EmoWOZ: A Large-Scale Corpus and Labelling Scheme for Emotion Recognition in Task-Oriented Dialogue Systems

Anonymous ACL submission

Abstract

The ability to recognise emotions lends a conversational artificial intelligence a human touch. While emotions in chit-chat dialogues have received substantial attention, emotions in task-oriented dialogues have been largely overlooked despite having an equally important role, such as to signal failure or success. Existing emotion-annotated task-oriented corpora are limited in size, label richness, and public availability, creating a bottleneck for downstream tasks. To lay a foundation for studies on emotions in task-oriented dialogues, we introduce EmoWOZ, a large-scale manually emotion-annotated corpus of task-oriented dialogues. EmoWOZ is based on MultiWOZ, a multi-domain task-oriented dialogue dataset. It contains more than 11K dialogues with more than 83K emotion annotations of user utterances. In addition to Wizard-of-Oz dialogues from MultiWOZ, we collect human-machine dialogues within the same set of domains to sufficiently cover the space of various emotions that can happen during the lifetime of a data-driven dialogue system. To the best of our knowledge, this is the first large-scale open-source corpus of its kind. We propose a novel emotion labelling scheme, which is tailored to task-oriented dialogues. We report a set of experimental results to show the usability of this corpus for emotion recognition and state tracking in task-oriented dialogues.

1 Introduction

Incorporating human intelligence into conversational artificial intelligence (AI) has been a challenging and long-term goal (Picard, 1997). Emotional intelligence, defined as the ability to regulate, perceive, assimilate, and express emotions, is a key component of general intelligence (Mayer et al., 1999). Such emotion awareness can help the conversational AI generate more emotionally and semantically appropriate responses (Zhou et al., 2017).

Dialogue systems generally fall into two classes. Task-oriented systems converse with users to help complete tasks. Chit-chat systems are set up to mimic the unstructured conversations or 'chats' characteristic of human-human interaction (Jurafsky and Martin, 2009). Chat-oriented systems are typically modelled in a supervised fashion with large available corpora (Vinyals and Le, 2015). In contrast, task-oriented systems track the user goal throughout the dialogue and a policy is typically trained via some form of reinforcement learning to conduct dialogue towards successful goal completion (Young, 2002). Moreover, the scope of the dialogue can also be extended during this process, e.g. by adding new domains to the dialogue system (Madotto et al., 2020) . Consequently, the distribution of data from which a task-oriented system learns can change.

Task-oriented dialogues and chit-chat dialogues contain different nuances of emotion due to emotions having inherently different roles. Chit-chat dialogues is a means to express emotion. Speakers may discuss emotional experiences (Li et al., 2017), or topics that induce emotions such as news broadcasts (Lubis et al., 2017). In task-oriented dialogues, emotion is centred around the user goal, making it more contextual and subtle. Therefore, besides inferring emotional states from dialogue utterances, an agent also needs to reason about emotion-generating situations (Poria et al., 2021).

Substantial research efforts in emotion recognition in conversations (ERC) have been invested in chit-chat dialogues. There are several public ERC corpora containing chit-chat dialogues (Li et al., 2017; Poria et al., 2018; Zahiri and Choi, 2017) and dialogue-like data (Zhou and Wang, 2017). These corpora can tremendously accelerate the building of emotional chatbots using data-driven approaches (Zhou et al., 2017). In task-oriented dialogues, emotions are equally important but have been largely overlooked. Existing corpora are small in size, and labels are limited to sentiment polarity. This creates a bottleneck for downstream tasks.

In this work, we present **EmoWOZ**, a large-scale manually labelled corpus for emotion in task-oriented dialogues. EmoWOZ is derived from MultiWOZ (Budzianowski et al., 2018), one of the largest multi-domain corpora and the benchmark dataset for various dialogue modelling tasks, from dialogue state tracking (Heck et al., 2020b) to policy optimisation (Zhao et al., 2019). We also collected and annotated human-machine dialogues as a complement. Our contributions are as follows:

- We construct a corpus containing task-oriented dialogues with emotion labels, comprising more than 11K dialogues and 83K annotated user utterances. To the best of our knowledge, this is the first large-scale open-source corpus & code for emotion recognition in task-oriented dialogues.
- We propose a novel labelling scheme, containing 7 emotion classes, adapted from the Ortony, Clore and Collins (OCC) model (Ortony et al., 1988), specifically tailored to capture a spectrum of emotions in relation to user goals in task-oriented dialogue.
- We report a series of emotion recognition baseline results to show the usability of this corpus. We also empirically show that the emotion labels can be used to improve the performance of other task-oriented dialogue system modules, in this case, a dialogue state tracker (DST).

2 Related Work

2.1 Emotion Models

Within the area of affective computing, emotion models are commonly grouped into two types: dimensional models and categorical models.

Dimensional models describe emotions as a combination of values across a set of dimensions. The longest established dimensions are valence and arousal, as proposed by Russell (1980) in the circumplex model of emotion. Valence measures the positivity, while arousal measures the activation. Happiness, for example, is an emotion with positive valence and high activation. Additional dimensions, namely dominance and expectancy (Fontaine et al., 2007), have also been proposed to further describe and distinguish complex emotions.

Categorical models group emotions into distinct categories. The "Big six" theory is one of the most well-known theories on universal emotions. Based on studies of facial expressions, Ekman (1992) proposed six basic human emotions which are influenced neither by culture nor other social influences: happiness, anger, sadness, disgust, fear, surprise. Parrott (2001) conceptualised over a hundred emotions into a tree-structured list and identified six primary emotions from it.

Ortony et al. (1988) proposed the Ortony, Clore and Collins (OCC) emotion model, which is explicitly developed for implementation in computers. In the OCC model, 22 emotion types are described as a valenced reaction to one of three cognitive elicitors: consequences of events, actions of agents, or aspects of objects. For example, *dissatisfied* is specified as disapproving of someone else's blameworthy action. These cognitive aspects are in line with the cognitive process of a computational agent, making the OCC model suitable for building emotional artificial agents. However, the use of this model for dialogue agents is not yet wide-spread.

Although there are corpora with real-valued annotation of multiple emotion dimensions (Preoţiuc-Pietro et al., 2016; Buechel and Hahn, 2017), researchers often focus on the valence dimension and annotate with discrete classes (Socher et al., 2013), often called sentiment polarity. Emotion datasets also consider emotions from various categorical models in the annotation scheme (Li et al., 2017; Poria et al., 2018), but some datasets create a unique set of domain-specific labels. For instance, Zhou and Wang (2017) leverage common emojis in social media posts. The Topical-Chat dataset (Gopalakrishnan et al., 2019) introduces *curious to dive deeper* in addition to other basic emotions.

As most corpora are not annotated by experts (see Table 1), emotion labels from everyday vocabulary provide more accessibility to crowd-source the annotation task. In this work, we propose to adapt the OCC model and map it into a novel set of 7 emotions. We aim for this scheme to capture the cognitive context of emotions while retaining the simplicity of labels that facilitate large-scale crowd-sourcing of emotion annotations.

2.2 Emotion Dialogue Datasets

Most existing ERC datasets focus on chit-chat dialogue. Chit-chat dialogue lends itself well to affective computing research due to its open-domain set-up, where conversations are often rich in emotion. One of the largest such corpora is DailyDialog (Li et al., 2017), which contains conversations between English learners on various topics

Metric	DailyDialog	MELD	EmoryNLP	DSTC1	SentiVA	TML	EmoWOZ(Ours)
Dialogue type		Chit-chat			Task-oriented		
# Dialogues	**13,118**	1,433	897	50	1,282	3,496	11,434
Total # turns	102,979	13,708	12,606	517	35,267	68,216	**167,260**
# Unique tokens	26,364	8052	8441	199	-	-	**28,417**
Avg. turns / dialogue	7.9	9.6	14.1	10.3	**27.5**	19.5	14.63
Avg. tokens / turn	**14.6**	10.4	14.3	2.3	-	-	12.78
Label type	Emo	**Sent, Emo**	**Sent, Emo**	Sent	Sent	Sent	**Sent, Emo**
# Classes	7	**3 and 7**	**3 and 7**	3	3	5	**3 and 7**
# Annotations	**102,879**	13,708	12,606	517	35,267	68,216	83,630
# Annotators / turn	3	3	**4**	-	3	2	3
Expert Annotator?	Yes	No	No	-	No	No	No
Agreement	0.789	0.43	0.14	-	**0.8**	0.79	0.602
Open-sourced?	**Yes**	**Yes**	**Yes**	**Yes**	No	No	**Yes**

Table 1: Comparison of our corpus to similar corpora. Values in bold indicate the best value for each metric. For label type, "Emo" stands for emotion categories and "Sent" stands for sentiment polarities. DSTC1, SentiVA, and TML refer to works by Shi and Yu (2018), Saha et al. (2020), and Wang et al. (2020), respectively.

ranging from relationships to money. Other similar datasets include EmoryNLP (Zahiri and Choi, 2017) and MELD (Poria et al., 2018). They contain multi-party dialogues from the TV show *Friends*. TV recordings in talk show format have also been utilised to collect emotion-rich and topic-specific dialogues (Lubis et al., 2015). Unfortunately, existing data suitable for task-oriented corpora, such as customer service chat logs, are typically not within the public domain.

There also exist a few corpora concerning the affective aspect of task-oriented dialogues. Wang et al. (2020) proposed a large-scale sentiment classification corpus containing customer service dialogues in Chinese. However, this dataset is not publicly available. Saha et al. (2020) annotated dialogues from bAbI (Bordes and Weston, 2016) with sentiment for policy optimisation. These dialogues are machine-generated, which may not match real human emotions well. In a similar spirit, Shi and Yu (2018) annotated the DSTC1 dataset with user sentiment. Unfortunately, containing only 50 dialogues, the dataset is very limited in terms of coverage. To summarise, existing corpora are either limited in size or not publicly available, limiting further works on emotions in task-oriented dialogue systems. Furthermore, sentiment annotations overlook the effect of goals on users' emotional states and may not sufficiently capture emotional nuances in task-oriented dialogues.

3 Dataset Construction

3.1 Task-oriented Dialogues

MultiWOZ: Our dataset covers the entirety of MultiWOZ, which was constructed using the Wizard-of-Oz framework (Kelley, 1984). Each dialogue was completed by two workers, each acting as the user or the operator, to achieve specified goals such as information retrieval or making reservations. There are 7 domains in total. A single dialogue or even a single turn can span multiple domains.

Complementary Dialogues: Most dialogues in MultiWOZ are successful, potentially creating a biased emotion coverage towards positive emotions. However, it is necessary for EmoWOZ to cover a variety of dialogues, since, during the life span of a data-driven task-oriented dialogue system, the distribution of emotions may change. We also envisage emotions be used as learning signal for dialogue system optimisation. It is thus crucial for emotion estimators to learn from both failed and successful dialogues. To cover negative emotions in failed dialogues, we complement MultiWOZ with human-machine **dial**ogues from a **s**ub-**o**ptimal **p**olicy (DialSoP). Instead of instructing human wizards to make machine-like mistakes, we let subjects directly interact with a sub-optimal policy, which, we believe, elicits more genuine reactions.

We trained a policy in a supervised fashion on MultiWOZ and achieved a task success rate of 55% when evaluated with the ConvLab-2 (Zhu et al., 2020) rule-based user simulator. Similar to Li et al. (2020), the policy uses a recurrent neural network (RNN) based model to produce multiple actions in a single turn, followed by the ConvLab-2 template-based NLG module for response generation. We launched a dialogue interactive task on Amazon Mechanical Turk, where workers are asked to retrieve information by interacting with the sub-optimal policy. Workers are not told to purposely express anger to the system. To obtain more diverse conversations, user feedback is used to further train the policy using RL. However, the policy

Elicitor	Valence	Conduct	OCC Emotion	Our 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 the dataset	
	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 the dataset	
		Impolite			
	Negative	Polite	Shame, remorse, hate	Apologetic	Apologising for causing confusion to the operator.
		Impolite		Not modelled	Insulting the operator for no reason.
Events, facts	Positive	Polite	Happy-for, gloating, love,	Excited, happy, anticipating	Looking forward to a good event (e.g. birthday party).
		Impolite	satisfaction, relief, joy	Not applicable to the dataset	
	Negative	Polite	Distress, resentment, hate, fears-	Fearful, sad, disappointed	Encountered a bad event (e.g. robbery).
		Impolite	confirmed, pity, disappointment	Not applicable to the dataset	
NA	Neutral	Polite	NA	Neutral	Describing situations and needs.
		Impolite		Not modelled	No emotion but rude (e.g. using imperative sentences).

Table 2: Comparison between the OCC model and our labelling scheme. Emotions that do not occur in our dataset are marked as "not applicable to our dataset". {User, negative, impolite} has too few instances and {neutral, impolite} is not strong enough to be considered as *abusive*. They are therefore not modelled for now. For simplicity, the emotion word in blue is used to represent each emotion category. The OCC model is illustrated in Appendix A.

remains sub-optimal throughout the data collection, reaching a final human-rated success rate of 73%.

3.2 Emotion Annotation Scheme

EmoWOZ focuses on user emotions rather than system ones. We believe recognising user emotions is the starting point for building emotion-aware task-oriented dialogue systems. We use the OCC model to arrive at specific emotion categories. For that, we consider the following aspects:

1. Elicitor or cause: The OCC model defines three main elicitors of emotion: events, agents, and objects. In task-oriented dialogues, events describe the situation which brings the user to interact with the system. For example, a user may be looking for a hotel for an upcoming trip or asking for the police information after a robbery. Agents are participants of the dialogue: the user and the system. Objects are equal to entities being talked about in the dialogue, such as the recommended hotel or the nearest police station. In our dataset, an object is always associated with either the operator, who proposes it, or an event, which drives the need for it. For this reason, we do not consider the object as an elicitor alone. On the other hand, within the agent category, it is important to distinguish between the user and the system. Therefore, we arrive at three elicitors for our annotation scheme: 1) the system, 2) the user, and 3) events (or facts).

2. Valence: In essence, the OCC model describes emotion as a valenced reaction towards an elicitor. Valence is a dimension which expresses the positivity or negativity of emotion. For example, successfully achieving a goal is likely to bring positive valence, while a misunderstanding with an agent is likely to cause negative valence. As EmoWOZ will demonstrate in a later section, valence is highly related to task success or failure, making it an important signal for a task-oriented system. We distinguish neutral and emotional utterances, and further separate emotional utterances into those with negative and positive valence.

3. Conduct: Conduct is not a part of the OCC model, but given the rising concern of how humans behave when interacting with virtual assistants (Cercas Curry and Rieser, 2018), we decide to include it. Conduct describes the politeness of users and is usually associated with emotional acts. Politeness can indicate the degree of valence. For example, the user can express very strong dissatisfaction through rudeness. It also helps distinguish emotions such as those associated with apology or abuse, which are both intrinsically negative.

Considering all combinations of these three aspects for annotation leads to a large number of classes. When choosing the final set of classes we were guided by whether or not a particular emotion category occurs in the database and the potential impact of that emotion category on the dialogue policy. We also carried out several trials and considered the ease of communicating to the annotator how to label such instances. We finally arrive at a set of 6 non-neutral emotion categories:

An emotion elicited by the operator is defined as *satisfied* if it is positive, and *dissatisfied* if it is negative. Positive emotion caused by an event gives us *excited*, and negative *fearful*. In terms of negative emotions expressed towards the system, we consider user conduct to distinguish between *dissatisfied* and *abusive*, since they require very different responses from the system (Curry and Rieser, 2019). In terms of the negative emotions that users may direct toward themselves, we single out *apologetic* behaviours since it features in human-human information-seeking dialogues. Emotion categories and their attributes in the above-mentioned aspects and their relation to the original OCC model are shown in Table 2.

3.3 Emotion Annotation Setup

We crowd-source the emotion annotation on Amazon Mechanical Turk in a controlled manner. Workers are shown the dialogue history up to the utterance they are required to label. Each emotion category is followed by a list of emotion words that best fit into the category and an explanation. Each dialogue is annotated by three different workers. We also implement several measures to ensure the quality of the emotion labels:

Qualification tests: The test contains fifteen questions, seven are straight-forward and eight are more complex. The test also serves as a tutorial. For difficult questions, hints are provided to guide the workers to identify implicit emotions and use contextual information (see Appendix B).

Hidden tests: We pre-label more than 1000 utterances containing obvious emotions and use them as sanity checks. The hidden tests serve as an indicator of worker reliability. If a worker scores above 80% on the hidden tests, we assume that the worker is reliable. Otherwise, the workers' submission is subject to manual review.

Review for outliers: We use a simple lexicon-based recogniser and manually annotate a small batch to have an estimate of the overall emotion distribution. If the label distribution in a worker's submissions deviates substantially from our prior belief, we mark them for manual review.

Annotation limit: We limit each worker to annotate at most 500 dialogues to ensure a diversity of workers and to avoid that workers adapt to our approval policy. Overall, we had 215 workers, each annotating 160 dialogues on average.

4 EmoWOZ Characteristics

4.1 Linguistic Style

Dialogues from MultiWOZ and DialSoP differ linguistically. As seen in Table 3, DialSoP has longer dialogues than MultiWOZ as it takes longer for the sub-optimal policy to accomplish user goals. Meanwhile, users use simpler and shorter sentences when talking to a machine. Poor system performance and its unnaturalness discourage users to converse with it (see sample dialogues with annotations in Appendix C). We will analyse the impact of these differences on emotion recognition in Section 5.1.3.

	MultiWOZ	**DialSoP**	**EmoWOZ**
# Dialogues	10,438	996	11,438
# Unique tokens	27,833	3,133	28,417
Avg. turns / dialogue	13.7	24.3	14.6
Avg. tokens / user turn	11.6	5.7	10.6
Avg. unique user tokens / dialogue	57.8	36.5	55.6

Table 3: Comparison of linguistic features in EmoWOZ.

4.2 Emotion Distribution

Emotion	**EmoWOZ**		**MultiWOZ**		**DialSoP**	
	Count	Prop.	Count	Prop.	Count	Prop.
Neutral	58,678	70.2%	51,417	71.9%	7,261	60.0%
Fearful	404	0.5%	385	0.5%	19	0.2%
Dissatisfied	5,053	6.0%	909	1.3%	4,144	34.2%
Apologetic	843	1.0%	838	1.2%	5	0.04%
Abusive	134	0.2%	46	0.1%	88	0.7%
Excited	991	1.2%	876	1.2%	115	1.0%
Satisfied	17,527	21.0%	17,053	23.8%	474	3.9%

Table 4: Count and prop(ortion) of emotion labels.

According to Table 4, the most common non-neutral emotion in EmoWOZ is *satisfied*, followed by *dissatisfied*. This is expected in task-oriented dialogues as users mainly express emotion in relation to their goals. While MultiWOZ contains more neutral utterances, it has a more diverse emotion distribution than DialSoP. MultiWOZ contributes most *satisfied* utterances whereas DialSoP contributes most *dissatisfied* utterances. This is in line with their respective dialogue-generating setup.

Sometimes users also express emotion to engage or provoke the operator. MultiWOZ contains more *apologetic* and less *abusive* utterances than DialSoP, suggesting that users tend to be more polite when talking to human operators. Dialogues from MultiWOZ also contain more event-elicited emotions than DialSoP. Users are more talkative when conversing with human operators. Users may describe a miserable situation they were experiencing, hoping to be helped and comforted. A human operator would naturally show empathy. In MultiWOZ, the operator sometimes asks if the user is alright when the user is looking for help from a robbery. When talking to machines, users tend not to express such chit-chat-style emotions due to the expected incapability of the machine to reciprocate. This indicates that an emotionally intelligent agent will allow dialogues that are emotionally richer and more nuanced, even in a task-oriented setting.

4.3 Inter-annotator Agreement

We measure the inter-annotator agreement by computing the Fleiss' Kappa (Fleiss, 1971). The Fleiss' Kappa for EmoWOZ is 0.602, suggesting a substantial agreement. The Fleiss' Kappa for MultiWOZ is 0.611, higher than 0.465 for DialSoP. Emotions in DialSoP are more challenging to annotate because users express emotion less explicitly when they know the system does not react to emotions. Annotators often have to infer the user's implicit emotions from dialogue history, for example, based on repetitions or misunderstanding.

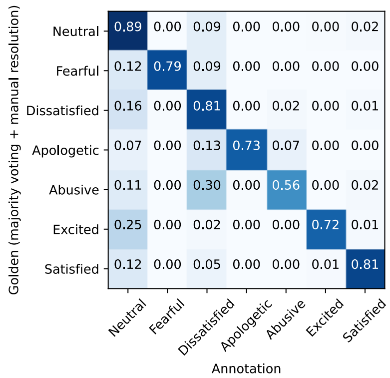

Figure 1: Confusion matrix of emotion annotations.

Among all utterances, 72.1% see a full agreement among three annotators, 26.4% see a partial agreement, and 1.5% see no agreement. The count of each case in each subset can be found in Appendix D. Utterances for which no agreement is reached are resolved manually.

Figure 1 illustrates the confusion matrix between annotators' labels and the golden labels. Most disagreements occur between non-neutral emotions and neutral, as well as *abusive* and *dissatisfied*. This is reasonable as workers adopt different valence or impoliteness thresholds when they make decisions. There is also confusion among emotions with the same polarity but different causes. This suggests that workers may have different interpretations of the emotion elicitor. For example, a user may express sadness after the agent informed that there is no attraction meeting the user's criteria. While the emotion is caused by the fact that there is no match, one can also argue that the operator failed to suggest alternative options.

5 Experiment

5.1 Emotion Recognition in Dialogue

Emotion recognition aims to recognise emotion within an utterance. Unlike utterances in isolation, emotion recognition in dialogues is highly contextual with respect to the dialogue history. To take dialogue context into account, RNN models or transformer models are typically used. As baselines, we compare two models originally developed for chit-chat emotion recognition as well as a couple of BERT-based models.

5.1.1 Baselines

BERT (Devlin et al., 2018): BERT is used as the utterance encoder. Each user turn is encoded in isolation without any dialogue context. The `[CLS]` token from a `bert-base-cased` model is used as the feature representation, which is then fed into a linear output layer for classification.

ContextBERT: The set-up is identical to that of BERT, except that the entire dialogue history and the current user utterance are concatenated to form one long sequence. We add "`User:`" and "`System:`" to mark the speaker of each turn.

DialogueRNN (Majumder et al., 2018): The model combines gated recurrent units (GRUs) with an attention mechanism to capture the long-term trajectory of the dialogue. We experiment with using GloVe embeddings (Pennington et al., 2014) or the `[CLS]` representation from BERT as input features. When GloVe is used, a convolutional neural network (CNN) layer is used as a feature extractor to generate utterance representations. This CNN layer is dropped when using BERT features.

COSMIC (Ghosal et al., 2020): This model also combines GRUs with the attention mechanism. In addition to utterance representations from a pre-trained language model (LM), it supplements input features with common-sense knowledge extracted from a pre-trained commonsense transformer model called COMET (Bosselut et al., 2019). Although the original paper uses RoBERTa as input features, we found that BERT results in a better sequence representation for emotion recognition on our data. Therefore we use BERT as the utterance encoder in our experiments.

5.1.2 Experimental Setup

We perform a recognition task on the 7 emotions proposed in our annotation scheme. All models are implemented in PyTorch (Paszke et al., 2019). For COSMIC and DialogueRNN, we use the code provided by the respective papers. We include more details on the training hyperparameters of each model in Appendix E. To split EmoWOZ into training, validation, and testing sets, we keep the original

Model	Feature	Ctx.	F1 of Each Emotion in EmoWOZ							EmoWOZ		MultiWOZ		DialSoP	
			Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	MacF1	WgtF1	MacF1	WgtF1	MacF1	WgtF1
BERT	BERT	No	91.3	40.5	33.3	**71.1**	12.8	46.2	**89.2**	48.9	74.0	48.5	**83.8**	53.6	39.1
ContextBERT	BERT	Yes	**92.7**	35.6	**62.2**	61.8	**25.3**	45.7	88.7	**53.2**	**79.9**	**48.7**	83.3	58.8	**71.2**
DialogueRNN	GloVe	Yes	88.3	25.9	58.3	55.6	16.0	37.4	88.0	46.9	77.9	44.6	81.7	56.6	63.2
DialogueRNN	BERT	Yes	85.8	36.7	45.3	67.5	11.3	44.3	88.8	49.0	76.2	44.5	82.4	56.1	63.5
COSMIC	BERT+COMET	Yes	89.6	**41.9**	46.0	69.4	10.4	**47.2**	89.1	50.7	76.8	46.2	83.1	**58.9**	60.3

Table 5: Comparison of baseline models. We report the F1 for each emotion label (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied) on EmoWOZ as well as **Mac**ro and **Weight**ed F1 on EmoWOZ and its subsets. Please refer to Appendix F.1 for more detailed results.

Example 1: *Dissatisfied*					Example 2: *Dissatisfied*				
U: I need to arrive by 15:15 **S:** I have train TR4068 leaving at 5:35 and arriving at 5:52. **U:** I want to confirm that I will arrive by 15:15? You stated, leaving at 5:35 and arriving at 5:52? [to classify]					**U:** I also need a taxi to go between the hotel and the restaurant. I'd like to leave the Gonville hotel by 09:15 **S:** When would you like to arrive by? **U:** I just mentioned that I would like to leave by 9:15 please. [to classify]				
BERT	ContextBERT	DialogueRNN (GloVe)	DialogueRNN (BERT)	COSMIC	BERT	ContextBERT	DialogueRNN (GloVe)	DialogueRNN (BERT)	COSMIC
✗ (neutral)	✓ (dissatisfied)	✓ (dissatisfied)	✓ (dissatisfied)	✓ (dissatisfied)	✗ (neutral)	✗ (neutral)	✗ (neutral)	✗ (neutral)	✗ (neutral)

Figure 2: Example dialogues from the test data and the emotion prediction for the last utterance by each model.

split of MultiWOZ and split DialSoP with a ratio of 8:1:1, leading to 9,234, 1,100, and 1,100 dialogues in each set. We run each task on 5 different seeds and report the average performance.

For all experiments discussed in the next section, we also performed the sentiment recognition task. Results can be found in Appendix F.2.

5.1.3 Results and Discussion

Table 5 summarises the performance of baseline models. Since almost 70% of the annotations are *neutral*, we exclude it when calculating average F1 scores. In general, models that take into account context information perform better on the full EmoWOZ. This shows the importance of context or dialogue-level features in emotion recognition in task-oriented dialogues. An exception is DialogueRNN with GloVe feature, which underperforms in EmoWOZ macro F1, likely due to the embedding used. On the other hand, BERT scores very well on MultiWOZ dialogues but performs poorly on DialSoP for both setups. This suggests that emotions in MultiWOZ are less context-dependent.

BERT performs best for *apologetic* and *satisfied*, potentially due to the existence of distinguishable keywords associated with these emotions such as "thank you" for *satisfied* and "sorry" for *apologetic*. These two emotion labels do not benefit from context. In contrast, BERT produces a significantly worse F1 on *dissatisfied*, probably because users tend to express dissatisfaction more implicitly, for instance via repetition or correction, making dialogue-level features necessary.

Figure 2 shows two dialogues with implicit emotions and predictions made by respective baseline models. In example 1, the system gives the wrong time of arrival, eliciting mild annoyance from the user. BERT predicts *neutral* because in isolation, the utterance has no words suggesting dissatisfaction. All other models correctly recognise *dissatisfied*, as they capture the misunderstanding occurs in previous dialogue turns. Example 2 presents a similar but more implicit case, where all models fail. This shows that EmoWOZ contains contextualised emotions that are more implicit and subtle, requiring more sophisticated features and models.

Table 6 presents cross-data experiments with ContextBERT, examining how well the two subsets complement each other. Complementing DialSoP with dialogues from MultiWOZ largely improves the performance across all emotion labels (macro F1). On the other hand, while complementing MultiWOZ with DialSoP slightly increases macro F1 on MultiWoZ, the scores for some emotions such as *satisfied* and *dissatisfied* are decreased.

We further investigate the drop in F1 of *dissatisfied* and *satisfied* by looking at the change in recall and precision after complementing MultiWOZ with DialSoP. As shown by Table 7, *dissatisfied* sees an increase in recall, whereas *satisfied* sees an increase in precision. We believe it is necessary to distinguish recall and precision, as for some emotions, one may be more important than the other. The relative importance of recall and precision for each emotion class depends on its implication to a task-oriented dialogue system and the consequence of false recognition. For example, a high recall of *dissatisfied* is desirable because the system should not

Training Data	Test on MultiWOZ				Test on DialSoP			
	Dissatisfied	Satisfied	Macro F1	Weigted F1	Dissatisfied	Satisfied	Macro F1	Weigted F1
MultiWOZ	**38.1**	**90.1**	47.7	**84.0**	10.3	54.6	42.5	16.9
DialSoP	14.6	78.7	18.1	67.5	73.2	57.4	33.7	70.0
EmoWOZ	32.0	89.7	**48.7**	83.3	**73.5**	**60.5**	**58.8**	**71.2**

Table 6: Performance of ContextBERT in cross-dataset experiments. To summarise, we report the F1 of *dissatisfied* and *satisfied*, the most common emotions in DialSoP and MultiWOZ respectively. We also report macro F1 and weighted F1 for overall model performance. For detailed results, please refer to Appendix F.1.

Metric	Dissatisfied	Satisfied
Recall	$34.0 \rightarrow 52.3$ (↑)	$91.2 \rightarrow 89.6$ (↓)
Precision	$43.5 \rightarrow 23.1$ (↓)	$89.0 \rightarrow 89.7$ (↑)

Table 7: Change in precision and recall on MultiWOZ by ContextBERT, after adding DialSoP to training. All changes have statistical significance ($p < 0.05$).

miss any failure in dialogues. On the other hand, a high precision may be more desirable for emotions such as *satisfied* to ensure proper affective response from the system. When the relative importance of recall and precision of the emotion is taken into account, complementing MultiWOZ with DialSoP is beneficial to *dissatisfied* and *satisfied*, the two most important emotions in task-oriented dialogues. Detailed results can be found in Appendix F.3.

	Dissatisfied	Satisfied
MultiWOZ Label	1.5%	24.0%
DialSoP (#token>11.6) Label	26.6%	1.7%
DialSoP (#token>11.6) Prediction	28.2%	2.3%
DialSoP Label	37.2%	4.4%
MultiWOZ (#token<5.7) Label	1.2%	37.7%
MultiWOZ (#token<5.7) Prediction	3.2%	37.6%

Table 8: Emotion distribution in labels and ContextBERT prediction. See Appendix F.4 for full results.

Due to different linguistic features and emotion distributions in MultiWOZ and DialSoP, one concern is that the models learn to predict emotion based on these statistical artifacts. According to Table 3, the most obvious difference is the average utterance length (5.7 in DialSoP and 11.6 in MultiWOZ). A naive model may simply recognise the data source from word count and predict the most likely emotion from that source. Table 8 presents how ContextBERT trained on EmoWOZ predicts emotion in long DialSoP and short MultiWOZ utterances. The emotion distribution in model prediction is vastly different from that in the complementing subset. Clearly, the model does not simply count words to decide on the underlying emotion.

5.2 Emotions for Dialogue State Tracking

In task-oriented dialogues, dialogue state tracking (DST) aims to continuously track the user's goal and intent as the dialogue progresses (Young et al., 2010). We hypothesise that the user emotion can help inform the system about their goal. To investigate this, we train a dialogue state tracker that incorporates an additional task to predict one of 7 emotional classes on the MultiWOZ dataset. We utilise the *out-of-task training* approach and the available code presented in (Heck et al., 2020a). We follow the multitask learning (MTL) algorithm, where on each training step, the same model is trained on two different batches, one from the main task (DST) and one from the auxiliary task (emotion recognition). Since neutral emotion provides limited information on the user goal, we remove a half of the neutral utterances when performing MTL. We show that additional emotion labels can lead to a significant improvement ($p < 0.02$) in the joint goal accuracy (JGA) of DST (see Table 9).

Training tasks	JGA
Dialogue state tracking	53.7
Dialogue state tracking & emotion recognition	**54.7**

Table 9: DST JGA for MultiWOZ 2.1 (Eric et al., 2019).

6 Conclusion

We present EmoWOZ, a corpus of task-oriented dialogues with emotion annotations. We propose a novel labelling scheme derived from the OCC model to capture a set of emotions in relation to user goals in dialogues. Labelled user emotions will allow us to work towards emotion-aware task-oriented dialogue systems, for dialogues closer to human-human interactions. Baseline results show the challenge to recognise context-dependent and implicit emotions from task-oriented dialogues. There is still room for improvement, for example, by leveraging dialogue-level features such as slot-value pairs and dialogue acts. We also demonstrate the usefulness of emotion labels in training other dialogue system modules. We hope this dataset can offer insights beyond the scope of emotion recognition and push the performance of downstream tasks in task-oriented dialogue modelling.

References

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. *CoRR*, abs/1906.05317.

Pawel Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. MultiWOZ - A large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. *CoRR*, abs/1810.00278.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Amanda Cercas Curry and Verena Rieser. 2018. #MeToo Alexa: How conversational systems respond to sexual harassment. In *Proceedings of the Second ACL Workshop on Ethics in Natural Language Processing*, pages 7–14, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Amanda Cercas Curry and Verena Rieser. 2019. A crowd-based evaluation of abuse response strategies in conversational agents. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 361–366.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, pages 169–200.

Mihail Eric, Rahul Goel, Shachi Paul, Abhishek Sethi, Sanchit Agarwal, Shuyang Gao, and Dilek Hakkani-Tür. 2019. MultiWOZ 2.1: Multi-domain dialogue state corrections and state tracking baselines. *CoRR*, abs/1907.01669.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. 2007. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. *CoRR*, abs/2010.02795.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020a. Out-of-task training for dialog state tracking models. *CoRR*, abs/2011.09379.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020b. TripPy: A triple copy strategy for value independent neural dialog state tracking. *CoRR*, abs/2005.02877.

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

J. F. Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Trans. Inf. Syst.*, 2(1):26–41.

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.

Ziming Li, Julia Kiseleva, and Maarten de Rijke. 2020. Rethinking supervised learning and reinforcement learning in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3537–3546, Online. Association for Computational Linguistics.

Nurul Lubis, Michael Heck, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2017. Processing negative emotions through social communication: Multimodal database construction and analysis. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 79–85.

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.

Andrea Madotto, Zhaojiang Lin, Zhenpeng Zhou, Seungwhan Moon, Paul Crook, Bing Liu, Zhou Yu, Eunjoon Cho, and Zhiguang Wang. 2020. Continual learning in task-oriented dialogue systems.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2018. DialogueRNN: An attentive

RNN for emotion detection in conversations. *CoRR*, abs/1811.00405.

John D Mayer, David R Caruso, and Peter Salovey. 1999. Emotional intelligence meets traditional standards for an intelligence. *Intelligence*, 27(4):267–298.

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

W. Gerrod Parrott. 2001. *Emotions in social psychology: essential readings*. Key readings in social psychology. Psychology Press, Philadelphia.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

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, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *CoRR*, abs/1810.02508.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations.

Daniel Preoţiuc-Pietro, H. Andrew Schwartz, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and Elisabeth Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 9–15, San Diego, California. Association for Computational Linguistics.

J.A. Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

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.

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.

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.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

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 multi-task learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9177–9184.

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.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. *CoRR*, abs/1708.04299.

Tiancheng Zhao, Kaige Xie, and Maxine Eskenazi. 2019. Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models. 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 1208–1218.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *CoRR*, abs/1704.01074.

Xianda Zhou and William Yang Wang. 2017. MojiTalk: Generating emotional responses at scale. *CoRR*, abs/1711.04090.

Qi Zhu, Zheng Zhang, Yan Fang, Xiang Li, Ryuichi Takanobu, Jinchao Li, Baolin Peng, Jianfeng Gao, Xiaoyan Zhu, and Minlie Huang. 2020. ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 142–149. Association for Computational Linguistics.

A The OCC Model

Figure A.1 summarises definitions of emotion groups in the OCC model.

Elicitor	Aspects of events or agents			OCC Emotion
Consequences of Events	Consequence for other	Desirable for other		happy-for resentment
		Undesirable for other		gloating pity
	Consequence for self	prospects relevant	confirmed	satisfaction fears-confirmed
			disconfirmed	relief disappointment
		prospects irrelevant		joy distress
Consequences of Events / Actions of Agents	Consequence for self, prospect irrelevant, and related to actions of agents.		self agent	gratification remorse
			other agent	gratitude anger
Actions of Agents	self agent			pride shame
	other agent			admiration reproach
Aspects of Objects				love hate

Figure A.1: The OCC Model

B Amazon Mechanical Turk Set-up

B.1 Qualification Test

Figure B.1 illustrates one example from our qualification test. Hints are provided for difficult questions containing implicit emotions as shown in the example.

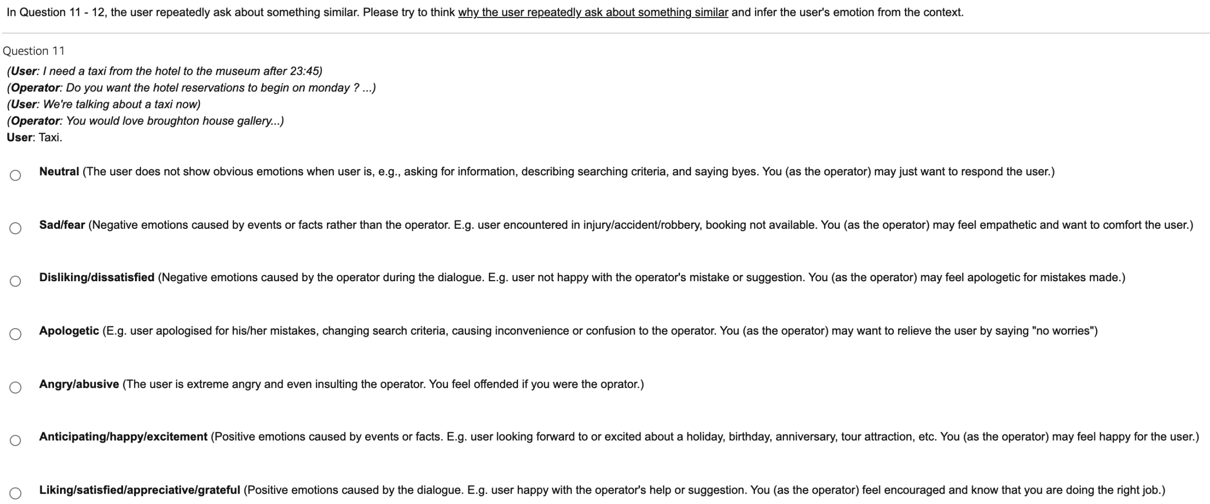

Figure B.1: One of fifteen questions in our qualification test

B.2 Main Task Page

Figure B.2 shows the task page for workers. Before arriving at this page, they will be prompted with a consent form and a message asking if they would like to go through a tutorial.

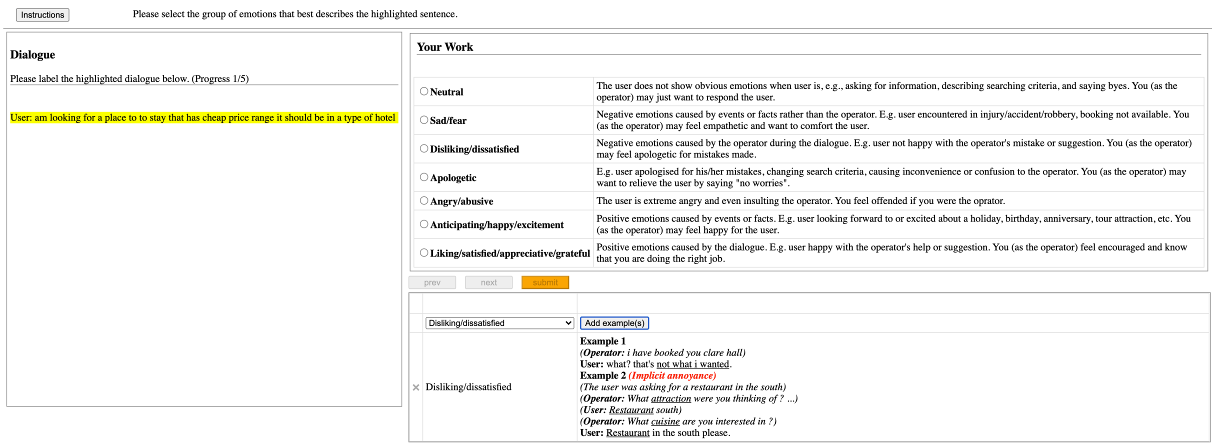

Figure B.2: Amazon Mechanical Turk main task page

C Dialogue Examples

Figure C.1 shows examples of how emotions are expressed by the user in EmoWOZ. Figure C.2 shows examples of annotated dialogues.

Emotion	Example	
Neutral	Operator:	What sort of food would you like it to be?)
	User:	You choose. Book me a table for 6 people at 12:00 on Thursday.
Fearful, sad, disappointed		(Start of the dialogue)
	User:	Can you help me please? I just got robbed and I need help!! Help!
Dissatisfied, disliking		(Explicit expression)
	Operator:	i have booked you clare hall
	User:	what? that's not what i wanted.
		(Implicit expression)
	User:	I need a taxi from the hotel to the museum after 23:45
	Operator:	Do you want the hotel reservations to begin on Monday?...
	User:	We're talking about a taxi now
	Operator:	You would love broughton house gallery
	User:	Taxi.
Apologetic	Operator:	I'm afraid that's all we've got.
	User:	I'm sorry I may have confused you. I would want to go from Petersborough to Cambridge on Saturday and arrive by 13:00, so anything that will get me there before that.
Abusive	Operator:	(talking about something irrelevant)
	User:	**** ***, would ya stupid thing
Excited, happy, anticipating	Operator:	There are 3 restaurants on the west side that serve British food. Do you prefer moderate or expensive priced?
	User:	Expensive, please, we're splashing out for my mother's birthday
Satisfied, liking, appreciative, grateful	Operator:	Any time. I was happy to help.
	User:	You are doing a great job! That's all I needed

Figure C.1: Example for each emotion label

Annotation Example 1 (MultiWOZ)	
Operator:	... Is there anything else I can help with today?
User:	Yes I would like to visit some architecture landmarks in the centre of the city. What would you recommend? [Excited]
Operator:	yes, there are four churches and a school in that area.
User:	Could you suggest one to me? I will also need to know the address and phone number, as well as the entrance fee. Thanks. [neutral]
Operator:	all saints church is free admission and is located at jesus lane. The phone number is 01223452587.
User:	Thanks so much! That is all I need help with today! [appreciative]
Operator:	You're quite welcome! Have a lovely day, goodbye!

Annotation Example 2 (DialSoP)	
Operator:	Where is your destination ?
User:	liverpool [neutral]
Operator:	The arrival time is 05:51 . The train is for sunday you are all set. The train will be departing from cambridge . The booking is for arriving in london kings cross .
User:	thats great...to bad the city is completely wrong but ok [dissatisfied]
	... ...
User:	yes book a return on Monday [neutral]
Operator:	Did you need any further assistance today ?
User:	yes please book a return on Monday [dissatisfied]
Operator:	Is there anything else I can help you with ?
User:	BOOK A RETURN TRAIN ON MONDAY [abusive]

Figure C.2: Annotation examples from EmoWOZ

D Annotator Agreement in EmoWOZ

Source	Fleiss' Kappa	#NA	#PA	#FA
MultiWOZ	0.611	1,016	17,270	53,238
DialSoP	0.465	259	4,778	7,069
EmoWOZ (overall)	0.602	1275	22,048	60,307

Table D1: Inter-annotator agreement and agreement count of EmoWOZ and its subsets. NA means no agreement—three annotators annotate with three different emotions. PA means partial agreement—only two annotators annotate with the same emotion. FA means full agreement—three annotators annotate with the same emotion.

E Hyperparameters for Model Training

Model	Optimiser	Learning Rate	L2 Reguliser Weight	Training Epochs
BERT	Adam	2e-5	0	10
ContextBERT	Adam	2e-5	0	10
DialogueRNN(GloVe)	Adam	1e-4	1e-5	60
DialogueRNN(BERT)	Adam	1e-4	1e-4	60
COSMIC	Adam	1e-4	3e-4	20

Table E1: Hyperparameters for model training

F Detailed Cross-dataset Experiment Results

F.1 Emotion Classification (7 classes)

Model	**Set-up**	**F1 for each Emotion Label**							**Average F1 w/o Neutral**			**Average F1 w Neutral**		
		Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied	Micro	Macro	Weighted	Micro	Macro	Weighted
BERT	D → D	70.71	10.00	**47.95**	0	0	**69.81**	**73.68**	**51.14**	33.57	**50.25**	**62.96**	38.88	**61.74**
	M → D	**73.47**	**64.00**	1.85	**100**	0	55.60	73.19	19.52	49.11	11.41	59.57	52.59	46.27
	E → D	73.14	50.00	34.62	**100**	0	66.71	70.15	41.01	**53.58**	39.10	62.55	**56.37**	58.22
	D → M	89.01	0	6.14	8.64	9.97	16.69	79.02	58.72	20.08	69.64	79.38	29.93	83.64
	M → M	95.29	**44.32**	**41.55**	**73.07**	**22.14**	38.19	**90.62**	**85.30**	**51.65**	**84.96**	**92.55**	**57.88**	**92.43**
	E → M	94.17	37.68	31.00	70.72	20.18	**41.59**	89.9	82.19	48.51	83.77	90.80	55.04	91.29
ContextBERT	D → D	82.65	7.14	73.20	0	0	61.85	57.44	70.54	33.67	69.95	77.45	40.65	77.03
	M → D	73.17	50.71	10.30	71.67	0	67.42	54.62	24.16	42.45	16.91	59.22	46.84	48.51
	E → D	**82.68**	**51.11**	**73.49**	**100**	0	**67.94**	**60.45**	**71.49**	**58.83**	**71.17**	**77.78**	**62.24**	**77.64**
	D → M	91.51	2.94	14.58	0	2.08	11.96	78.68	65.78	18.14	67.54	84.83	28.58	84.70
	M → M	**94.97**	**32.24**	**38.09**	**65.39**	21.43	39.15	**90.06**	**84.44**	47.73	**83.97**	**92.05**	54.48	**91.92**
	E → M	94	31.23	31.99	61.27	**37.09**	**40.72**	89.66	81.72	**48.66**	83.25	90.50	**55.14**	91.02
DialogueRNN (GloVe)	D → D	57.95	0	**63.40**	0	0	62.58	67.11	62.59	32.18	62.37	**60.78**	35.86	**59.89**
	M → D	**71.94**	31.94	22.48	93.33	0	59.14	71.14	34.01	46.34	28.46	59.79	49.99	52.88
	E → D	43.27	**35.88**	62.69	**100**	0	**67.24**	**74.00**	**63.18**	**56.63**	**63.16**	55.89	**54.73**	51.99
	D → M	84.01	0	6.62	0	0	13.93	85.83	66.27	17.73	75.12	77.80	27.20	81.55
	M → M	91.25	23.88	28.76	**59.29**	15.06	30.40	**88.47**	78.23	40.98	81.49	87.08	48.16	88.55
	E → M	**92.44**	**24.36**	**32.98**	55.17	**34.60**	**32.11**	88.46	**80.05**	**44.61**	**81.68**	**88.59**	**51.45**	**89.46**
DialogueRNN (BERT)	D → D	71.81	29.65	55.32	0	**20.89**	71.04	70.88	57.27	41.30	56.66	**65.62**	45.66	**65.17**
	M → D	**73.11**	**62.12**	7.15	90.00	0	62.10	**71.34**	24.92	48.79	15.89	59.59	52.26	48.03
	E → D	57.47	32.14	**63.55**	**100**	0	**72.44**	68.37	**63.80**	**56.08**	**63.45**	61.20	**56.28**	60.09
	D → M	89.71	3.83	7.78	21.61	3.85	19.68	83.68	66.06	23.41	74.33	82.30	32.88	85.45
	M → M	**94.16**	**42.96**	**38.99**	**69.53**	**26.97**	36.77	**89.88**	**83.17**	**50.85**	**84.02**	**90.95**	**57.04**	**91.35**
	E → M	88.86	37.69	17.20	67.06	18.21	**37.46**	89.46	73.78	44.51	82.38	83.91	50.85	87.06
COSMIC	D → D	71.43	0	58.87	0	**5.33**	71.15	71.60	59.96	34.49	59.28	66.31	39.77	66.10
	M → D	**73.47**	**64.00**	3.47	**100**	0	62.19	**73.95**	22.35	50.60	13.07	59.78	53.87	46.99
	E → D	71.34	50.00	**59.25**	**100**	0	**73.30**	70.88	**60.79**	**58.91**	**60.28**	**66.71**	**60.68**	**66.49**
	D → M	89.04	0	7.76	0	7.01	21.66	84.09	64.20	20.09	73.97	81.02	29.94	84.87
	M → M	**95.04**	**45.62**	**41.69**	**71.53**	**22.14**	38.79	**90.56**	**84.81**	**51.72**	**84.90**	**92.16**	**57.91**	**92.23**
	E → M	92.11	39.78	23.50	68.94	15.05	**40.14**	89.7	78.16	46.19	83.09	87.94	52.75	89.61

Table F1: Performance of baseline models on emotion classification including cross-dataset experiments. For cross-dataset experiments, the "X → Y"s in the 'Set-up' column represents the training and evaluation set-up, where X is the training set and Y is the test set. E stands for EmoWOZ, M stands for MultiWOZ, and D stands for DialSoP. M → D, for example, means to train on MultiWOZ and test on DialSoP. Extreme values for "Apologetic" and "Abusive" in DialSoP ("* → D"s) are caused by their rarity in the test set (1 and 5 occurrences respectively).

F.2 Sentiment Classification (3 classes)

Model	Feature	Ctx.	F1 of Each Sentiment in EmoWOZ			Average F1 w/o Neutral					
						EmoWOZ		MultiWOZ		DialSoP	
			Neutral	Negative	Positive	Macro	Weighted	Macro	Weighted	Macro	Weighted
BERT	BERT	No	91.92	42.73	**89.46**	66.09	77.06	**70.84**	**86.31**	53.54	40.53
ContextBERT	BERT	Yes	**92.44**	59.34	88.00	73.67	**80.40**	68.08	84.90	64.00	65.79
DialogueRNN	GloVe	Yes	89.62	**62.46**	86.67	**74.56**	80.25	70.59	83.80	**69.97**	**66.67**
DialogueRNN	BERT	Yes	81.35	42.90	88.32	65.61	76.28	56.17	82.41	67.73	65.03
COSMIC	BERT+COMET	Yes	90.54	50.24	89.43	69.83	79.03	66.49	85.35	65.11	58.80

Table F2: Summarised performance of baseline models on sentiment classification.

Model	Set-up	F1 for each Sentiment Label			Average F1 w/o Neutral			Average F1 w Neutral		
		Neutral	Negative	Positive	Micro	Macro	Weighted	Micro	Macro	Weighted
BERT	$D \rightarrow D$	70.56	**49.35**	**73.20**	**53.04**	**61.27**	**52.44**	63.56	**64.37**	**62.59**
	$M \rightarrow D$	73.29	2.84	72.95	20.37	37.89	11.92	59.53	49.69	46.29
	$E \rightarrow D$	**74.07**	35.99	71.10	42.67	53.54	40.53	**63.81**	60.38	59.32
	$D \rightarrow M$	89.06	16.50	83.16	67.98	49.83	76.41	82.37	62.91	85.54
	$M \rightarrow M$	**95.51**	**57.45**	**90.37**	**87.32**	**73.91**	**87.03**	**93.26**	**81.11**	**93.15**
	$E \rightarrow M$	94.80	51.43	90.25	85.42	70.84	86.31	92.15	78.82	92.43
ContextBERT	$D \rightarrow D$	**81.89**	**72.25**	58.55	**70.16**	**65.40**	**70.48**	**76.95**	**70.90**	**76.87**
	$M \rightarrow D$	72.33	4.69	59.45	20.84	32.07	11.79	57.87	45.49	45.70
	$E \rightarrow D$	79.51	66.42	**61.58**	65.68	64.00	65.79	73.88	69.17	73.48
	$D \rightarrow M$	91.88	16.52	81.72	74.62	49.12	75.11	87.25	63.37	87.21
	$M \rightarrow M$	**95.16**	**51.57**	**89.70**	**86.40**	**70.63**	**85.83**	**92.74**	**78.81**	**92.57**
	$E \rightarrow M$	94.27	46.98	89.17	84.10	68.08	84.90	91.40	76.81	91.66
DialogueRNN (GloVe)	$D \rightarrow D$	**72.45**	**67.55**	69.82	**67.89**	68.68	**67.85**	**70.25**	**69.94**	**70.42**
	$M \rightarrow D$	72.07	18.04	58.01	30.06	38.02	23.21	58.67	49.37	50.58
	$E \rightarrow D$	53.57	65.52	**74.43**	66.47	**69.97**	66.67	61.36	64.51	59.33
	$D \rightarrow M$	90.40	9.97	85.58	76.22	47.77	77.92	86.03	61.98	86.92
	$M \rightarrow M$	92.37	51.83	86.01	82.14	68.92	82.55	89.18	76.74	89.64
	$E \rightarrow M$	**93.19**	**54.01**	**87.16**	**83.57**	**70.59**	**83.80**	**90.26**	**78.12**	**90.58**
DialogueRNN (BERT)	$D \rightarrow D$	**72.79**	57.39	71.26	59.53	64.32	59.18	**67.25**	**67.15**	**66.80**
	$M \rightarrow D$	70.93	11.51	70.69	26.66	41.10	19.17	58.34	51.04	48.16
	$E \rightarrow D$	50.75	**64.09**	**71.38**	**64.81**	**67.73**	**65.03**	59.35	62.07	57.03
	$D \rightarrow M$	89.78	20.57	85.49	72.83	53.03	78.92	84.49	65.28	86.76
	$M \rightarrow M$	**92.98**	**47.79**	**89.20**	**83.00**	**68.49**	**85.00**	**89.95**	**76.66**	**90.76**
	$E \rightarrow M$	84.58	23.34	89.00	70.34	56.17	82.41	79.48	65.64	83.98
COSMIC	$D \rightarrow D$	68.20	**61.92**	72.34	**63.34**	**67.13**	**63.27**	66.22	67.49	66.04
	$M \rightarrow D$	73.21	3.15	71.99	21.10	37.57	12.07	59.43	49.45	46.32
	$E \rightarrow D$	**73.77**	56.60	**73.63**	59.35	65.11	58.80	**67.94**	**68.00**	**67.18**
	$D \rightarrow M$	86.53	17.74	85.78	68.85	51.76	78.89	80.58	63.35	84.40
	$M \rightarrow M$	**95.30**	**57.66**	**90.20**	**86.99**	**73.93**	**86.90**	**92.97**	**81.05**	**92.96**
	$E \rightarrow M$	92.96	42.84	90.14	82.31	66.49	85.35	89.76	75.31	90.84

Table F3: Detailed results of baseline models on sentiment classification including cross-dataset experiments. For cross-dataset experiments, the "$X \rightarrow X$"s in the 'Set-up' column represents the training and evaluation set-up. E stands for EmoWOZ, M stands for MultiWOZ, and D stands for DialSoP. $M \rightarrow D$, for example, means to train on MultiWOZ and test on DialSoP.

F.3 Change in precision and recall on MultiWOZ after Complementing MultiWOZ with DialSoP in Training

		Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied
ContextBERT	Recall	**94.9 → 92.9**	28.8 → 28.8	**34.0 → 52.3**	62.8 → 60.8	**13.3 → 33.3**	39.2 → 38.1	**91.2 → 89.6**
	Precision	95.0 → 95.1	37.2 → 35.1	**43.5 → 23.1**	68.6 → 65.2	60.0 → 42.0	39.1 → 44.3	**89.0 → 89.7**
	F1	**95.0 → 94.0**	32.2 → 31.2	**38.1 → 32.0**	**65.4 → 61.3**	**21.4 → 37.1**	39.2 → 40.7	**90.1 → 89.7**

Table F4: Precision, recall and F1 score of ContextBERT for all emotions when trained on MultiWOZ and EmoWOZ respectively, and tested on MultiWOZ. $A \rightarrow B$ represents how the value change after complementing MultiWOZ with DialSoP in training. A is the value when trained on MultiWOZ and B is the value when trained on EmoWOZ. Values with statistical significance ($p < 0.05$) are bolded and colored where red indicates a drop and green indicates an improvement. For recognising user emotions in task-oriented dialogues, a high precision is more desirable for *neutral*, *apologetic*, *abusive*, *excited*, and *satisfied* where as a high recall is more desirable for *fearful* and *dissatisfied*.

F.4 Emotion Distribution in Model Predictions

Test Set	Model	Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied
MultiWOZ Label		72.31	0.22	1.45	0.98	0.08	1.00	23.96
DialSoP (#token>11.8) Label		64.74	0	26.59	0.58	0	6.36	1.73
DialSoP (#token > 11.8) Prediction	BERT	88.55	0	5.43	0.58	0	4.05	1.39
	ContextBERT	62.20	0	28.21	0.58	0	6.71	2.31
	DialogueRNN-GloVe	33.29	0.23	52.49	0.58	0.23	11.45	1.73
	DialogueRNN-BERT	43.58	0.12	42.54	0.58	0	11.45	1.73
	COSMIC	66.24	0	20.23	0.58	0	11.45	1.50
DialSoP Label		56.17	0.47	37.18	0.08	0.4	1.34	4.35
MultiWOZ (#token<5.8) Label		60.76	0	1.21	0	0	0.3	37.73
MultiWOZ (#token < 5.8) Prediction	BERT	61.06	0	1.85	0.03	0	0.27	36.79
	ContextBERT	58.82	0.06	3.21	0.03	0	0.33	37.55
	DialogueRNN-GloVe	55.39	0.27	1.36	0.15	0	1.18	41.64
	DialogueRNN-BERT	47.76	0.06	12.79	0.12	0	0.67	38.61
	COSMIC	56.79	0	5.97	0.15	0.03	0.45	36.61

Table F5: Emotion distribution in model predictions (trained on EmoWOZ).