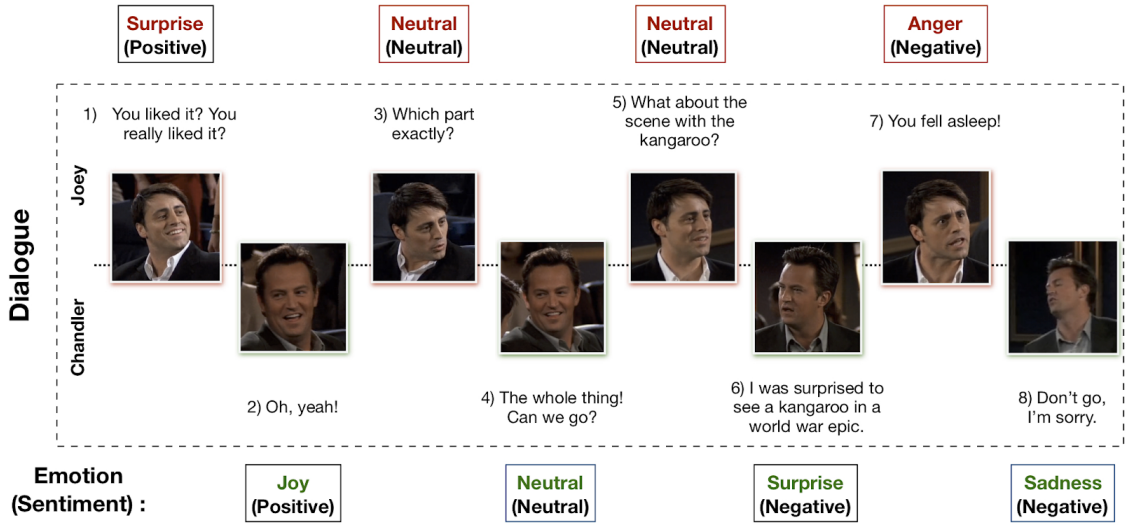

Figure 1.1: Reprinted from [1]. An excerpt of a conversation with each utterance tagged with corresponding emotion and sentiment label (which can be either positive, negative, or neutral sentiment). This is a simple example of a conversation between two parties. The task holds for multi-party conversations as well [1, 2].

This thesis is organized as follows: This chapter first gives an overview of the ERC task, states its problem definition, and details about the research challenges. Chapter 2 reviews the key contributions and prior work in the ERC domain. Chapter 3 elaborates on the proposed framework. Chapter 4 describes the training and testing setup for the analysis of the methodologies. Inferences are drawn in Chapter 5. The thesis concludes by discussing the results, challenges and future scope in Chapter 6.

1.1 Problem Definition

The ERC (Emotion Recognition in Conversation) task aims to learn a function that would take as input the transcript of a conversation and speaker information for each constituent utterance and identify the emotion of each utterance from a set of predefined emotions. An utterance is defined as a unit of speech bound by breathes or pauses in a conversation [5]. Any conversational unit - phrases, clauses or sentences can be an utterance.

Mathematically, the objective is to predict the emotion label e_i of each utterance u_i, provided the input sequence of N utterances $[(u_1, p_1), (u_2, p_2), ..., (u_N, p_N)]$, where each utterance $u_i =$

$[u_{i1}, u_{i2}, ..., u_{iT}]$ consists of T words u_{ij} spoken by party p_i.

Classification of Emotions

Emotions can be represented in two ways: either as discrete types or as a point in multi-dimensional continuous space. In the first kind of representation, there are two widely known emotion category models: Plutchik's wheel of emotions defining eight emotion types [6] and Ekman's model of six basic emotions [7]. Emotions are mapped to one of the discrete emotion labels present in the model. In the second representation, emotions are a measure of the two or three dimensions which can be levels of arousal, pleasure, relaxation etc [8]. Although the dimensional approach can map an emotion more precisely using vectors, they are hard to be annotated.

In this thesis, we use IEMOCAP [9] dataset which categorize utterances into six emotion labels: angry, frustrated, happy, sad, excited, and neutral. However, the research works referenced may use a different set of emotions.

1.2 Research Challenges

ERC is an arduous task to address because of its many research challenges. The following sub-sections describe the challenges as well as the factors governing the emotion of an utterance. These factors inherently also drive the conversation.

1.2.1 Annotation of emotion labels in datasets

While building a corpus, labeling of emotions depends on the annotator's perspective. This adds a complication in annotation. Real-time labeling of unscripted interactions is potentially impractical because that would hinder the communication flow. To resolve this, multiple annotators are involved to label an emotion. Hence, improving inter-annotator agreement is also an ongoing research problem [10].

1.2.2 Context Modelling

Initial research treated individual utterances independently while predicting the emotion of that utterance. Later studies showed that the structure of the dialogue, the neighboring sentences,

and previous utterances play a major role in determining the emotion conveyed in a particular utterance. The preceding utterances at time $< t$ along with its temporal sequence can be viewed as a context of the utterance at time t. The computation and representation of this context experience considerable difficulties due to the emotional dynamics involved. Research studies have made it apparent that the information derived from the textual representations of the context of an utterance has helped ERC models to elevate their prediction accuracies. It has also helped in generalizing the overall model of emotion recognition [11].

1.2.3 Interlocutor Profiling

People have their own subtle ways of communicating their sentiments. Depending on the character of the speaker, the personalities, motives, perspectives of the interlocutors, behavior towards each other, etc., a phrase or an utterance can carry different emotional strength and polarity. This highlights the need to do user profiling for better results because this shall provide the appropriate prior knowledge about the interlocutor [3, 12, 13].

1.2.4 Emotion Shift

Emotions can fluctuate with each dialogue and misidentifying such labels is found to be a significant limitation causing reduced accuracies of ERC models. Fig 1.1 shows the shifts in emotions of Joey and Chandler throughout their conversation. It is observed that current ERC models are unable to discern spontaneous emotional changes and thus are not capable to comprehend the subtle distinctions between classes of emotions that are closely linked to each other [14, 15]. Examples of such emotions include anger v/s frustration, happy v/s excited, etc.

1.2.5 Reasoning

In conversational history, emotion reasoning not only discovers the contextual utterances that activate that emotion but also decides the role of such contextual utterances on the target utterance. An opinion-holder may also have an emotional bias against the entity/topic in question. For example, in Fig. 1.1, the knowledge of Joey being angry comes from him discerning that Chandler deceived him which is not evident from the utterances. The lack of rich annotated datasets con-

taining the cause of emotion makes it very difficult to frame meaningful structures of conversation that would have made the AI dialogue systems more empathetic.

1.2.6 Multi-modal Analysis

Our day-to-day conversations are not always explicit. This ambiguity is dealt with by facial gestures and/or the tone of the dialogue. Hence in some cases, multi-modal data becomes essential for drawing the emotion predictions. Using only textual data for such conversation can result in false classifications of emotions. User-generated videos are now available through online platforms aiding in the development of multi-modal datasets. Fusion techniques of these multi-modal features are being actively researched. Some studies have proved enhancement of emotion recognition employing the multi-modal conversation as input features compared to baseline text-based models [15, 16, 17].

2. RELATED WORKS

For several years, emotion recognition has become an active research area and has been studied in interdisciplinary areas. Recently, a growing number of models for solving ERC using various deep learning structures have been proposed. Traditionally, convolutional neural network (CNNs) architectures [18] were used for utterance level emotion detection but they failed to model the contextual information. The recent trend in the identification of conversational emotions is to conduct context modeling using deep-learning-based algorithms in either textual or multi-modal environments. The advent of recurrent neural networks (RNNs) made it seem easier to model context due to the sequential nature but has been unsuccessful in the literature [11] owing to its poor performance to capture long distance contextual information. However, memory networks, RNNs, and attention mechanisms in a hybrid architecture have been used to comprehend contextual knowledge [3, 13].

CMN [19] used memory networks in dyadic conversations for emotion recognition, where two different memory networks enabled inter-speaker interaction. Further, to model emotional dynamics, DialogueRNN [3] proposed a GRU-based model focussing on party state and global state, employing an attention mechanism to extract information for each target utterance. A recent work by Sheng et al. [20] has focused on phrase-level semantic relations to understand emotion changes. Combating the issue of long sequences, GCN [21] as well as transformer [22] architectures became popular as they were able to exploit contextualized utterance representations.

In KET [23], its ERC model uses an external knowledge base to enrich an utterance with its related entities which aides in determine the emotion associated with the utterance. Such incorporation of knowledge is being in done in other related NLP tasks like dialogue systems [24], sentiment analysis [25]. Thus it can be said that humans also rely on some common-sense knowledge to convey feelings. Awareness of common-sense is important for interpreting discussions and producing effective answers [12]. It has been seen in many studies that transformers can effectively integrate contextual data as well as external information bases in the ERC models. Competitive

performance against advanced state-of-the-art techniques has been demonstrated by simple baselines using BERT [26, 27]. Several works have investigated strengthening the structure of transfer learning through model pretraining by either changing the learning method or adapting weight in the ERC downstream task [12, 28].

2.1 Experiments and Analysis

In our experiments, DialogueRNN with utterance embeddings extracted from RoBERTa [29] fine-tuned for the emotion recognition task have shown promising results. However, in some cases the model lacks common sense reasoning. One such example is shown in Table 2.1. This indicates that there is scope for improving performance by explicitly incorporating Common-Sense Knowledge (CSK). This is attempted in COSMIC [12].

Speaker	Utterance	True Emotion	Predicted Emotion
Person A	Hi, um- I think my baggage was lost, and I need to I guess file a claim or I don't know what.	Neutral	Neutral
Person B	Okay, what flight were you on?	Neutral	Neutral
Person A	Um-Seventeen. Coming from --	Neutral	Neutral
Person B	On what airline?	Neutral	Neutral
Person A	It was Virgin Atlantic coming from London.	Neutral	Neutral
Person B	From London, okay. Um- And you went through all your proper checkpoints?	Neutral	Neutral
Person A	Yeah, I don't know where it was lost. I was coming from Africa, so.	**Frustrated**	**Neutral**

Table 2.1: Misclassifications by DialogueRNN with textual features from RoBERTa.

COSMIC uses COMET [4], an automatic knowledge graph generation language model, to extract CSK from the speaker's utterance. For a given subject, COMET can generate objects for nine different relations. Figure 2.1 shows COMET's response for three relations when given the

prompt, 'My luggage was lost'.

COSMIC incorporated this knowledge by maintaining the speaker's internal states as per CSK. They intend to model the speaker's intent, their reaction, and how they are perceived by others using GRUs and update them using CSK relations. All these relations are present in COMET and are taken as input in the COSMIC model. The state vectors from these GRUs are then concatenated to update the emotional state of individual parties in the conversation. Finally, a fully connected layer is used to determine the emotion of the party.

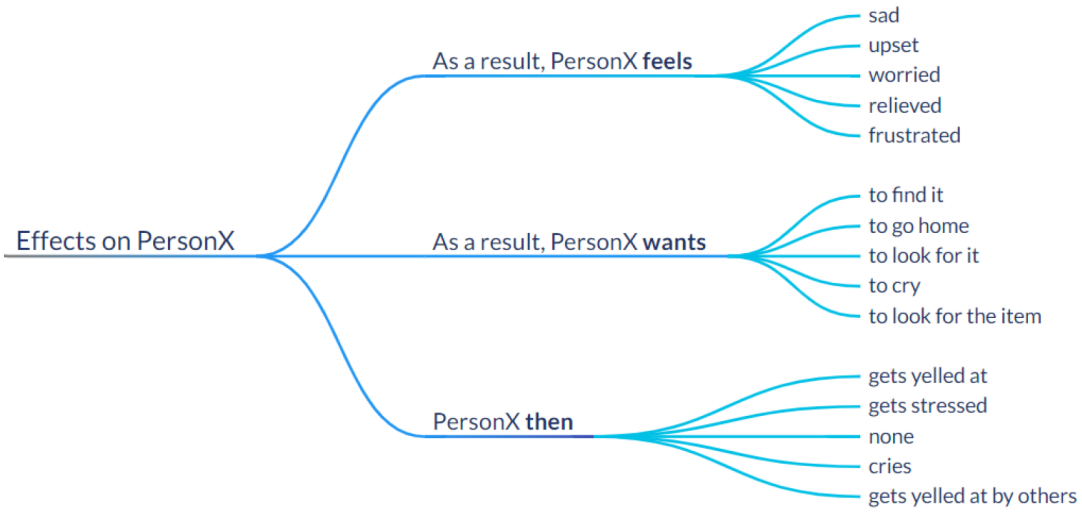

Figure 2.1: Reprinted from demo web page of Mosaic Knowledge Graphs built by Allen Institute for AI. The diagram shows the COMET Output for three relations for 'My luggage was lost' subject. Web Link: Mosaic KG

.

The results from COSMIC show a marginal improvement over the DialogueRNN with features extracted from RoBERTa (shown in Table 2.2). Only a marginal improvement is observed over the DialogueRNN results. Moreover, the accuracy of identifying emotion shifts has degraded. Possible justifications for COSMIC's underwhelming performance could be noisy common-sense attributes or incorrect method of incorporating them. In this work, I study and explore ways to incorporate common-sense knowledge differently and effectively.

Model	Accuracy	Accuracy on Emotion Shifts
$DRNN_{Glove}$	63.40	47.50
$DRNN_{RoBERTa}$	64.66	51.14
$COSMIC$	65.11	50.63

Table 2.2: ERC results comparison between DialogueRNN_RoBERTa and COSMIC.

Common-sense knowledge incorporation is attempted by Chang et al. (2020) [30] for SocialIQA, a social commonsense reasoning task. They present two approaches for incorporating CSK into their model. In the first approach, they fine-tune RoBERTa with the tuples acquired from large common-sense knowledge graphs like ATOMIC [31] and ConceptNet [32]. In the second approach, they attend over different common sense tuples acquired from the graphs to incorporate common sense knowledge selectively. Their experiments show promising results and improvement upon the corresponding baseline models.

In my proposed models, I adapt the approaches used by Chang et al. for ERC to incorporate common sense knowledge more effectively. For implicit incorporation we fine-tune language models using the common-sense knowledge for the given utterance. For explicit incorporation, we apply cross-attention on common-sense knowledge representation vector for each utterance. We empirically show that implicit incorporation improves performance over the baseline models and is better at predicting emotion shifts, a key challenge in ERC task.

3. METHODOLOGY

In conversational emotion recognition, the task is to classify each of the constituting utterances into their appropriate emotion category. Like most modern NLP tasks, ERC involves two main stages. First stage is generating context-independent representation vector for the utterance. Being context-independent indicates that it doesn't utilize any other utterances from the dialogue. The second stage is contextual modelling. Here, we use recurrent neural models that uses information from the previous utterances to make use of the context in predicting emotions.

Our framework proposes changes in both these stages for incorporating common-sense knowledge (CSK) in ERC. First, we propose implicitly incorporating CSK by directly embedding it in the vector representations of the utterances. This will help us use natural language CSK thereby making it extendable to enhancements or changes in external CSK. Furthermore, it provides more visibility in the CSK being incorporated. Also, we can filter noisy CSK before incorporation through pertinent NLP models.

Incorporating CSK explicitly in the model is the second approach. In this, we use the vector representation of CSK as a feature in the model. The intention is to use a particular relation in CSK that give the effect on speaker directly in the model. This also allows us to use relevant CSK from previous utterances in determining the emotion through attention networks. We believe that given the nature of the conversations, the effect from the central event shows up after a few utterances. Finally, we want to check how well do the two methods compliment each other.

3.1 Implicit Incorporation of CSK

In this approach, we attempt to incorporate CSK in the context-independent vector representations of utterances. For extracting CSK, this work uses COMET [4], an automatic common-sense generation model trained on ATOMIC [31]. ATOMIC is an collection of common-sense if-then descriptions in the form of a knowledge graph. It contains 877K tuples covering a variety of social CSK around specific event prompts (e.g., "X brings gifts"). Each event prompt has nine CSK

dimensions covering the event's causes (e.g., "X needs to purchase them"), its effects on the agent (e.g., "X feels good about themselves") and its effect on other direct (or implied) participants (e.g., "Others feel grateful"). The dimensions in ATOMIC are detailed in Table 3.1.

Event	**Description**	**Example Completion:**
		Person X puts Person X's trust in Person Y
`oEffect`	The effect the event has on others besides Person X	is considered trustworthy is believed gains Person X's loyalty
`oReact`	The reaction of others besides Person X to the event	trusted honored trustworthy
`oWant`	What others besides Person X may want to do after the event	work with Person X partner with Person X to help Person X
`xAttr`	How Person X might be described given their part in the event	faithful hopeful trusting
`xEffect`	The effect that the event would have on Person X	gets relieved stays faithful Is betrayed
`xIntent`	The reason why X would cause the event	to be trusting his or her help/guidance/advice to be friends
`xNeed`	What Person X might need to do before the event	to be friends with Person Y to have heard a lot of good things about Person Y to get to know Person Y
`xReact`	The reaction that Person X would have to the event	trusting safe, not alone understood
`xWant`	What Person X may want to do after the event	to rely on Person Y to go into business with Person Y to make sure that their heart feeling is right

Table 3.1: Reprinted from [4]. Definition of the relations in ATOMIC. Events in ATOMIC center around the personal situations of a central figure, Person X, with potentially more participants.

For training COMET, the event prompt from ATOMIC tuple is taken as the subject s , the dimensions are treated as relations r and the causes/effects are the objects o. COMET is trained to generate the object phrase o from concatenated subject phrase s and relation phrase r for a triplet $\{s, r, o\}$. Sample output from COMET was presented in Figure 2.1.

Each utterance is fed as a subject and objects for four relations, `xIntent`, `xAttr`, `xNeed`, and `xWant`, are extracted. `xIntent` and `xNeed` are the causes for the subject and could also be the causes behind speaker's emotion. `xAttr` is the speaker's attributes based on the utterance which could help in implicitly modelling the personality of the speaker. The intention behind including `xWant` is to allow capturing turns in emotion, for e.g. regret. The objects extracted from COMET for the utterances are then appended to pre-defined templates for each relation to form meaningful sentences (See Table 3.2 for templates with example sentences). When COMET cannot find a suitable object for the subject relation pair, it returns *none* as the object and that CSK tuple is filtered out.

Relation	**Template**	**Example**
`xIntent`	*I am trying ...*	*I am trying* to be trusting
`xNeed`	*I needed ...*	*I needed* to be friends with Person Y
`xAttr`	*I am ...*	*I am* trying to be trusting
`xWant`	*I want ...*	*I want* to rely on Person Y

Table 3.2: Sentence templates for the relations with examples. Examples uses objects for prompt 'Person X puts Person X's trust in Person Y'.

Finally, the generated CSK sentences are appended with the utterance to form an extended utterance. Few extended utterances are shown in Table 3.3.

RoBERTa model [29] is used to extract context independent utterance level feature vectors. RoBERTa is based on the BERT model [26] which is a deep bidirectional transformer intended to pre-train over huge unlabelled text corpus to learn language representations. BERT uses novel pre-

Utterance	Extended Utterance
I will do no such thing.	I will do no such thing. *I am trying to be a good person. I am stubborn. I want to be left alone.*
Is that your phone?	Is that your phone? *I am curious. I want to call someone.*
Do you want my jacket?	Do you want my jacket? *I am trying to be warm. I am cold. I want to go outside.*
Awesome. We're both going to be in L.A.	Awesome. We're both going to be in L.A. *I am trying to be in a different place. I am adventurous. I needed to buy a plane ticket. I want to go to the beach.*

Table 3.3: Sentence templates for the relations with examples. Examples uses objects for prompt 'Person X puts Person X's trust in Person Y'.

training tasks called Masked Language Modelling (MLM) and Next Sentence Prediction (NSP). RoBERTa Large follows the original BERT Large architecture having 24 layers, 16 self-attention heads in each block and a hidden dimension of 1024, resulting in a total of 355M parameters. RoBERTa improves upon BERT by removing NSP objective and training for more epochs with much larger mini-batches and learning rates.

RoBERTa model was pre-trained on five large unlabelled language datasets for the MLM task: BookCorpus, a dataset of 11K unpublished books, English Wikipedia , CC-News, a dataset of 63 millions crawled news articles, OpenWebText, an opensource WebText dataset, and Stories, a subset of CommonCrawl filtered for story-like texts. The combined training corpus size of these datasets turns out to 160 GB. MLM pre-training task converts the text into tokens and uses the token representation as an input and output for the training. A random subset of the tokens (15%) are masked, i.e. hidden during the training, and the objective function is to predict the correct identities of the tokens.

For fine-tuning RoBERTa on the ERC task, two approaches are proposed: first, fine-tuning for MLM task using the utterances and second, fine-tuning for context independent emotion recognition task. In the first approach, the CSK appended extended utterances are used as indepen-

dent textual documents to fine-tune RoBERTa Large pre-trained on the 160 GB language corpus (See Figure 3.1). Extended utterances are generated from the training split of ERC datasets using COMET. A subset of tokens from the extended utterances are masked and RoBERTa is trained to predict those masked labels. The language model is expected to implicitly learn common-sense knowledge specific to our task through this approach.

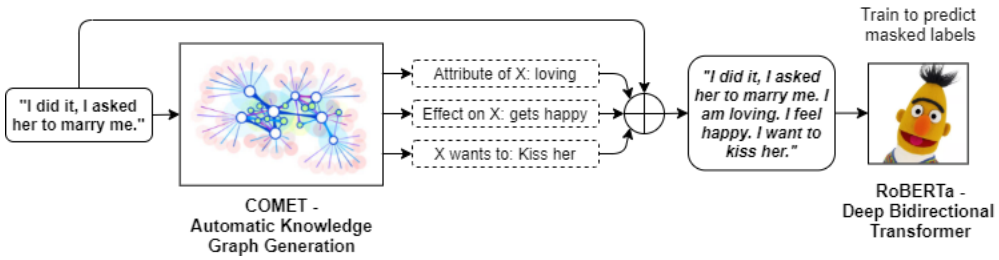

Figure 3.1: Implicitly incorporating CSK for utterances by fine-tuning RoBERTa using extended utterances generated by appending CSK retreived from COMET to the utterances.

The second approach is a more generalized approach for fine-tuning RoBERTa on any supervised classification task. In this, RoBERTa is fine-tuned by training to predict the emotion of a given extended utterance. Let an extended utterance x consists of a sequence of tokens $x1, x2, ..., xN$, with emotion label E_x. A special token $[CLS]$ is appended at the beginning of the extended utterance to create the input sequence for the model: $[CLS], x1, x2, ..., xN$. This sequence is passed through the model. The $[CLS]$ token activation from the last layer are then used in a feed-forward network to classify it into its emotion class E_x. The errors are back-propagated through RoBERTa to update the weights. I experiment with the pre-trained weights from the large language corpus and the MLM fine-tuned RoBERTa as the initial RoBERTa weights for the emotion classification fine-tuning.

To extract the feature representations of the utterances, $[CLS]$ is appended to the utterance and passed to fine-tuned RoBERTa model. Activations from the final four layers for the $[CLS]$ token are averaged out to get a 1024 dimensional feature vector for the utterance.

3.2 Explicit Incorporation of CSK

In the second approach, CSK is used in the ERC model explicitly. A state vector $k_{X,t}$ is introduced to model the common-sense effect on a person X at a particular utterance t. It is obtained by attending over all the previous common-sense effect vectors using the current context vector. Finally, it is used to update the party state in DialogueRNN model.

COMET is a generative model and it produces a discrete sequence of tokens conditioned on the subject and relation phrase. But for explictily using CSK, vector representations are required. For that, we take the pre-trained COMET model on ATOMIC knowledge graph and discard the phrase generating decoder module. The utterance t is treated as the subject and is concatenated with relations `xReact` and `oReact` separately and passed through the COMET encoder. Activations from the last stage of the encoder are extracted for each relation. This gives two different vectors, one for the effect on self for the utterance obtained using `xReact` and one for the effect on others for the utterance obtained using `oReact`. These vectors have 768 dimensional.

Let us introduce nomenclature for our CSK effect vectors. Let the two parties in the conversation be, *Person A* and *Person B* and let $f_{X,t}$ be the common-sense effect on person X by the utterance t. If person X is the speaker of utterance t, $f_{X,t}$ is the effect on self (`xReact`) vector retrieved from COMET and if the person X is the listner, $f_{X,t}$ is the effect on others (`oReact`) vector. An example is shown in the figure. Here, *Person A* is the speaker which implies *Person B* is the listner as we are dealing with just dyadic conversations. Thus, effect on self is *Person A's* CSK effect vector $f_{A,t}$ and effect on others is *Person B's* CSK effect vector $f_{B,t}$ (See Figure 3.2).

Person X's common-sense effect vectors for from utterance 1 to $t-1$, $(f_{X,1}, f_{X,2}, ..., f_{X,t-1})$, are concatenated to generate person X's common-sense effect matrix, F_X. We attend over F_X using the current context vector from DialogueRNN model. This gives higher weightage to the effect vectors that are relevant to the current context. The rationale behind introducing an attention network is that the effect of certain utterances show up a bit later in the conversation. Also, strong effects alter the emotional state throughout the remaining conversation. To capture these, we utilize the history of effect vectors for person X in determining their emotional state as per CSK. The

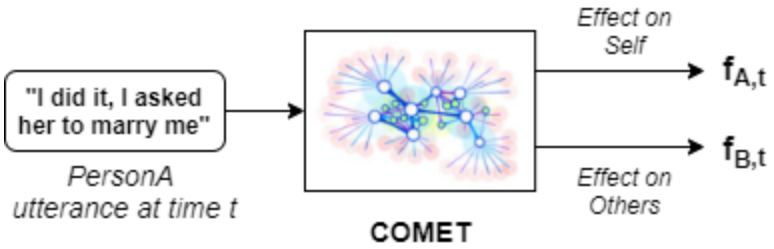

Figure 3.2: Extracting common-sense effect feature vectors for utterance t from COMET. These feature vectors are used to explicitly incorporate CSK in the model.

attended common-sense vector for person X is denoted as $k_{X,t}$ as shown in Figure 3.3.

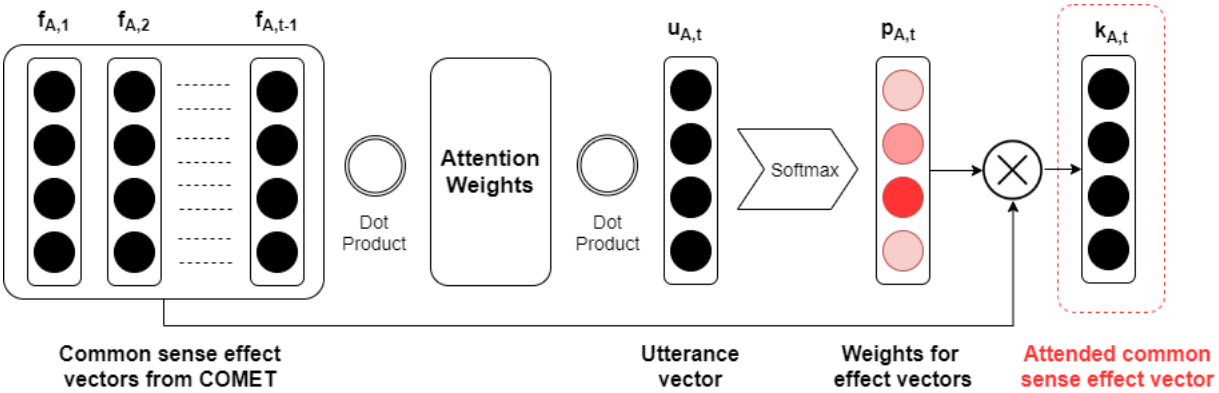

Figure 3.3: Attention network used to generate context relevant common-sense effect vector $k_{X,t}$.

DialogueRNN is used as the baseline model for our experiments. DialogueRNN is designed based on the assumption that there are three major factors affecting the emotion of an utterance: the speaker, the context of the conversation and the emotions of preceding utterances. DialogueRNN attempts to model these factor using GRUs. Figure 3.4 shows DialogueRNN architecture with the additional vectors added in our approach shown as red arrows.

- GRU_P is the party state GRU intended to track the emotional dynamics of individual parties' involved in the conversation. Each party/person has its individual party state GRU and it is updated when that person utters. GRU_P is updated using the utterance representation u_t and

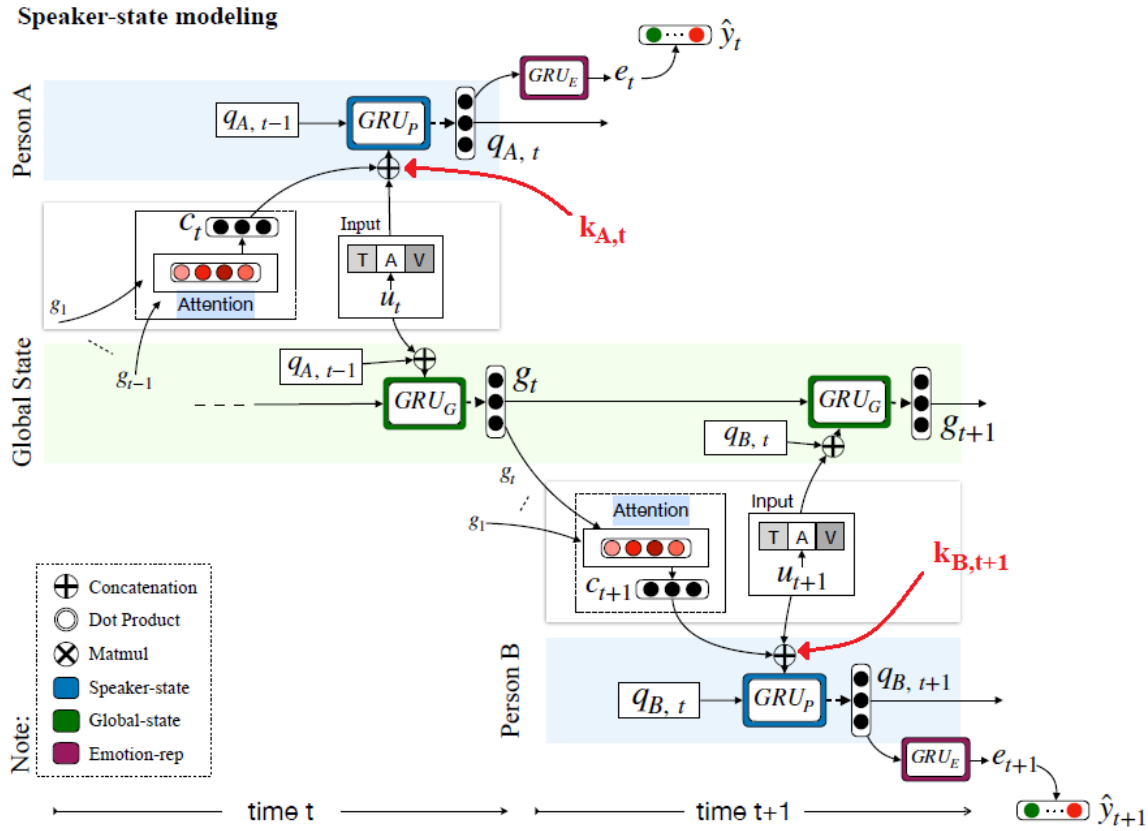

Figure 3.4: Adapted from [3]. Modified DialogueRNN architecture for a dyadic conversation between Person A and Person B. Note that although represented separately, the emotional state GRU_E is shared between the parties. Contextual CSK knowledge vectors $k_{A,t}$ and $k_{B,t}$ are added for explicit incorporation of CSK.

the contextual vector c_t. Attention is applied over all the preceding global states using the current utterance u_t to generate context vector c_t. This give higher weightage to the global states that are more relevant to the current utterance.

- GRU_G is the global state GRU which captures the conversation context. It is shared amongst all the parties and updated on each utterance. GRU_G uses the current utterance u_t and the speaker's state $q_{X,t}$ to update the global state g_t. As it is updated using the states of all the parties involved, it captures inter-speaker dependencies to produce improved contextual representations.

- GRU_E is the emotional state GRU which tracks the emotional state e_t of the conversation. This models the final important factor of the emotion of an utterance which is the emotions of preceding utterances. A single GRU_E is used to track the emotions for all the parties. The emotional state e_t is inferred from the from the speaker's state $q_{X,t}$ and the previous emotional state e_{t-1}. As the previous emotional state was inferred from the previous speaker's state, the inter-party emotional dynamics are captured through GRU_E.

Finally, a two-layer feed-forward network with a soft-max output layer is used to calculate emotion-class probabilities from emotion representation e_t. Then the utterance u_t is assigned the emotion class with the highest probability.

The specific variant of DialogueRNN that performs the best and is used in our experiments is the bi-directional DialogueRNN with emotional attention. In Bidirectional DialogueRNN two different DialogueRNN cells are used for forward and backward pass of the input sequence. For the backward DialogueRNN, the utterance sequence of the dialogue is reversed. The output emotion representations e_t from both the forward and backward cells are concatenated at the utterance level before passing into the feed-forward network for emotion classification. This concatenated emotion representation contains information from both the past as well as future utterances in the dialogue, similar to a bidirectional RNN. Another addition in this variant is the emotional attention. Instead of directly using the emotion representation e_t, attention is applied over all the emotion representations in the dialogue. Experiment results show significant improvement in final emotion prediction on using emotional attention [3].

To explicitly incorporate CSK in DialogueRNN, the attended common-sense vector $k_{X,t}$ will be used to update the party state GRU. It is appended with the context vector c_t and the utterance u_t to update GRU_P. For a conversation between Person A and Person B, two common-sense vectors $k_{A,t}$ and $k_{B,t}$ respectively are added to DialogueRNN architecture. These are in-turn generated by attending over common-sense matrices F_A and F_B using the context vector c_t. The motivation behind incorporating common-sense vectors for party state update is that the use of common-sense knowledge about how a person feels at time t would augment the contextual modelling.

4. EXPERIMENTAL SETUP

4.1 Dataset

IEMOCAP, SEMAINE, Emotionlines, MELD, DailyDialog, EmoContext are a few publicly accessible datasets for the task of conversational emotion recognition.

We shall use IEMOCAP corpus for our experiments. Interactive emotional dyadic motion capture database (IEMOCAP) [9] is published by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). It contains videos of two-way conversations of ten unique actors. The actors engage in scripted as well as spontaneous conversations for hypothetical scenarios designed to elicit specific emotions. The dialogues are between two parties only and are separated into utterances. Each utterances is annotated with one of the following emotion labels - happy, sad, neutral, angry, excited, and frustrated by three human annotators. Although this is a multi-modal corpus containing textual, audio and visual features, we will be using only textual modality in our experiments.

IEMOCAP contain 151 dialogues which have a total of 7433 utterances. Train and test split strategy is adopted from DialogueRNN where they do a 80/20 split with no overlapping speaker among the train and test dialogues. The final train set contains 120 dialogues with 5810 utterances and test set contains 30 dialogues with 1539 utterances. The distribution of emotions in the training and test sets are shown in Figure 4.1.

4.2 Baselines

DialogueRNN [21]: This work models the individual speakers separately using GRUs along with a global state GRU and emotional state GRU. Bidirectional variant of the model along with additional attention on emotion is reported to give the best results and is used for comparison. Originally, it used glove vectors to extract textual feature vectors for the utterances. The baseline that we compare against uses textual features extracted from RoBERTa fine-tuned for labelled emotion classification task on the IEMOCAP dataset.

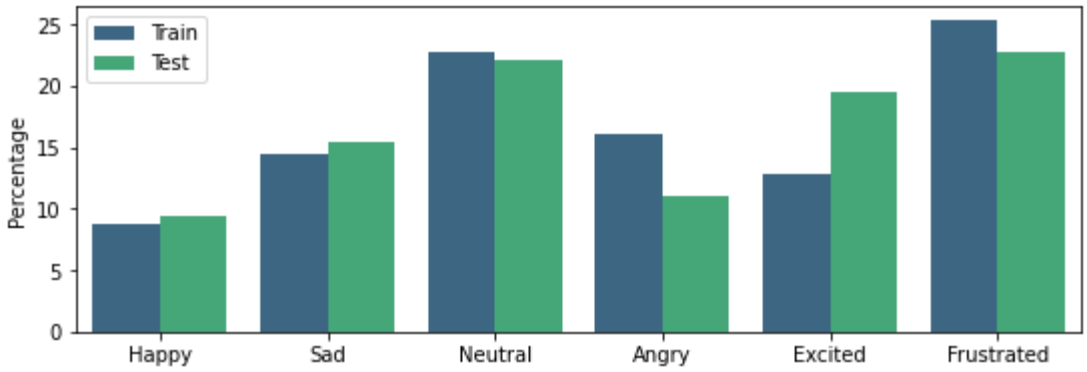

Figure 4.1: Percentage of each emotion in the train and test split of IEMOCAP dataset.

COSMIC [12]: This paper attempts to utilize common-sense knowledge in emotion recognition. The authors of this work intend to model additional states for each party involved in the conversation to capture the common-sense knowledge of their latent emotional state and their emotional state as observed by others. For this, common-sense knowledge from COMET is extracted using specific relations. It gives a marginal improvement above the DialogueRNN with RoBERTa textual features as reported in the paper.

4.3 Feature Extraction

We are just using the textual modality from IEMOCAP dataset. Thus, we need to extract representation vectors for the utterances present in the dataset. Also, for explicit incorporation of CSK in the model, the common-sense feature vectors for each utterance are required. The process followed to extract them is detailed chronologically in this section.

4.3.1 Common-sense knowledge extraction

We implicitly incorporate CSK in the utterance representations during the fine-tuning phase. Thus, first of all we query COMET to extract the necessary CSK relations. We use the released pre-trained COMET weights on the ATOMIC knowledge graph. COMET is queried for `xIntent`, `xAttr`, `xNeed` and `xWant` reactions using a *greedy* sampling algorithm. The returned objects are appended with the sentence templates shown in Table 3.2. Finally, all the generated sentences

are appended to the utterance to form an extended utterance.

4.3.2 Fine-tuning: MLM

We use RoBERTa for textual feature extraction with different fine-tuning setups. Facebook's open-source library, *fairseq* [33], is used to do all the RoBERTa specific tasks. A pre-trained RoBERTa Large released with the *fairseq* is used as the initial weights for all the experiments.

Fine-tuning RoBERTa on MLM task is undertaken to incorporate implicit CSK in the utterance representations. For this, extended utterances for all the utterances along with the appended CSK sentences are dumped in raw text files. The train, valid and test utterances are dumped in separate text files and encoded using the GPT-2 BPE encoder. Finally, they are binarized using the *fairseq* GPT2 dictionary to make them RoBERTa compatible. RoBERTa is trained on the training utterances corpus for masked token prediction task for 12500 updates with a peak learning rate of 5×10^{-4} and a 0.2 dropout rate. Details about all the other hyper-parameters are provided in the Appendix. The evaluation at checkpoints is done using the validation set and the weights at checkpoint with the best accuracy are saved for use in the next step of fine-tuning.

4.3.3 Fine-tuning: Labelled emotion classification

We perform RoBERTa fine-tuning for labelled emotion classification in all the experiments. The difference is in the utterances that we use for the task. In the *base* labelled fine-tuning we use the utterances as it from the IEMOCAP dataset. In the *CSK* labelled fine-tuning, we use the extended utterances generated by appending the CSK sentences from COMET.

For both the cases, the process for fine-tuning is similar. First, we extract out the utterances from IEMOCAP and write them to a new file with every utterance on a single line. A label file is also generated with the true emotions of the utterances on the corresponding lines. This is done for all the train, valid and test splits. RoBERTa implementation by *fairseq* expects data in this format for training on the classification task. Similar to MLM training, the data and labels are BPE encoded using the GPT-2 encoder and binarized using the GPT-2 dictionary. RoBERTa is trained to predict the emotion classes on the training set for 30 epochs through the IEMOCAP dataset with

a peak learning rate of 1×10^{-6} and a 0.2 dropout rate. All the other hyper-parameters are provided in the Appendix. The weights at checkpoint that gives the best validation accuracy are saved for extracting textual features.

4.3.4 Utterance representation extraction

To extract the representation vectors for the utterances, we use the RoBERTa model in evaluation mode. First, RoBERTa is initialized with the desired fine-tuning weights. Utterances from IEMOCAP dialogues are encoded and a $[CLS]$ token is appended at the start of each utterance. We use the same utterances that were use to fine-tune RoBERTa for the labelled emotion classification task. The processed utterances are then passed to the language model. Activations from the last 4 layers for the $[CLS]$ token are extracted and passed as features to be used in the ERC models.

4.3.5 Common-sense representation extraction

COMET is an encoder-decoder model and returns objects in the form of texts. To explicitly incorporate common-sense effect knowledge, we need representation vectors for them. To get that, we ignore the decoder head of COMET and extract the internal activations as common-sense representation vectors. For each utterance, we append the desired relations separately and pass them to the COMET encoder. The activations returned by the encoder are directly used as common-sense representation vectors. Utterances with multiple sentences are broken down into individual sentences. The common-sense activations for the constituent sentences are averaged to get the representation for the utterance. Representation vectors for all nine available relations are extracted this way for each utterance. These are then used as needed in the ERC models.

4.4 Training

We use DialogueRNN for all the experiments of incorporating implicit CSK. For, explicit incorporation, placeholder for CSK feature vectors and the attention network are added to the DialogueRNN model and the attended common-sense vector is concatenated with the party GRU update vectors. In the DialogueRNN model, the party state dimension is 500, the global state GRU is 500 and the emotional state GRU is 300. For incorporating the textual features from RoBERTa,

we added 1d batch normalization layers for each of the four layer activations. The final input to DialogueRNN is the average of the normalized RoBERTa vectors. Same setup is used for all experiments with implicit CSK incorporation. DialogueRNN is trained for 60 epochs 1×10^{-4} learning rate and 0.1 dropout rate. All hyper-parameters are kept consistent through all the experiments and are provided in Appendix for reproduction.

COSMIC uses smaller dimensions for party state (150) and global state (150) internal representations introduces two new states of the same dimension with the intention of capturing commonsense knowledge about the speaker. For training COSMIC, the extracted CSK vectors are used in conjunction with the utterance feature vectors. COSMIC utilizes five relations from the total nine relations available from COMET. COSMIC is trained for 60 epochs 1×10^{-4} learning rate and 0.25 dropout rate. All hyper-parameters are kept consistent through all the experiments and are provided in Appendix for reproduction.

COMET and RoBERTa feature extractions are performed only once per setup. The model training is performed 10 times and the results are averaged to compute the final results. Random seeds are not controlled to get a better approximation of the true expected values for metrics.

4.5 Evaluation

F1 score is calculated and recorded for each emotion label. The weighted average of these F1 scores is the metric we use to evaluate models. Taking a weighted average gives classes importance in proportion to their sizes in the dataset. Confusion matrix gives us the information about which emotions are mixed up more by the models. Overall accuracy of the predictions is also tracked for each experiment.

Predicting the emotions correctly when the emotions of a person change during the conversation is particularly tricky. Moreover, there are additional applications that capturing emotion turns enable. Thus, in the post modelling step, we calculate the model's accuracy in capturing the correct emotion after emotion shifts. An emotion shift happens when a particular person's emotion changes from their previous emotion during the conversation. After predicting the labels, we filter out the test utterances where the speaker's emotion has shifted and observe the accuracy of the

predicted emotions for that subset.

Another challenge pertaining to ERC tasks is the ability of models to distinguish between fine-grained emotions. From the emotions available in IEMOCAP dataset, *(Happy/Excited)* and *(Angry/Frustrated)* are two such pair's which are most often confused by the model. In an attempt to gain more understanding about this problem, we track the model's accuracy at fine-grained emotion prediction. A confusion matrix involving only the fine grained classes is generated and the total correctly predicted count is divide by the total count of samples that involved only the pair of emotions. The results from both *(Happy/Excited)* and *(Angry/Frustrated)* pairs are combined to get the final fine-grained emotion accuracy for the predictions.

5. RESULTS

5.1 COMET Common-Sense Knowledge

The results of Common-Sense Knowledge extraction for utterances from COMET are discussed below. Table 5.1 summarize the results across all the relations. For each utterance in the train, valid and test split of IEMOCAP dataset, COMET is queried for each of the nine relations. The relations intended to capture effects on others namely `oWant`, `oEffect` and `oReact`, have high percentage of no results each. A no result indicate that the decoder algorithm could not sample an appropriate object. COMET is not able to provide much useful information regarding the effect on others from the utterances. These relations are not used in implicit CSK incorporation. `oEffect` is used in explicit CSK incorporation with the rationale that although the sampling algorithm couldn't provide an object, the internal representation does constitute CSK that could be leveraged.

Relation	**Count of unique objects**	**Mean of object frequencies**	**S.D. of object frequencies**	**No result%**
`xNeed`	419	2.3	3.5	87.1
`xIntent`	443	6.5	24.7	61.4
`xAttr`	409	16.7	39.8	0.2
`xWant`	1249	5.5	22.1	0.9
`xEffect`	419	2.0	4.3	87.8
`xReact`	164	24.2	86.1	1.6
`oWant`	142	3.0	5.7	94.2
`oEffect`	6	2.0	2.2	99.6
`oReact`	15	21.4	32.4	95.9

Table 5.1: COMET result summary for IEMOCAP dataset. Total utterances in the train, valid and test splits combined are 7433. No results count are excluded while counting mean and standard deviation (S.D.).

The self relations, xNeed, xIntent and xEffect, have a good percentage of no-results as well. However, they do provide substantial unique objects with a good distribution as seen from the mean and standard deviation of the object frequencies. These relations, when present, provides unique common-sense inferences for the utterances. xNeed and xIntent are causal relations, while xEffect is an effect relation. xNeed and xIntent are used for implict CSK incorporation and xEffect is used for explicit incorporation.

xAttr and xWant have negligible no-results (<1%) and return a diverse object distribution. Especially, xWant provides variegated common-sense reasoning for the utterances with an object repeating only 5.5 times throughout the dataset. Both of these relations are used in implicit CSK incorporation to augment in language model fine-tuning. xReact although providing results for most of the utterances, is often repetitive. Many of the frequently repeated objects are overlapping the emotions that we intend to predict, as shown in Table 5.2. As a result, this relation is avoided from both implicit and explicit knowledge incorporation as it might end up inducing noise.

xNeed		xIntent		xAttr	
none	6468	none	4562	determined	610
to have a job	42	to be a good friend	393	careless	398
to have a phone	29	to be a good person	210	curious	352
to get in the car	23	to be helpful	175	smart	278
xEffect		**xWant**		**xReact**	
to be successful	625	none	6523	happy	3492
gets yelled at	54	to get a drink	585	sad	922
gets drunk	33	to be left alone	224	relieved	336
cries	32	to be a good person	185	satisfied	301

Table 5.2: Top-4 objects results per relation from COMET for IEMOCAP dataset utterances.

5.2 Implicit CSK incorporation

Incorporating CSK implicitly is attempted by fine tuning RoBERTa using two different tasks, MLM training and labelled emotion classification. The baseline is obtained by fine-tuning RoBERTa on basic IEMOCAP without CSK for labelled emotion classification task. In all the experiments, vector representations of utterances are extracted from RoBERTa and DialogueRNN model is trained on them to generate the output. The results are shown in Table 5.3.

Model	Fine-tuning		Per emotion F1 score						Total	
	MLM	**Labelled**	**Happy**	**Sad**	**Neutral**	**Angry**	**Excited**	**Frustrated**	**W.Avg. F1**	**Acc.**
$DRNN_{Glove}$	-	-	36.61	78.80	59.21	**65.28**	**71.86**	**58.91**	63.40	62.75
$DRNN_{RoBERTa}$	-	Base	**51.56**	83.14	67.26	59.56	64.22	57.47	64.58	64.66
$DRNN_{Impl.CSK}$	CSK	Base	49.89	77.05	64.91	59.42	61.98	56.93	62.42	62.43
$DRNN_{Impl.CSK}$	CSK	CSK	44.98	82.00	65.89	60.31	**66.10**	55.11	63.39	63.48
$DRNN_{Impl.CSK}$	-	CSK	47.56	**85.16**	**68.16**	**61.41**	64.53	**58.23**	**65.15**	**65.06**

Table 5.3: Implicit CSK incorporation results for DialogueRNN on IEMOCAP. Fine-tuning shows the data used for two stages of RoBERTa fine-tuning, MLM and Labelled emotion classification. Base denotes just the utterances and CSK denotes extended utterances appended with CSK knowledge from COMET. All results are average of 10 runs. Test scores are calculated on the best validation F-score.

MLM training using the extended utterances hurts the performance of the model and it dips below the baseline model. The intuition behind MLM not working might be noisy CSK. The analysis of CSK extraction results (Table 5.1) showed repetitive and overlapping common-sense objects for many input utterances. These would introduce unnecessary biases in the language representations and that ends up degrading the performance compared to the baseline which performs only labelled emotion classification fine-tuning using the utterances. Performing labelled emotion classification using CSK appended utterances after the MLM fine-tuning seems to abate part of the negative impact from MLM.

In the last experiment, the MLM training is skipped. RoBERTa is fine-tuned directly on the labelled emotion classification task using the CSK appended utterances. This setup outperforms the

baseline by a small margin in both weighted average F1 score (0.57) and accuracy (0.40) metrics. Further, it improves the F1-score across all but one emotions. As we perform labelled emotion classification training using the CSK extended utterances, RoBERTa fine-tuning is aided by the additional diversity in the dataset. Previously ambiguous utterances will now have more tokens in them. These tokens would be different enough for RoBERTa to manipulate the utterance representations significantly. As we use the same extended utterance for feature extraction, previously ambiguous utterances are now more easily distinguishable. Thus, the marginal improvement in performance when using CSK appended utterances can be attributed to the addition of new diverse set of tokens. Table 5.4 shows an example where the correct label was predicted with the help of appended CSK.

Speaker	**Utterance**	**True Emotion**	**Predicted Emotion**	
			w/o CSK	**w CSK**
Person A	Yeah, they're going to do some sort of memorial service or something.	sad	sad	sad
Person B	Cool, Well, If you want me to go with you, I will. *{CSK: I am friendly. I want to go with them.}*	neu	**sad**	**neu**
Person A	Thanks.	sad	sad	sad
Person B	No problem.	neu	neu	neu

Table 5.4: Excerpt from a test set dialogue with emotion prediction with and without CSK (implicit). The actual CSK sentences appended to the utterance are shown as well.

5.3 Explicit CSK incorporation

We explicitly incorporate CSK in models by extracting and using CSK feature vectors directly. Explicit incorporation of CSK in model is attempted by COSMIC. They maintain a person's internal states using GRUs and update them using CSK vectors from COMET. Our approach is to apply contextual attention on the CSK knowledge throughout the conversation to get a relevant CSK state vector. Then, we use it to update the single party state. IEMOCAP results on apply-

ing explicit CSK are summarized in Table 5.5. We compare the baseline with the two methods for incorporating CSK. We also experiment explicit CSK models with the feature vectors that have implicit CSK incorporated. Based on the results from the Implicit incorporation experiments, we use the feature vectors extracted from RoBERTa fine-tuned for labelled emotion classification using utterances extended with CSK.

Model	**Happy**	**Sad**	**Neutral**	**Angry**	**Excited**	**Frustrated**	**W.Avg. F1**	**Acc.**
$DRNN_{RoBERTa}$	**51.56**	83.14	67.26	59.56	64.22	57.47	64.66	64.58
$DRNN_{Impl.CSK}$	47.56	85.16	**68.16**	**61.41**	64.53	58.23	**65.15**	65.06
$COSMIC$	42.56	81.70	63.83	57.79	**69.15**	**63.41**	64.87	**65.11**
$COSMIC_{Impl.CSK}$	43.16	80.63	64.50	58.73	67.92	62.47	64.56	64.67
$DRNN_{Expl.CSK}$	49.75	84.64	66.15	58.24	61.56	56.10	63.42	63.44
$DRNN_{Impl.+Expl.CSK}$	45.85	**85.33**	66.73	60.29	61.42	56.59	63.59	63.57

Table 5.5: Explicit CSK incorporation results comparison with Implicit CSK and COSMIC. All results are average of 10 runs. Test scores are calculated on the best validation F-score.

When we use explicit CSK incorporation using the attention method proposed in this work, the performance degrades over baseline. Both the accuracy and weighted F1-score are impacted. However, COSMIC is able to improve upon the baseline model by utilizing explicit CSK. This corroborates that the CSK knowledge, albeit noisy, is useful in emotion classification. Hence, we attribute the performance degradation to our method of CSK incorporation. The attention network might be redundant given that the global state GRU is already attended over and it is inferred from the party state. Further, noisy nature of CSK seems to be affecting the performance of the model adversely. Particularly, in the party state GRU, the update vector size has increased substantially on introducing the 768 dimensional CSK vector. This forces approximations to a lower order thereby losing valuable information from the feature vectors.

Using implicit and explicit incorporation together degrades the performance of the COSMIC model. One contrasting observation is that the base DialogueRNN model with implicit CSK outperforms both methods where we attempt to incorporate explicit CSK above implicit CSK. Once

incorporated implicitly, the explicit CSK information seems to be redundant to the model.

5.4 Implicit vs Explicit

Best results through explicit CSK incorporation are observed from COSMIC and the best results using implicit CSK incorporation are observed by labelled fine-tuning on utterances extended with CSK. Comparing the results from these two methods, shows marginal differences. Weighted F1 score of the implicit method is more than explicit method by a small percentage while accuracy show negligible difference.

Analyzing the differences in CSK extraction for these two methods might give us intuition about the difference in F1-score. For implicit incorporation we use a *greedy* sampler and discard all the *none* results thereby filtering huge portion of low confidence CSK. This is not the case with explicit CSK incorporation where we use internal activations with no way of knowing the model's confidence in the CSK for particular utterances. Even with this difference, the performance of the two methods is comparable in our experimental setup.

5.5 Emotion Shift

Model	**Emotion shift**	**Fine-grained emotions**	**Emotion shifts in predictions (%)**
$DRNN_{RoBERTa}$	51.14	69.46	19.0
$COSMIC$	50.63	**72.90**	**20.5**
$DRNN_{Impl.CSK}$	**52.37**	68.67	17.9

Table 5.6: Accuracy in predicting emotion shifts of a speaker and distinguishing fine-grained emotions *(Happy, Excited)* and *(Angry, Frustrated)* of different ERC models.

Emotion shifts are utterances where the emotion of the speaker changes. These are particularly challenging to capture as sometimes the evidence for emotion change is subtle. The accuracy in predicting emotion shifts of the implicit CSK model improves compared to COSMIC and DialogueRNN (See Table 5.6). The model itself relies heavily on the context and requires

strong evidence for shifting the predicted emotion of an utterance from the person's preceding emotion. Our understanding is that the context-independent common-sense knowledge about how the speaker feels biases the features more towards their true label during fine-tuning. This helps the downstream model in identifying a shift in speaker's emotion.

5.6 Fine-grained emotion analysis

The emotion pairs *(Happy,Excited)* and *(Angry, Frustrated)* are difficult to distinguish between and pose a challenge in emotion classification task. Table 5.6 shows accuracy results in distinguishing pairs of fine-grained emotions. COSMIC is significantly better than DialogueRNN and implicit CSK incorporation is worse than DialogueRNN. It is observed that the performance on fine-tuned emotions is inverse of the performance on emotion shifts. To get more insight into what might be happening here, we take a look at the % of emotion turns in the predictions of the models. COSMIC has the most turns in its prediction and implicit CSK has the least. COSMIC has a greater tendency to shift emotions, often to a close emotion. This flexibility seems to helping COSMIC in distinguishing fine-grained emotions.

6. SUMMARY

6.1 Conclusion

Incorporating automatically generated common-sense is a difficult task due to the challenges posed by the noise in generated knowledge and interfacing it with the many stages in the downstream NLP task. Despite of these challenges, incorporating CSK does improve results over the baseline in our experiments. This substantiate the fact that common-sense inclusion is useful in the conversational emotion recognition task. Interacting with the different experiment stages and analyzing the results has afforded us several conclusions which are detailed in this section.

Common-sense knowledge generation has not yet matured and has limited distribution in terms of output samples. This exacerbates when the knowledge base used for training the common-sense knowledge generator differs from the target domain. However, good sampling algorithms acting as filters could be used to query quality common-sense inferences. These instances, albeit small in number, has the potential of adding value to the input features.

Fine-tuning language models for masked token prediction on fabricated data is best avoided as they are extremely prone to noise. However, fine-tuning for labelled sentence classification task is much more robust to noisy data. Even with noisy common-sense knowledge and rudimentary sentence formation, fine-tuning language model on labelled emotion classification improves performance in the downstream task.

Our experiments show that implicit incorporation of common-sense knowledge is fairly simple and way of leveraging common-sense. It gives comparable results to complex explicit incorporation methods. Fine-tuning language models for labelled classification task helps in subsiding the noise in generated common-sense thus freeing the model from this responsibility. Thus, it leverages future advances in automatic knowledge-generation directly without substantial modelling efforts.

Incorporating common-sense knowledge explicitly has to be carefully modelled with a proper

understanding of the inferences that can be made from the available common-sense knowledge. When implemented alongside implicit CSK incorporation, explicit CSK incorporation ends up being redundant and even degrades the model's performance. As the input features have to compete with the CSK features for relevance in inference, some information infused into the feature vectors through language models is dropped impacting the overall performance.

6.2 Challenges

High variance was observed in results of subsequent runs of the same model. Because of this, it is particularly hard to attribute predictions on particular utterances to the models. As a lot of information is incorporated in the feature vectors itself using language models, the results from changes in modelling approach provide very little insight. Validating the correctness of different modelling approaches like speaker state modelling, listener modelling, global context modelling, etc have to rely heavily on the final result. Also, validating the effectiveness of these factors independently require ablation techniques for analysis.

IEMOCAP dataset is a small dataset for training large language models like RoBERTa especially for training on MLM task. The size of the dataset becomes a challenge as the model starts over-fitting after a few epochs given the high dimensionality compared to the number of samples.

6.3 Future Scope

- The approaches mentioned in this work should be tested on other ERC datasets like MELD, DailyDialog, EmoContext, etc. to solidify the findings and gather more insights.

- Better common-sense knowledge has to be incorporated to truly understand its impact in ERC task. COMET has the ability to train on any knowledge graph. Using a conversation specific knowledge graph with relations like ATOMIC that captures the causes and effect of the utterance would be an ideal approach to develop rich common-sense knowledge for ERC. From COMET, combinations of different relations would help in segregating a highly effective set of relations specific to ERC. Currently, sentence formation using CSK from comet is rudimentary. Generating more coherent sentences with the CSK and utterances

could help in fine-tuning using MLM task.

- To comprehend the effect of explicit CSK incorporation, the internal state changes and their impact on emotion classification shall be studied in depth. This is will help in understanding the interactions between the utterances and CSK features. An effective approach could then be devised for explicit incorporation of common-sense in the ERC task.

- ERC task could take a huge leap in performance if emotion shifts could be predicted more accurately. Experiments with the model could be accompanied with experiments in data sampling thereby putting more emphasis on dialogues with emotion shifts. Auxiliary network intended to capture just emotion shifts, with integration in the contextual models, could be explored.

REFERENCES

[1] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019.

[2] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[3] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 6818–6825, 2019.

[4] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4762–4779, 2019.

[5] D. Olson, "From utterance to text: The bias of language in speech and writing," *Harvard educational review*, vol. 47, no. 3, pp. 257–281, 1977.

[6] R. Plutchik, "A psychoevolutionary theory of emotions," 1982.

[7] P. Ekman, "Facial expression and emotion.," *American psychologist*, vol. 48, no. 4, p. 384, 1993.

[8] I. Bakker, T. Van Der Voordt, P. Vink, and J. De Boon, "Pleasure, arousal, dominance: Mehrabian and russell revisited," *Current Psychology*, vol. 33, no. 3, pp. 405–421, 2014.

[9] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[10] I. D. Wood, J. P. McCrae, V. Andryushechkin, and P. Buitelaar, "A comparison of emotion annotation approaches for text," *Information*, vol. 9, no. 5, p. 117, 2018.

[11] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883, 2017.

[12] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pp. 2470–2481, 2020.

[13] C. Huang, A. Trabelsi, and O. R. Zaiane, "Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 49–53, 2019.

[14] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100943–100953, 2019.

[15] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," *IEEE Transactions on Affective Computing*, 2020.

[16] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 2594–2604, 2018.

[17] A. Shenoy, A. Sardana, and N. Graphics, "Multilogue-net: A context aware rnn for multimodal emotion detection and sentiment analysis in conversation," *ACL 2020*, p. 19, 2020.

[18] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.

[19] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, vol. 2018, p. 2122, NIH Public Access, 2018.

[20] D. Sheng, D. Wang, Y. Shen, H. Zheng, and H. Liu, "Summarize before aggregate: A global-to-local heterogeneous graph inference network for conversational emotion recognition," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4153–4163, 2020.

[21] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2020.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[23] P. Zhong, D. Wang, and C. Miao, "Knowledge-enriched transformer for emotion detection in textual conversations," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 165–176, 2019.

[24] J. He, B. Wang, M. Fu, T. Yang, and X. Zhao, "Hierarchical attention and knowledge matching networks with information enhancement for end-to-end task-oriented dialog systems,"

IEEE Access, vol. 7, pp. 18871–18883, 2019.

[25] A. Kumar, D. Kawahara, and S. Kurohashi, "Knowledge-enriched two-layered attention network for sentiment analysis," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 253–258, 2018.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[27] Y.-H. Huang, S.-R. Lee, M.-Y. Ma, Y.-H. Chen, Y.-W. Yu, and Y.-S. Chen, "Emotionx-idea: Emotion bert–an affectional model for conversation," *arXiv preprint arXiv:1908.06264*, 2019.

[28] D. Hazarika, S. Poria, R. Zimmermann, and R. Mihalcea, "Conversational transfer learning for emotion recognition," 2020.

[29] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[30] T.-Y. Chang, Y. Liu, K. Gopalakrishnan, B. Hedayatnia, P. Zhou, and D. Hakkani-Tur, "Incorporating commonsense knowledge graph in pretrained models for social commonsense tasks," in *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 74–79, 2020.

[31] M. Sap, R. LeBras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," 2019.

[32] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," 2018.

[33] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 48–53, 2019.

[34] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions in text using deep learning and big data," *Computers in Human Behavior*, vol. 93, pp. 309–317, 2019.

[35] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "Semeval-2019 task 3: Emocontext contextual emotion detection in text," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 39–48, 2019.

APPENDIX A

REPRODUCTION DETAILS

A.1 Hyperparameters

A.1.1 Fine-tuning RoBERTa MLM

Parameter	**Value**
Total # of training steps	12500
# of updates for learning rate warmup	1000
Max sequence length	512
# of positional embeddings	512
# of sequences per batch	8
Update frequency	4
Dropout	0.2
Weight decay	0.1
Clip norm	0.0
Peak learning rate	0.0005
Optimizer	Adam
Adam coefficients	0.9, 0.98

Table A.1: Hyperparameters for fine tuning RoBERTa for MLM task

A.1.2 RoBERTa Labelled Emotion Classification

Parameter	Value
Total # of training steps	4840
# of updates for learning rate warmup	290
# of classes	6
# of sequences per batch	8
Max positions	512
Max positions	5000
Max epochs 30	
Update frequency	4
Dropout	0.2
Weight decay	0.1
Clip norm	0.0
Peak learning rate	1e-6
Optimizer	Adam
Adam coefficients	0.9, 0.98

Table A.2: Hyperparameters for fine tuning RoBERTa for labelled emotion classification task

A.1.3 DialogueRNN

Parameter	Value
active_listener	False
attention	'general'
batch_size	30
class_weight	True
dropout	0.1
epochs	60
l2	1e-05
lr	0.0001
no_cuda	False
rec_dropout	0.1
tensorboard	False

Table A.3: Training parameters for DialogueRNN

A.1.4 COSMIC

Parameter	Value
active_listener	False
attention	'general2'
batch_size	16
class_weight	False
dropout	0.25
epochs	60
l2	0.0003
lr	0.0001
mode1	2
no_cuda	False
norm	3
rec_dropout	0.1
residual	False
seed	100
tensorboard	False

Table A.4: Training parameters for COSMIC