From Chatter to Matter: Addressing Critical Steps of Emotion Recognition Learning in Task-oriented Dialogue

Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gašić
Heinrich Heine University Düsseldorf, Germany
{fengs,lubis,ruppik,geishaus,heckmi,linh,niekerk,revuk100,gasic}@hhu.de

Abstract

Emotion recognition in conversations (ERC) is a crucial task for building human-like conversational agents. While substantial efforts have been devoted to ERC for chit-chat dialogues, the task-oriented counterpart is largely left unattended. Directly applying chit-chat ERC models to task-oriented dialogues (ToDs) results in suboptimal performance as these models overlook key features such as the correlation between emotions and task completion in ToDs. In this paper, we propose a framework that turns a chit-chat ERC model into a task-oriented one, addressing three critical aspects: data, features and objective. First, we devise two ways of augmenting rare emotions to improve ERC performance. Second, we use dialogue states as auxiliary features to incorporate key information from the goal of the user. Lastly, we leverage a multi-aspect emotion definition in ToDs to devise a multi-task learning objective and a novel emotion-distance weighted loss function. Our framework yields significant improvements for a range of chit-chat ERC models on EmoWOZ, a large-scale dataset for user emotion in ToDs. We further investigate the generalisability of the best resulting model to predict user satisfaction in different ToD datasets. A comparison with supervised baselines shows a strong zero-shot capability, highlighting the potential usage of our framework in wider scenarios.

1 Introduction

Emotion recognition in conversations (ERC) is a crucial task in conversational artificial intelligence research because it lays the foundation for affective abilities in computers such as empathetic response generation (Picard, 1997). Over years, it has shown values in downstream applications such as opinion mining (Colneric and Demšar, 2020) and human-like dialogue modelling (Zhou et al., 2018).

Dialogue systems can be broadly categorised into two categories: (1) chit-chat or open-domain systems and (2) task-oriented dialogue (ToD) systems. Chit-chat systems are set up to mimic human behaviours in a conversation (Jurafsky and Martin, 2009). There are no particular goals associated with the dialogue and the system aims to keep the user engaged with natural and coherent responses. On the other hand, ToD systems are concerned with fulfilling user goals, such as information retrieval for hotel booking (Young, 2002).

```
A: We have a holiday next week, don't we ?
B: Yes, on Monday .
A  What're you going to do ?
B: I'm probably going to spend the day looking at cars.
   ...
```

(a) Chit-chat dialogue from Li et al. (2017)

```
U: I am planning a vacation and really could use some
   help finding a good place to stay in town. I've
   never been to Cambridge before.
S: which side of town do you prefer and what is the
   price range?
U: It doesn't matter. What do you recommend?
S: alexander bed and breakfast is a guest house in the
   centre area. Would you like to book a room?
   ...
```

(b) Task-oriented dialogue from Budzianowski et al. (2018)

Figure 1: Comparison of dialogues about holiday in chit-chat dialogues and task-oriented dialogues.

Recently, the difference between chit-chat and ToD systems have been blurred by the utilisation of pre-trained language models as back-bone to both types of systems. However, emotions in ToDs and chit-chat dialogues play different roles and are therefore expressed differently (Feng et al., 2022). This highlights the need for dedicated emotion modelling methods for each system.

As illustrated in Figure 1, in chit-chat dialogues, speakers make use of emotions to facilitate communication by, for example, raising empathy as a result of emotion-eliciting situations or topics. On the other hand, emotions in ToDs are centred around the user's goal, and therefore emotion cues lie in both the user's wording and the task performance.

Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue, pages 85–103
September 11–15, 2023.

While many large-scale corpora for emotions in chit-chat dialogues exist (Busso et al., 2008; McKeown et al., 2012; Lubis et al., 2015; Li et al., 2017; Zahiri and Choi, 2018), there are considerably fewer resources for emotions in ToDs. EmoWOZ, which evolved from MultiWOZ, a widely used ToD dataset, is one notable exception (Feng et al., 2022). It contains a novel emotion description that is designed for ToDs and inspired by the Ortony-Clore-Collins (OCC) model (Ortony et al., 1988). Emotion is described in terms of three aspects: **valenced** (positive or negative) reactions towards **elicitors** (operator, user, or event) in a certain **conduct** (polite or impolite). However, due to the nature of ToDs, the occurrence of some emotions (e.g. users expressing feelings about their situations) are very rare, leading to a class imbalance in the corpus.

Similarly, advancements on the ERC task are mainly focused on chit-chat dialogues, involving an array of diverse factors from speaker personality (Majumder et al., 2019) to commonsense knowledge (Ghosal et al., 2020). Nevertheless, since these models are designed for chit-chat dialogues, they overlook how emotions are triggered and expressed with respect to goal completion in task-oriented context. The work of Devillers et al. (2003) is among one of the earliest and very few to address emotion detection in ToDs but uses generic unigram models instead of dedicated approaches.

In this work, we tackle critical steps of ERC in ToDs from three angles: the data, the features, and the learning objective. In particular,

Data: we address the poor ERC performance of particularly rare emotions in ToDs via two strategies of data augmentation (DA),

Features: we leverage dialogue state information and sentiment-aware textual features,

Objective: we exploit the three aspects of emotions, namely valence, elicitor, and conduct, in two ways: as a multi-task learning (MTL) objective and to define a novel emotion-distance-weighted loss (*EmoDistLoss*).

To the best of our knowledge, our work is the first to provide dedicated methods for emotion recognition in ToDs. Our experiments and analyses show that our framework leads to significant improvements for a range of chit-chat ERC models when evaluated on EmoWOZ.

We further investigate the generalisability of the best resulting model to predict user satisfaction in various ToD datasets under zero-shot transfer. Our model achieves comparable results as supervised baselines, demonstrating strong zero-shot capability and potential to be applied in wider scenarios.

2 Related Work

2.1 ERC Datasets

Early work on ERC relied on small scale datasets (Busso et al., 2008; McKeown et al., 2012; Lubis et al., 2015). More recently, a few large-scale datasets have been made available to the research community. They contain dialogues from emotion-rich and spontaneous scenarios such as daily communications (Li et al., 2017) and situation comedies (Zahiri and Choi, 2018).

For ToDs, the majority of available datasets address only one particular aspect of emotions such as sentiment polarity (Saha et al., 2020; Shi and Yu, 2018), user satisfaction (Schmitt et al., 2012; Sun et al., 2021), and politeness (Hu et al., 2022; Mishra et al., 2023). For more fine-grained emotions, Singh et al. (2022) constructed EmoInHindi for emotion category and intensity recognition in mental health and legal counselling dialogues in Hindi, and Feng et al. (2022) released EmoWOZ, which concerns user emotions in human-human and human-machine in information-seeking dialogues. Among these datasets, EmoWOZ has the largest scale, accompanied with a label set tailored to the task-oriented scenario.

2.2 Data Augmentation (DA)

DA is an effective approach to improve model performance by improving data diversity without explicitly collecting more data. While textual DA can be performed in the feature space via interpolation and sampling (Kumar et al., 2019), it is commonly performed in the data space for controllability. Rule-based methods involve operations such as insertion and substitution (Wei and Zou, 2019). While they are easy to implement, the diversity in augmented samples depends on the complexity of the rules. On the contrary, model-based methods are more scalable. These typically include the use of language models (Jiao et al., 2020), translation models (Xie et al., 2020a), and paraphrasing methods (Hou et al., 2018).

Additional training samples can also be obtained from unlabelled data via weak supervision (Ratner et al., 2017). To generate the automatic labels, a single model or an ensemble of models may

be used. This method can be interpreted as self-augmentation (Xu et al., 2022), self-training (Xie et al., 2020b), or distillation (Radosavovic et al., 2017).

DA has also been also deployed in ToD modelling. Hou et al. (2018) generated samples by paraphrasing delexicalised utterances. Gritta et al. (2021) conceptualised ToDs into transitional graphs and generate new dialogue paths by sampling. Heck et al. (2022) proposed a weak supervision framework to address the lack of fine-grained span labels for dialogue state tracking. DA for emotions in ToDs requires careful considerations to avoid emotion mismatch and is not yet explored.

2.3 ERC Models and Features

Text-based ERC is in essence a text classification problem with an emphasis on contextual modelling. Poria et al. (2017) proposed a recurrent neural network (RNN) for multimodal ERC. The follow-up work of Majumder et al. (2019) considered speaker-specific context. ERC performance has been continuously improved by techniques such as incorporating external knowledge (Ghosal et al., 2020) and contrastive learning (Song et al., 2022).

Sentiment-aware Embeddings Word-vector embeddings tailored for a particular natural language processing task can effectively improve the performance for that task (Naseem et al., 2021). In a similar vein, Tang et al. (2014) incorporated sentiment classification objectives in the training of the word embedding model of Collobert and Weston (2008) specifically for sentiment analysis. Yu et al. (2017) refined static word embeddings with the aid of a sentiment lexicon. Later, many sentiment-aware variants of pre-trained language models were obtained by incorporating sentiment-related objectives in training (Xu et al., 2019; Yin et al., 2020; Zhou et al., 2020). They successively achieved state-of-the-art performance in sentiment analysis tasks among language representation models.

2.4 Learning Objectives for ERC Models

ERC is often considered a single-label sequential classification problem. Using softmax cross-entropy loss has been the norm in the training of deep learning ERC models for categorical emotions (Poria et al., 2017; Zhong et al., 2019; Ghosal et al., 2020; Kim and Vossen, 2021) or quantised emotion dimensions (Cerisara et al., 2018; Wang et al., 2020). However, this simplistic cross-entropy loss ignores the inter-class relations and output probabilities on incorrect classes.

Chen et al. (2019) proposed to suppress the output probabilities of incorrect classes equally while minimising the standard cross-entropy loss. Hou et al. (2016) proposed squared earth mover's distance to penalise the misclassifications according to a ground distance matrix that quantifies the dissimilarities between classes for image age estimation and aesthetics estimation.

Although highly suitable for emotions, learning from misclassifications is rarely considered because the distance between emotion classes is hard to quantify. Therefore, we propose to leverage the structured label definition of EmoWOZ to model inter-class similarity.

Multi-task Learning (MTL) is a technique for learning tasks in parallel using a shared representation. It aims to improve generalisation by using the information in training signals of related tasks as an inductive bias (Caruana, 1997). In emotion recognition, auxiliary tasks include topic classification (Wang et al., 2020) and personality traits (Li et al., 2021). When co-labels are not available, it is also possible to leverage aspects of emotion for additional labels such as valence-arousal (Kim et al., 2017). In this work, we exploit the valence-elicitor-conduct labels in EmoWOZ for MTL.

3 Background

3.1 User Emotion Recognition

We formulate the task as recognising one emotion class e_t from a set of n discrete emotions $E = \{e^1, e^2, ..., e^n\}$ in the user turn u_t, given a dialogue history $H_t = [u_t, s_{t-1}, u_{t-1}, ..., s_1, u_1]$, where s denotes system turns and u denotes user turns. Unlike existing chit-chat ERC models, which are often built for static analysis on the dialogue as a whole, real-time ERC in ToDs does not consider future utterances in dialogue.

3.2 User Satisfaction Prediction

User satisfaction prediction aims to predict one satisfaction level c_t from a set of m discrete levels $C = \{c^1, c^2, ..., c^m\}$ in the user turn u_t, given all previous turns $P_t = [s_{t-1}, H_{t-1}]$. This task differs from ERC in that the user turn u_t is not available as a part of model input. Since user satisfaction is highly correlated with the valence aspect in user emotion, this task can also be viewed as

user emotion prediction. This is an important task in building ToD systems and has been used for user simulation and system evaluation (Sun et al., 2021).

4 Emotion Recogniser for Task-oriented Dialogues (ERToD)

In this section, we propose our ERToD framework that adapt chit-chat ERC models to the task-oriented domain, as illustrated in Figure 2.

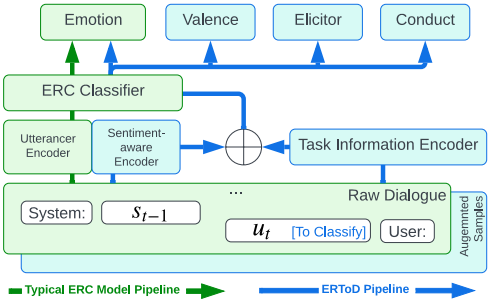

Figure 2: Our proposed ERToD Framework.

4.1 Data Augmentation

Unlike emotions in chit-chat dialogues, resources for emotions in ToDs are very limited. In addition, the data scarcity not only lies in the lack of linguistic diversity but also in the limited domains and actions in which emotions are expressed.

In ToDs, user's emotional expressions have different degrees of connection to the dialogue task. For example, a user can express dissatisfaction towards the system by pointing out the system's mistake. In such a case, simply replacing or paraphrasing the user's utterance based on emotion can potentially break the consistency of the task flow in the context. Such emotions are *context-dependent*.

On the other hand, *context-independent emotions* are expressed without any connection to the user goal, such is the case with abusive utterances. Due to the lack of connection, a simple replacement with a different abusive sentence can fit into the context well without impairing the consistency of task flow in the dialogue.

To obtain augmented samples with meaningful and coherent context, we adopt two different strategies of DA according to the degree of context dependency of emotional expressions.

Context-independent Emotions To augment samples for a target emotion e, we select a user utterance u' with the equivalent label from other dialogue datasets. We then use it to replace the user utterance u_t having label e in the training data while keeping the original context $[s_{t-1}, u_{t-1}, ..., s_1, u_1]$. The new sample is obtained as $H'_t = [u', s_{t-1}, u_{t-1}, ..., s_1, u_1]$.

Context-dependent Emotions We first sample a pool of unlabelled candidate dialogues $H'_t = [u'_t, s'_{t-1}, u'_{t-1}, ..., s'_1, u'_1]$ from other ToD datasets. We train a classifier with an uncertainty estimator to identify the emotion label e_t of the user utterance u_t and its confidence in each candidate:

$$p(e_t), \text{conf}(e_t) = \text{UncertaintyClassifier}(H'_t) \quad (1)$$

The candidate is selected for emotion e_t only if $\text{conf}_t(e)$ is above a confidence threshold θ.

4.2 Task Information Encoder

We use a dialogue state tracker (DST) to determine the status of goal completion at each turn. In ToDs, the dialogue state describes the system's understanding of the the user's goal up to that point in the dialogue (Young et al., 2010). It encodes dialogue progress in an abstractive manner.

Here as a proof of concept, we use an ontology-dependent DST, which means the concepts that the system can talk about are pre-determined. While we can eliminate the ontology dependency by, for example, using an ontology-independent DST and extracting task features from dialogue state description in natural language, this goes beyond the scope of this work. The DST takes the dialogue history to determine SemDS_t, the current dialogue state in semantic form. It is stored as a dictionary that records slots and filled values. SemDS_t is then converted into a vector of 0/1's, indicating whether a particular slot has been filled.

$$V_t = \text{Vectoriser}(\text{SemDS}_t) \quad (2)$$

To account for the change of dialogue state, which depicts how the system performs locally, we concatenate dialogue states of three consecutive turns to obtain a contextual dialogue state vector.

$$\widetilde{V}_t = V_t \oplus V_{t-1} \oplus V_{t-2} \quad (3)$$

$\text{V}_{t\leq 0}$ are zero vectors, representing the state before the dialogue starts. $\widetilde{V}_t$ is then fed into a trainable fully connected (FC) layer.

$$S_t = \text{FC}(\widetilde{V}_t) \quad (4)$$

Feature Fusion for Emotion Classification For a chit-chat ERC model with an arbitrary utterance encoder, $R_t = \text{Encoder}(H_t)$, i.e. R_t is the encoded representation of the dialogue history H_t. The utterance encoder is replaced with a sentiment-aware

encoder in our framework (see Figure 2).

The utterance and the task information encodings are fused via concatenation and fed into the emotion classifier. The output probability of all emotion classes in utterance u_t is given by:

$$p_t = \text{Softmax}(\text{Classifier}(R_t \oplus S_t)) \quad (5)$$

4.3 Learning Objectives

4.3.1 Emotion-Distance Weighted Loss

Emotion classification is a very challenging task due to the subjectivity in the perception of emotion. Since some emotions are more similar to each other than others, it may be advantageous to distinguish marginally wrong recognitions (satisfied vs excited) from extremely wrong ones (satisfied vs dissatisfied). Furthermore, different misclassifications can elicit different user reactions to the dialogue agent. For example, perceiving satisfaction when the user is neutral may or may not annoy the user, but accusing the user of abusive behavior by mistake is a serious offense to the user. Therefore, it is intuitive to penalise misclassifications according to (1) the distance from the label and (2) output probabilities on incorrect labels.

Defining the Emotion Distance Since emotion labels in EmoWOZ are defined in three aspects, we can define the distance between emotion labels in terms of their distance on each aspect. A matrix D is defined where each element $D(i, j)$ is a vector containing the distance between emotion label i and j in each of three aspects (valence, elicitor, and conduct). The matrix D is symmetric with vector-valued entries.

$$D(i, j) = [d_{val}(i, j), d_{eli}(i, j), d_{con}(i, j)] \quad (6)$$

The final distance is obtained by the sum of the distance in each aspect, followed by an addition of 1 and smoothing with the log operator. The addition of 1 ensures that the log distance is still 0 for identical labels.

$$\widetilde{D}(i, j) = \log(\text{sum}(D(i, j)) + 1) \quad (7)$$

Considering Misclassification Probabilities For each sample including the dialogue history H_t, we look at the softmax output from the model.

$$p_t = \text{Classifier}(H_t) \quad (8)$$

We aim to minimise the probability of each misclassification $p_t(e = e_i)$ where $e_i \neq \text{label}_t$. This is done by maximising $1 - p_t(e = e_i)$, the probability of the utterance *not* being wrongly recognised as e_i. We then calculate the log of this probability so that in the case of a perfectly correct recognition, the penalty from misclassification will be 0.

$$f(p_t) = \log(1 - p_t) \quad (9)$$

Obtaining Weights for Misclassifications We obtain the relevant row in matrix D that contains the distance between each emotion and the ground-truth label j of utterance u_t, followed by a normalisation to obtain a vector $w_{t,j}$ of normalised emotion-distance weights for all emotions.

$$o_{t,j} = \text{onehot}(\text{label}_t = j) \quad (10)$$

$$\widetilde{D}(:, j) = \widetilde{D} \times o_{t,j} \quad (11)$$

$$w_{t,j} = \widetilde{D}(:, j)/\text{sum}(\widetilde{D}(:, j)) \quad (12)$$

EmoDistLoss The final loss, which we name *EmoDistLoss*, is calculated from the negative weighted sum of log terms from Equation 9. Since the distance, hence the weight, between identical labels is 0, this calculation does not involve the output probability of the correct label.

$$\textit{EmoDistLoss}_t = -w_{t,j} \cdot f(p_t) \quad (13)$$

4.3.2 MTL via Emotional Aspects

In addition to the emotion classification head, we have a classification head for each emotion aspect from the label definition, namely the valence, the elicitor, and the conduct.

The overall classification loss L is a weighted sum of the loss from softmax outputs of four classification heads $L_{emo}, L_{val}, L_{eli}, L_{con}$ with a hyperparameter α.

$$L = \alpha L_{emo} + \frac{1}{3}(1-\alpha)(L_{val} + L_{eli} + L_{con}) \quad (14)$$

5 User Emotion Recognition in ToDs with ERToD

5.1 Experimental Set-up

5.1.1 Dataset

We train and test our models on EmoWOZ. It contains user emotion annotations for all dialogues from MultiWOZ (Budzianowski et al., 2018) and additional 1000 human-machine dialogues. It contains 7 emotion groups (see Table 1 and Appendix A for details). Four emotion classes are considerably rare: *fearful*, *apologetic*, *abusive*, and *excited*. DA examples can be found in Appendix B. Our primary aim of DA is to address the poor ERC performance on rare emotions rather than building a balanced dataset. While the later aim can be achieved with the aid of large language models for example, this is out of the scope of our work.

Class Name	Valence	Elicitor	Conduct	Count (%)
Neutral	Neutral	Don't Care	Polite	58,656 (70.1%)
Satisfied	Positive	Operator	Polite	17,532 (21.0%)
Dissatisfied	Negative	Operator	Polite	5,117 (6.1%)
Excited	Positive	Event/Fact	Polite	971 (1.2%)
Apologetic	Negative	User	Polite	840 (1.0%)
Fearful	Negative	Event/Fact	Polite	396 (0.5%)
Abusive	Negative	Operator	Impolite	105 (0.2%)

Table 1: EmoWOZ Emotion definition and distribution.

Augmenting Abusive Utterances The user sometimes becomes abusive towards the system. While this correlates with failure to satisfy the user goal, exact abusive expressions uttered by the user are usually independent of the context. Therefore, we apply our DA method for context-independent emotions for *Abusive*. We utilise ConvAbuse, a dataset for nuanced abusive behaviours in chit-chat conversations (Cercas Curry et al., 2021), for more diverse abusive expressions. In ConvAbuse, user utterances are labelled with type, target, strength, and directiveness. We filter for abuses on the system's intellectuality (labelled as `type=intellectual` and `target=system`) to better suit ToD context. We combine each selected utterance with the context of a random abusive utterance in EmoWOZ, resulting in 273 augmented samples.

Augmenting Fearful, Apologetic, and Excited Utterances Expressions of these emotions usually contain task information. *Fearful* and *Excited* usually co-occur with a description of the situation that prompts the user to interact with the system. *Apologetic* is frequently associated with a correction of search criteria. There is a strong connection between these emotion expressions and the progression of the task in the dialogue history. Therefore, we apply our DA method for these context-dependent emotions. We look for samples with desired emotions from other ToD datasets using automatic labels. We train a ContextBERT on EmoWOZ (see Section 5.1.2) with a 30% dropout on the BERT output. We train the model with 10 different seeds and run inferences on the training set of existing ToD datasets: Schema-Guided Dialogue (SGD, Rastogi et al. 2019), Taskmaster-1 (TM-1), and Taskmaster-2 (TM-2) (Byrne et al., 2019). In addition, we filter for common domains of EmoWOZ: *Hotels, RideSharing, Travel, Restaurants* in SGD, *RestaurantTable, PizzaOrdering, CoffeeOrdering, UberLyft* in TM-1, and *HotelSearch, Restaurants, FoodOrdering* in TM-2. The classification confidence is measured by votes from 10 models. We use a confidence threshold of 0.7 and cap the number of augmented samples at 1000 for each emotion, resulting in 268 *fearful*, 872 *apologetic*, and 1000 *excited* samples.

5.1.2 Baselines

We implement ERToD to a range of ERC models that have been used to benchmark EmoWOZ, as listed in Table 2. ContextBERT (Feng et al., 2022) and EmoBERTa (Kim and Vossen, 2021) are simple yet robust transformer-based ERC models, and they have similar spirits except that they respectively use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) as utterance encoder. They are both built on top of BERT by additionally considering dialogue context and speaker roles in the input. DialogueRNN (Majumder et al., 2019) and COSMIC (Ghosal et al., 2020) are RNN-based models. Following (Feng et al., 2022), we use DialogueRNN with either {GloVe(Pennington et al., 2014)+Convolutional Neural Network} or BERT as the utterance encoder. COSMIC additionally extracts features with a pre-trained commonsense model (Bosselut et al., 2019)[1]. It is important to note that after replacing the original utterance encoder with the sentiment-aware encoder (as described in Section 5.1.3), two variants of DialogueRNN essentially become the same model, and so do EmoBERTa and ContextBERT.

5.1.3 Training

In our task information encoder, we use SetSUMBT DST (van Niekerk et al., 2021) from ConvLab-3 toolkit (Zhu et al., 2022). SetSUMBT is a strong DST considering uncertainty with a joint goal accuracy of 52.26% on MultiWOZ 2.1 (Eric et al., 2020). The FC layer in Equation 4 has input/output dimensions of 1083 and 256 respectively and hyperbolic tangent activation (TanH, LeCun et al. 2015). We further replace the utterance encoders of chit-chat ERC models with SentiX, a sentiment-adapted BERT (Zhou et al., 2020).

We use our proposed *EmoDistLoss* for the emotion classification head and cross-entropy loss for MTL heads (valence, elicitor, and conduct). Since the elicitor of *Neutral* emotion is not distinguishable and therefore not explicitly defined in EmoWOZ, we mark the elicitor of *Neutral* samples

[1]COSMIC requires future utterances in recognising the current emotion whereas other models can be configured as either bidirectional or unidirectional. While we use unidirectional set-ups where possible to comply with our task formulation in Section 3.1, we are also interested in how ERToD improves COSMIC for static dialogue analysis in ToDs.

as *don't care*, and their loss in from elicitor classification is ignored. α in Equation 14 is set to 0.4 based on several rounds of hyperparameter tuning.

To calculate the *EmoDistLoss*, we use 1 as the unit distance and define the distance for each emotional aspect as illustrated in Appendix C. For valence, it is commonly adopted to consider negative and positive as two polarities and neutral in the middle (Socher et al., 2013). Therefore, the distance is 2 between positive and negative, and 1 between non-neutral and neutral. For emotion elicitors, we set the distance between *don't care* to any specific elicitor as 0.5 to penalise a "lazy" classifier that wrongly recognises the emotion as neutral. Doing so also results in a consistent shortest distance of 1 between any pair of specific elicitors.

We follow the default training set-up of each model except for ContextBERT. We reduce the context size of ContextBERT from 512 to 128, resulting in stronger performance and faster training.

5.1.4 Evaluation

We report F1 for each emotion. For overall performance, we report both macro F1 and weighted F1. Macro F1 considers each emotion equally and reflects the model's ability to recognise rare emotions. Weighted F1 is the weighted sum of F1 scores of each label. Weights are determined by the proportion of each emotion in the dataset. We exclude *Neutral* from calculating the averages as it makes up more than 70% of labels.

In addition, we also calculate the average emotion distance (AED) between the recognised emotion and the label to quantify how wrong the model is when it misclassifies. The AED of an emotion e is calculated from the average of $\widetilde{D}(\text{label}{=}e, \text{recognised\_emotion})$ of samples whose label is e (see Equation 7). Lower AED means less severe consequences from mistakes, and is therefore more desirable. All experiments are repeated with 10 different seeds.

5.2 ERC Results

Table 2 shows the change in the emotion recognition performance of the selected chit-chat ERC models after incorporating our ERToD framework. ERToD achieves significant improvement in average F1 scores of all models (see Appendix D for examples of model outputs, Appendix E for F1 of individual emotions).

	Base Model		+ ERToD		Difference	
	MF1	WF1	MF1	WF1	MF1	WF1
BERT	50.1	73.5	61.4	77.3	+11.3	+3.8
DialogueRNN+GloVe	40.1	74.6	56.5	78.5	+16.4	+3.9
DialogueRNN+BERT	52.1	75.5	56.5	78.5	+4.4	+3.0
COSMIC	56.3	77.1	57.4	79.6	+1.1	+2.5
EmoBERTa	57.9	**83.0**	**65.9**	**83.9**	+9.0	+0.9
ContextBERT	**59.1**	81.9	**65.9**	**83.9**	+6.8	+2.0

Table 2: Macro- and weighted-average F1 (MF1, WF1) of ERC models before and after incorporating ERToD. Best average F1s are marked in **bold**. All differences are significant with $p < 0.05$.

	Model	Neu.	Sat.	Dis.	Exc.	Apo.	Fea.	Abu.
F1 Score (↑)	ContextBERT	93.5	89.1	69.7	45.6	69.6	33.3	47.0
	+ DA	†**94.2**	†90.5	†71.0	45.3	†72.1	‡38.3	†67.4
	+ DS	†**94.2**	†90.5	†71.3	45.7	†72.7	35.3	†69.4
	+ SentiX	†**94.2**	†**90.6**	†72.2	‡47.1	†73.2	†39.0	†66.1
	+ MTL	†**94.2**	†90.4	†**72.3**	‡47.2	†**73.4**	†41.0	†67.9
	+ ERToD	†94.1	†**90.6**	†**72.3**	†**47.6**	†72.0	†**42.4**	†**69.8**
AED Score (↓)	ContextBERT	0.058	0.094	0.304	0.497	0.269	0.605	0.554
	+ DA	†**0.049**	†0.080	0.312	0.493	‡0.292	0.593	†0.339
	+ DS	†0.053	†0.075	0.296	0.481	0.277	0.582	†0.300
	+ SentiX	†0.052	†0.077	‡0.286	†0.454	0.287	0.596	†0.283
	+ MTL	†0.054	†0.075	‡**0.284**	‡0.456	0.277	0.585	†**0.258**
	+ ERToD	0.056	†**0.070**	0.296	†**0.435**	**0.244**	**0.571**	†0.277

Table 3: F1 (↑) and AED (↓) scores of **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, and **Abu**sive. † indicates statistically significant difference with $p < 0.05$ and ‡ indicates $p < 0.1$ when comparing with ContextBERT. Best scores are marked in **bold**.

5.3 Ablation Study on ERToD

We perform an ablation study on the best resulting model, ContextBERT-ERToD (Table 3). We add each technique in the order of data-related, feature-related, and loss-related approaches. Averaged scores can be found in Appendix F.

Impact of DA DA helps improve almost all F1 scores even with a relatively small number of additional samples. There is a small and insignificant drop in the F1 of *Excited*, which is also frequently confused among human annotators. Further work to resolve the ambiguities would be beneficial.

Impact of Dialogue State (+DS) Adding dialogue state features further improves most other non-neutral emotions. Although it does not bring advantages for the F1 of *Fearful*, the AED of it continues to improve, showing that the system is making less severe mistakes.

Impact of SentiX Initialising BERT with SentiX parameters further improves the recognition of all other non-neutral emotions except for *Abusive*. This suggests that the sentiment information encoded in SentiX is useful for resolving ambiguity. We suspect that, while SentiX is good at distinguishing the valence of emotion, its effect is

limited for user conduct, the hallmark of *Abusive*.

Impact of MTL MTL improves F1 for all non-neutral emotions except for *Satisfied*. It also achieves the best AED for *Abusive*. This suggests that MTL heads, especially the conduct classification head, help identify emotions in the simpler valence-elicitor-conduct space. There is a slight drop in the F1 score of *Satisfied*, but it is compensated by the improvement in its AED.

Impact of *EmoDistLoss* (+ERToD) The final version of the model achieves the best F1 score in {*Satisfied, Dissatisfied, Excited, Fearful, Abusive*} and the best AED score in {*Satisfied, Excited, Apologetic, Fearful*}, leading to best averaged scores (Table F8). This shows penalising misclassifications according to emotion distance, which is only possible thanks to the emotion model, further helps recognise ambiguous emotions.

For the degradation of both scores in *Neutral*, we hypothesise that the model recognises non-neutral emotions more boldly than annotators, who are more cautious about subtle emotional cues.

6 Zero-shot User Satisfaction Prediction

6.1 Experimental Set-up

6.1.1 Dataset

We evaluate our model with **User Satisfaction Simulation** (USS) dataset where user utterances are annotated with 5-level satisfaction ratings (Sun et al., 2021). Dialogues in USS come from 5 different ToD datasets:

Jing Dong Dialogue Corpus (JDDC, Chen et al., 2020) is a multi-turn Chinese dialogue dataset for E-commerce customer service. USS contains 54.5k user satisfaction annotations for 3300 dialouges sampled from JDDC. Since JDDC is in Chinese, we translated it into English with Google Translate API first.

Schema-guided Dialogues (SGD, Rastogi et al., 2020) is a multi-domain, task-oriented conversations between a human and a virtual assistant. These conversations involve interactions with services and APIs spanning 20 domains, such as banks, events, media, calendar, travel, and weather. USS contains 13.8k user satisfaction annotations for 1000 dialogues sampled from SGD. Although we use SGD for DA, our DA samples do not overlap with SGD dialogues in USS.

Recommendation Dialogue (ReDial, Li et al., 2018) is an annotated dataset of dialogues, where users recommend movies to each other. USS contains 11.8k user satisfaction annotations for 1000 dialogues sampled from ReDial.

Coached Conversational Preference Elicitation (CCPE, Radlinski et al., 2019) is a dialogue dataset where the " assistant" is tasked with eliciting the "user" preferences about movies collected in the Wizard-of-Oz framework. USS contains 6.8k user satisfaction annotations for 500 dialogues sampled from CCPE.

MultiWOZ (Budzianowski et al., 2018) is a multi-domain task-oriented dialogue dataset collected in the Wizard-of-Oz framework spanning 7 domains such as restaurant, hotel, and attraction. USS contains 12.5k user satisfaction annotations for 1000 dialogues sampled from MultiWOZ. Since we trained our ERC model on EmoWOZ, which was based on MultiWOZ, we excluded it in our evaluation.

6.1.2 Baselines

We compare our zero-shot results with supervised models of Sun et al. (2021) and Kim and Lipani (2022). HiGRU (Yang et al., 2016) and BERT (Devlin et al., 2019) were the best two models trained by Sun et al. (2021) to benchmark USS dataset when it was first released. SatAct and SatActUtt are T5-based models (Raffel et al., 2020). SatAct is trained to predict user satisfaction and user action in a MTL set-up, whereas SatActUtt additionally incorporates user utterance generation. For satisfaction prediction, these models were set up to predict a 5-level rating during training.

These baseline models were trained on each one of the five ToD subsets in USS with a 10-fold cross-validation. Although non-3 ratings were upsampled by 10 times in their training, the training data size is still smaller than that of ContextBERT-ERToD (68.9k emotion annotations, EmoWOZ and DA samples altogether).

6.1.3 Zero-shot Inference

We experimented with ContextBERT-ERToD, the best resulting model from ERC training. After training the model for ERC, we fixed its parameters and ran inference with USS dataset for zero-shot user satisfaction prediction. To adapt to user satisfaction prediction set-up, we excluded information about the user turn at t from the model

input as well as the dialogue state. Specifically, for utterance encoding, we excluded u_t from the dialogue history to have $H_t = [s_{t-1}, u_{t-1}, ..., s_1, u_1]$. For task information encoding, we shifted the context window in Equation 3 by one and have $\widetilde{V}_t = V_{t-1} \oplus V_{t-2} \oplus V_{t-3}$ as the new contextual dialogue state vector.

6.1.4 Evaluation

In the works of baseline models, satisfaction ratings {1,2} were considered the negative class and {3,4,5} as the positive. To map the emotion prediction from our ERC model to binary satisfaction ratings, it is intuitive to leverage the valence aspect of emotions. Emotion classes with a negative valence were considered *Not Satisfied* and those with a positive valence as *Satisfied*. The emotion *Apologetic* is an exception among emotions with a negative valence. Since its elicitor is the user him/herself, it should not be considered as a sign of user dissatisfaction. Regarding the emotion class *Neutral*, we mapped it to *Satisfied* because the original evaluation set-up of baseline models considered the medium satisfaction rating, 3, as the positive class.

Overall, we considered {*Neutral, Apologetic, Excited, Satisfied*} as the positive class and {*Fearful, Dissatisfied, Abusive*} as negative.

6.2 Results

	JDDC	SGD	ReDial	CCPE
HiGRU (Sun et al., 2021)	17.1	8.6	8.3	27.4
BERT (Sun et al., 2021)	18.5	4.8	12.5	24.5
SatAct (Kim and Lipani, 2022)	-	71.3	-	16.5
SatActUtt (Kim and Lipani, 2022)	-	**84.7**	-	73.4
ContextBERT-ERToD (0-shot)	**50.8**	78.8	**78.1**	**77.6**

Table 4: Binary F1 scores on different USS subsets. Best scores are marked in **bold**.

Following existing work, we first report binary F1 for direct comparison. In Table 4, ContextBERT-ERToD performs comparably with SatActUtt and significantly outperforms other models. This shows that our ERToD framework in combination with the ERC model generalises well to user satisfaction prediction.

7 Conclusion

In this work, we propose ERToD, a framework to address three critical steps in learning and effectively adapt chit-chat ERC models to recognise emotions in ToDs. We propose two strategies of DA for different emotions to improve ERC performance in ToDs on rare emotions. We further leverage dialogue state and sentiment-aware embeddings for a richer feature representation. In addition, we apply MTL and devise a novel loss function, *EmoDistLoss*, which take the similarities between emotions into account. Our framework significantly improves existing chit-chat ERC models' performance in recognising user emotions in ToDs. By further applying our best resulting model to perform the task of user satisfaction prediction, we show that our method generalises well on other similar valence-related classification tasks in ToDs.

As more sophisticated and powerful dialogue systems such as ChatGPT arise, there is an urge to recognise, understand and handle the emotion of the user, especially in the age where online abuse is omnipresent. The long-term aim of this work is to obtain valuable insight for downstream ToD modelling tasks. This allows further investigation of emotion regulation strategies on the system side to improve task performance and user satisfaction, and to prevent undesirable user behaviours.

8 Acknowledgements

S. Feng, N. Lubis, M. Heck, and C. van Niekerk are supported by funding provided by the Alexander von Humboldt Foundation in the framework of the Sofja Kovalevskaja Award endowed by the Federal Ministry of Education and Research, while C. Geishauser, H-C. Lin, B. Ruppik, and R. Vukovic are supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018804636). Computing resources were provided by Google Cloud.

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim (Abe) Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Bill Byrne, Karthik Krishnamoorthi, Chinnadhurai Sankar, Arvind Neelakantan, Ben Goodrich, Daniel Duckworth, Semih Yavuz, Amit Dubey, Kyu-Young Kim, and Andy Cedilnik. 2019. Taskmaster-1: Toward a realistic and diverse dialog dataset. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4516–4525, Hong Kong, China. Association for Computational Linguistics.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. 2021. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 745–754, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hao-Yun Chen, Pei-Hsin Wang, Chun-Hao Liu, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. 2019. Complement objective training. In *International Conference on Learning Representations*.

Meng Chen, Ruixue Liu, Lei Shen, Shaozu Yuan, Jingyan Zhou, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020. The JDDC corpus: A large-scale multi-turn Chinese dialogue dataset for E-commerce customer service. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 459–466, Marseille, France. European Language Resources Association.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *International Conference on Machine Learning*.

Niko Colneric and Janez Demšar. 2020. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 11:433–446.

L. Devillers, L. Lamel, and I. Vasilescu. 2003. Emotion detection in task-oriented spoken dialogues. In *2003 International Conference on Multimedia and Expo. ICME '03. Proceedings (Cat. No.03TH8698)*, volume 3, pages III–549.

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, Adarsh Kumar, Anuj Goyal, Peter Ku, and Dilek Hakkani-Tur. 2020. MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 422–428, Marseille, France. European Language Resources Association.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4096–4113, Marseille, France. European Language Resources Association.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online. Association for Computational Linguistics.

Milan Gritta, Gerasimos Lampouras, and Ignacio Iacobacci. 2021. Conversation graph: Data augmentation, training, and evaluation for non-deterministic dialogue management. *Transactions of the Association for Computational Linguistics*, 9:36–52.

Michael Heck, Nurul Lubis, Carel van Niekerk, Shutong Feng, Christian Geishauser, Hsien-Chin Lin, and Milica Gašić. 2022. Robust Dialogue State Tracking with Weak Supervision and Sparse Data. *Transactions of the Association for Computational Linguistics*, 10:1175–1192.

Le Hou, Chen-Ping Yu, and Dimitris Samaras. 2016. Squared earth mover's distance-based loss for training deep neural networks. *ArXiv*, abs/1611.05916.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1234–1245, Santa Fe, New

Mexico, USA. Association for Computational Linguistics.

Zhiqiang Hu, Roy Kaa-Wei Lee, and Nancy F. Chen. 2022. Are current task-oriented dialogue systems able to satisfy impolite users? *ArXiv*, abs/2210.12942.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Daniel Jurafsky and James H. Martin. 2009. *Speech and Language Processing (2nd Edition)*. Prentice-Hall, Inc., USA.

Nam Kyun Kim, Jiwon Lee, Hun Kyu Ha, Geon Woo Lee, Jung Hyuk Lee, and Hong Kook Kim. 2017. Speech emotion recognition based on multi-task learning using a convolutional neural network. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 704–707.

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-aware emotion recognition in conversation with RoBERTa. *ArXiv*, abs/2108.12009.

To Eun Kim and Aldo Lipani. 2022. A multi-task based neural model to simulate users in goal oriented dialogue systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2115–2119, New York, NY, USA. Association for Computing Machinery.

Varun Kumar, Hadrien Glaude, Cyprien de Lichy, and William Campbell. 2019. A closer look at feature space data augmentation for few-shot intent classification.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature*, 521(7553):436.

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

Yang Li, Amirmohammad Kazameini, Yash Mehta, and Erik Cambria. 2021. Multitask learning for emotion and personality detection. *ArXiv*, abs/2101.02346.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Nurul Lubis, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2015. Construction and analysis of social-affective interaction corpus in english and indonesian. In *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, pages 202–206. IEEE.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An attentive RNN for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2023. Genpads: Reinforcing politeness in an end-to-end dialogue system. *PLOS ONE*, 18(1):1–20.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 20(5).

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

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 (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached conversational preference elicitation: A case study in understanding

movie preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 353–360, Stockholm, Sweden. Association for Computational Linguistics.

Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. 2017. Data distillation: Towards omni-supervised learning. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4119–4128.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *AAAI Conference on Artificial Intelligence*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré . 2017. Snorkel. *Proceedings of the VLDB Endowment*, 11(3):269–282.

Tulika Saha, Sriparna Saha, and Pushpak Bhattacharyya. 2020. Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PLOS ONE*, 15(7):1–28.

Alexander Schmitt, Stefan Ultes, and Wolfgang Minker. 2012. A parameterized and annotated spoken dialog corpus of the CMU let's go bus information system. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3369–3373, Istanbul, Turkey. European Language Resources Association (ELRA).

Weiyan Shi and Zhou Yu. 2018. Sentiment adaptive end-to-end dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1509–1519, Melbourne, Australia. Association for Computational Linguistics.

Gopendra Vikram Singh, Priyanshu Priya, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2022. EmoInHindi: A multi-label emotion and intensity annotated dataset in Hindi for emotion recognition in dialogues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5829–5837, Marseille, France. European Language Resources Association.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Weiwei Sun, Shuo Zhang, Krisztian Balog, Zhaochun Ren, Pengjie Ren, Zhumin Chen, and Maarten de Rijke. 2021. Simulating user satisfaction for the evaluation of task-oriented dialogue systems. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2499–2506, New York, NY, USA. Association for Computing Machinery.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland. Association for Computational Linguistics.

Carel van Niekerk, Andrey Malinin, Christian Geishauser, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. Uncertainty measures in neural belief tracking and the effects on dialogue policy performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7901–7914, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multitask learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9177–9184.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020a. Unsupervised data augmentation for consistency training. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. 2020b. Self-training with noisy student improves imagenet classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Yifei Xu, Jingqiao Zhang, Ru He, Liangzhu Ge, Chao Yang, Cheng Yang, and Ying Nian Wu. 2022. Sas: Self-augmentation strategy for language model pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11586–11594.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Da Yin, Tao Meng, and Kai-Wei Chang. 2020. SentiBERT: A transferable transformer-based architecture for compositional sentiment semantics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3695–3706, Online. Association for Computational Linguistics.

Steve Young. 2002. Talking to machines (statistically speaking). In *Seventh International Conference on Spoken Language Processing*.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. Refining word embeddings for sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark. Association for Computational Linguistics.

Sayyed Zahiri and Jinho D. Choi. 2018. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks. In *Proceedings of the AAAI Workshop on Affective Content Analysis*, AFFCON'18, pages 44–51, New Orleans, LA.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. 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)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Jie Zhou, Junfeng Tian, Rui Wang, Yuanbin Wu, Wenming Xiao, and Liang He. 2020. SentiX: A sentiment-aware pre-trained model for cross-domain sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 568–579, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qi Zhu, Christian Geishauser, Hsien-chin Lin, Carel van Niekerk, Baolin Peng, Zheng Zhang, Michael Heck, Nurul Lubis, Dazhen Wan, Xiaochen Zhu, Jianfeng Gao, Milica Gašić, and Minlie Huang. 2022. Convlab-3: A flexible dialogue system toolkit based on a unified data format.

A Emotion Definitions in EmoWOZ

Elicitor	Valence	Conduct	OCC Emotion Tokens	EmoWOZ Emotion	Implication of User
Operator	Positive	Polite	Admiration, gratitude, love	Satisfied, liking, appreciative	Satisfied with the operator because the goal is fulfilled.
		Impolite		Not applicable to EmoWOZ	
	Negative	Polite	Reproach, anger, hate	Dissatisfied, disliking	Dissatisfied with the operator's suggestion or mistake.
		Impolite		Abusive	Insulting the operator when the goal is not fulfilled.
User	Positive	Polite	Pride, gratification	Not applicable to EmoWOZ	
		Impolite			
	Negative	Polite	Shame, remorse, hate	Apologetic	Apologising for causing confusion to the operator.
		Impolite		Not modelled in EmoWOZ	Insulting the operator for no reason.
Events, facts	Positive	Polite	Happy-for, gloating, love, satisfaction, relief, joy	Excited, happy, anticipating	Looking forward to a good event (e.g. birthday party).
		Impolite		Not applicable to EmoWOZ	
	Negative	Polite	Distress, resentment, hate, fears-confirmed, pity, disappointment	Fearful, sad, disappointed	Encountered a bad event (e.g. robbery and option not available).
		Impolite		Not applicable to EmoWOZ	
-	Neutral	Polite	-	Neutral	Describing situations and needs.
		Impolite		Not modelled in EmoWOZ	No emotion but rude (e.g. using imperative sentences).

Table A1: EmoWOZ labels and similar emotions tokens from the OCC emotion model. For simplicity, emotion words in blue are used to represent each emotion category.

B Examples of Augmented Samples

B.1 Augmentation with Automatic Label

Source: Taskmaster-1
Dialogue ID: dlg-02edb443-9d6f-4553-af6e-f69778eb0fc5

...
S: Any other restaurant you were thinking about?
U: Yes, what about Char's at Tracy Mansion?
S: You mean the one on 1829 N Front St, Harrisburg, PA 17102-2213
U: **Yes, I've heard great things about that [DA Candidate]**

ContextBERT Ensemble Prediction: (Emotion = Excited, Confidence = 100%)

Figure B.1: DA sample for emotion *Excited*.

Source: SGD
Dialogue ID: 93_00124

...
U: I plan to check in on the 7th of this month.
S: Okay. You want to reserve 1 room at 1 Hotel Brooklyn Bridge in New York, for a 1 day stay on March 7th?
U: **Sorry, I misremembered the date. I actually want to check into one room on Saturday this week. [DA Candidate]**

ContextBERT Ensemble Prediction: (Emotion = Apologetic, Confidence = 100%)

Figure B.2: DA sample for emotion *Apologetic*.

Source: SGD Dialogue ID: 1_00028
U: **Hey! Today feeling so tired... Wish to have some food outside in a good restaurant... Can you find me one?** **[DA Candidate]** ...
ContextBERT Ensemble Prediction: (Emotion = Fearful, Confidence = 90%), (Emotion = Excited, Confidence = 10%)

Figure B.3: DA sample for emotion *Fearful*. Please note that although this class is called "fearful" for simplicity, user's negative emotion due to any undesirable events that is out of the control of the operator also belongs to this category in EmoWOZ according to Table A1.

B.2 Augmentation with Existing Dataset and Utterance Replacement

Source: ConvAbuse
DA Candidates:
1. Are you an idiot 2. You are stupid privitive word-repeater 3. Why do you not answer my question 4. Foolish ..
EmoWOZ sample for *Abusive* Dialogue ID: DMAGE2157.json
... U: need a good resturant near macy's in nyc S: You are more than welcome! U: can you help S: The address is Parkside, Cambridge . The phone is 01223358966 . The postcode is none . Thank you for using our services . U: **your horrible and not helpful at all** [Label = Abusive]

Figure B.4: DA sample for emotion *Abusive*. Candidate DA samples from ConvAbuse can be used to replace the abusive user utterance in EmoWOZ without causing any context inconsistency.

C Emotional Aspect Distance Definition

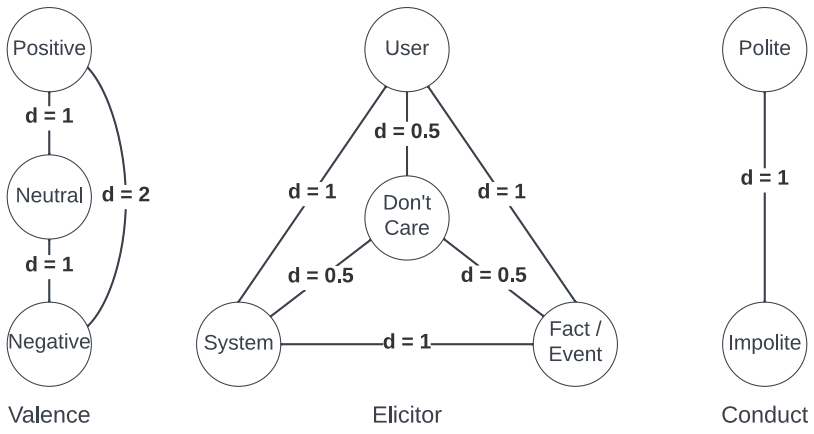

Figure C.1: Distance definition for Equation 6.

D Examples of Model Recognitions

Dialogue ID: DMAGE3777.json

...
U: Can you help me?
S: The phone is 01223358966. The postcode is none.
U: Ok
S: The address is Parkside, Cambridge.
U: **Nice** **[To classify, label = Dissatisfied]**

	BERT	DialRNN-G	DialRNN-B	COSMIC	EmoBERTa	ContextBERT
Base	Satisfied	Dissatisfied	Satisfied	Satisfied	Satisfied	Neutral
+ ERToD	Neutral	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied

Figure D.1: Model Recognitions on dialogue DMAGE3777 in EmoWOZ.

Dialogue ID: PMUL2437.json

...
S: There are 21 restaurants available in the centre of town. How about a specific type of cuisine?
U: I need to know the food type and postcode and it should also have multiple sports
S: I am sorry I do not understand what you just said. Please repeat in a way that makes sense.
U: **Get me the food type and the post code** **[To classify, label=Dissatisfied]**

	BERT	DialRNN-G	DialRNN-B	COSMIC	EmoBERTa	ContextBERT
Base	Neutral	Dissatisfied	Neutral	Neutral	Dissatisfied	Neutral
+ ERToD	Neutral	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied	Dissatisfied

Figure D.2: Model Recognitions on dialogue PMUL2437 in EmoWOZ

E Detailed ERC Performance on Each Emotion

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	89.8	88.8	35.1	42.9	70.4	36.2	27.5
DialogueRNN+GloVe	83.5	86.4	51.4	32.7	57.7	12.7	0.0
DialogueRNN+BERT	86.9	87.6	47.5	39.4	**71.5**	41.3	25.6
COSMIC	89.8	88.4	50.7	44.4	70.9	**52.0**	31.6
EmoBERTa	**94.0**	**90.3**	**71.0**	44.9	70.6	31.3	39.3
ContextBERT	93.5	89.1	69.7	**45.6**	69.6	33.3	**47.0**

Table E2: F1 scores of selected chit-chat ERC models BEFORE incorporating ERToD framework. The best score for each emotion is marked in **bold**.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	90.3	89.3	88.4	89.2	38.9	38.6	**47.7**	39.1	69.7	71.5	47.7	30.0	42.1	22.4	55.7	48.5	74.5	74.5
DialRNN-GloVe	**97.6**	73.0	78.5	**95.9**	36.5	**87.6**	22.2	**65.7**	44.7	**82.5**	11.2	18.9	0	0	32.2	**58.4**	65.0	**91.4**
DialRNN-BERT	94.0	80.7	84.7	90.7	34.8	75.3	36.5	42.9	68.3	75.0	46.7	37.5	28.6	23.5	49.9	57.5	70.4	84.2
COSMIC	93.1	86.8	86.2	90.7	42.3	64.4	43.7	45.3	**71.9**	70.1	**65.0**	**43.3**	**77.3**	20.0	**64.4**	55.6	74.0	81.7
EmoBERTa	94.2	**94.0**	**88.7**	92.2	**74.6**	69.5	45.6	42.6	73.0	70.3	37.9	27.2	54.0	24.7	62.3	54.4	**82.9**	83.8
ContextBERT	93.4	93.7	88.5	89.8	72.6	67.2	46.4	45.4	68.3	71.6	37.9	30.0	64.5	**37.6**	63.0	57.0	82.3	81.8

Table E3: **P**recision and **R**ecall scores of selected chit-chat ERC models BEFORE incorporating ERToD framework. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	92.4	90.4	43.7	**49.7**	75.4	39.5	**69.7**
DialogueRNN+GloVe	92.6	90.1	51.4	43.9	**77.6**	42.4	33.8
DialogueRNN+BERT	92.6	90.1	51.4	43.9	**77.6**	42.4	33.8
COSMIC	91.1	89.5	58.1	45.6	73.3	36.3	41.6
EmoBERTa	**94.0**	**90.5**	**72.3**	47.9	71.9	**43.4**	**69.7**
ContextBERT	**94.0**	**90.5**	**72.3**	47.9	71.9	**43.4**	**69.7**

Table E4: F1 scores of selected chit-chat ERC models AFTER incorporating ERToD framework. The best score for each emotion is marked in **bold**.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	91.0	93.8	88.9	92.0	57.5	35.5	**51.2**	48.9	**81.6**	70.3	48.1	33.9	**74.8**	65.9	67.0	57.7	79.8	76.3
DialRNN-GloVe	91.3	**94.0**	**89.7**	90.5	60.9	41.5	44.4	45.6	76.5	77.3	42.6	**38.3**	54.3	30.0	61.4	53.9	80.6	76.5
DialRNN-BERT	91.3	**94.0**	**89.7**	90.5	60.9	41.5	44.4	45.6	76.5	77.3	42.6	**38.3**	54.3	30.0	61.4	53.9	80.6	76.5
COSMIC	**94.4**	88.3	86.9	92.3	51.6	**68.9**	38.7	**57.4**	68.2	**79.3**	36.2	**38.3**	44.7	38.8	54.4	62.5	75.9	**84.6**
EmoBERTa	94.3	93.9	88.9	**92.4**	**75.6**	68.0	45.7	50.7	70.8	74.4	**54.6**	35.6	72.4	**68.2**	**68.0**	**64.9**	**83.5**	84.3
ContextBERT	94.3	93.9	88.9	**92.4**	**75.6**	68.0	45.7	50.7	70.8	74.4	**54.6**	35.6	72.4	**68.2**	**68.0**	**64.9**	**83.5**	84.3

Table E5: **P**recision and **R**ecall scores of selected chit-chat ERC models AFTER incorporating ERToD framework. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

Model	Neutral	Satisfied	Dissatisfied	Excited	Apologetic	Fearful	Abusive
BERT	+2.6	+1.6	+8.6	+6.8	+5.0	+3.3	+42.2
DialogueRNN+GloVe	+9.1	+3.7	+0.0	+11.2	+19.9	+29.7	+33.8
DialogueRNN+BERT	+5.7	+2.5	+3.9	+4.5	+6.1	+1.1	+8.2
COSMIC	+1.3	+1.1	+7.4	+1.2	+2.4	**-15.7**	+10.0
EmoBERTa	0.0	+0.2	+1.3	+3.0	+1.3	+12.1	+30.4
ContextBERT	+0.5	+1.4	+2.6	+2.3	+2.3	+10.1	+22.7

Table E6: Change of F1 scores of selected chit-chat ERC models after incorporating ERToD framework. The only degradation in performance is marked in **bold**.

In terms of F1 scores, ERToD results in improvement in all emotions except for *fearful* in COSMIC (Table E6). We further investigate this exception. While most of fearful utterances are located at the beginning

of the dialogue in the training and development set in EmoWOZ, the position of such utterances are more evenly distributed in the test set as well as the augmented samples. Upon toggling the development set and the test set for evaluation, we observe that the F1 of fearful by COSMIC drops significantly (52.0% $\rightarrow$ 28.8%) while that of COSMIC-ERToD remains roughly unchanged (35.5% $\rightarrow$ 37.6%). The trend in all other results remains unchanged.

The drastically different performance of COSMIC on the development and the test set suggests that COSMIC develops a positional bias from the training set of EmoWOZ. At the same time, COSMIC-ERToD performs similarly on both non-training sets, likely relying more on textual and task information. The limited performance of COSMIC-ERToD is likely due to the extra false-positives at the later stage of dialogues.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
BERT	+0.7	+4.5	+0.5	+2.8	+18.6	-3.1	+3.5	+9.8	+11.9	-1.2	+0.4	+3.9	+32.7	+43.5	+11.3	+9.2	+5.3	+1.8
DialRNN-GloVe	-6.3	+21.0	+11.2	-5.4	+24.4	-46.1	+22.2	-20.1	+31.8	-5.2	+31.4	+19.4	+54.3	+30.0	+29.2	-4.5	+15.6	-14.9
DialRNN-BERT	-2.7	+13.3	+5.0	-0.2	+26.1	-33.8	+7.9	+2.7	+8.2	+2.3	-4.1	+0.8	+25.7	+6.5	+11.5	-3.6	+10.2	-7.7
COSMIC	+1.3	+1.5	+0.7	+1.6	+9.3	+4.5	-5.0	+12.1	-3.7	+9.2	-28.8	-5.0	-32.6	+18.8	-10.0	+6.9	+1.9	+2.9
EmoBERTa	+0.1	-0.1	+0.2	+0.2	+1.0	-1.5	+0.1	+8.1	-2.2	+4.1	+16.7	+8.4	+18.4	+43.5	+5.7	+10.5	+0.6	+0.5
ContextBERT	+0.9	+0.2	+0.4	+2.6	+3.0	+0.8	-0.7	+5.3	+2.5	+2.8	+16.7	+5.6	+7.9	+30.6	+5.0	+7.9	+1.2	+2.5

Table E7: The difference in **P**recision and **R**ecall scores of selected chit-chat ERC models before and after incorporating ERToD framework. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. Neutral is excluded when calculating the averaged scores. For better presentation, DialogueRNN is shortened to DialRNN.

F Averaged Scores for the Ablation Study

	Model	Macro Avg	Weighted Avg
F1 Score (↑)	ContextBERT	59.1	81.9
	+ DA	†64.1	†83.4
	+ DS	†64.1	†83.5
	+ SentiX	†64.8	†83.7
	+ MTL	†65.3	†83.7
	+ ERToD	†**65.7**	†**83.9**
AED Score (↓)	ContextBERT	0.387	0.168
	+ DA	†0.351	†0.159
	+ DS	†0.335	†0.151
	+ SentiX	†0.331	†0.149
	+ MTL	†0.322	†0.147
	+ ERToD	†**0.316**	†**0.145**

Table F8: Ablation Study of ERToD. † indicates statistically significant difference with $p < 0.05$ when comparing with ContextBERT. The best score in each category is in **bold**. For each of the additional methods: DA = Data Augmentation, DS = Dialogue State Features, SentiX = Sentiment-aware Text Embedding, MTL = Multi-task Learning. Neutral is excluded when calculating the averaged scores.

	Neu.		Sat.		Dis.		Exc.		Apo.		Fea.		Abu.		M-Avg		W-Avg	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R
ContextBERT	93.4	93.9	88.5	**92.4**	72.6	68.0	46.4	50.7	68.3	74.4	37.9	**35.6**	64.5	68.2	63.0	64.9	82.3	84.3
+ DA	93.9	94.4	89.4	91.6	75.6	67.2	47.2	44.6	75.0	70.0	53.1	30.6	70.1	65.9	**68.4**	61.6	84.0	83.1
+ DS	93.8	**94.6**	**90.1**	90.9	74.5	68.4	**47.9**	44.6	75.8	69.0	50.7	27.8	69.9	69.4	68.1	61.7	**84.2**	82.9
+ SentiX	94.1	94.3	89.5	91.7	76.0	69.1	47.5	49.3	**76.7**	70.3	50.9	32.2	66.0	66.5	67.8	63.2	84.1	83.9
+ MTL	94.2	94.0	88.9	91.5	**76.4**	**70.6**	45.7	**49.8**	76.6	**71.6**	51.2	35.0	67.0	**72.4**	67.6	**65.1**	83.8	**84.2**
+ ERToD	**94.3**	94.1	88.9	91.9	75.6	69.3	45.7	48.8	70.8	70.8	**54.6**	34.4	**72.4**	70.0	68.0	64.2	83.5	84.1

Table F9: Ablation study on **P**recision and **R**ecall scores of ERToD. We report scores of each emotion: **Neu**tral, **Sat**isfied, **Dis**satisfied, **Exc**ited, **Apo**logetic, **Fea**rful, **Abu**sive, as well as **M**acro- and **W**eighted **Av**era**g**ed scores. The best score for each emotion is marked in **bold**. For each of the additional methods: DA = Data Augmentation, DS = DialogueState Features, SentiX = Sentiment-aware Text Embedding, MTL = Multi-task Learning.. Neutral is excluded when calculating averaged scores.