Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4096–4113
Marseille, 20-25 June 2022
© European Language Resources Association (ELRA), licensed under CC-BY-NC-4.0

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

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk and Milica Gašić
Heinrich Heine University Düsseldorf
Universitätsstraße 1, 40225 Düsseldorf, Germany
{fengs, lubis, geishaus, linh, heckmi, niekerk, gasic}@hhu.de

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 remain largely unaddressed. This is despite emotions and dialogue success having equally important roles in a natural system. 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.

Keywords: Emotion Recognition in Conversations, 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., 2018).

Dialogue systems generally fall into two classes. Task-oriented systems converse with users to help complete tasks determined by user goals. 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 (RL) 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., 2021). Consequently, the distribution of data from which a task-oriented system learns can change.

Emotions appear in both chit-chat and task-oriented dialogues. However, the cause of emotion may differ as well as their role. Chit-chat dialogues are 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, the user is primarily interested in achieving their goal. While an emotional situation may be a reason to interact with the system, e.g. the user just missed a flight and needs to rebook one, the emotion the user exhibits is more often a reaction to potential goal completion or failure. Since the emotion is centred around the user goal, it is 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 (Majumder et al., 2019; Ghosal et al., 2020). There are several public ERC corpora containing chit-chat dialogues (Li et al., 2017; Poria et al., 2019; Zahiri and Choi, 2018) and conversational data from social media (Zhou and Wang, 2018). These corpora can tremendously accelerate the building of emotional chatbots using data-driven approaches (Zhou et al., 2018). In task-oriented dialogues, recognising emotions is equally important but remains largely unaddressed. Using RL to optimise a dialogue policy necessitates a feedback signal. While it is accepted that the feedback signal needs to correlate with user satisfaction (Ultes et al., 2017), this feedback signal is often based on hand-coded rules. Could an emotional model instead be directly used to provide such a feedback signal? Could it also be used to support emotion-aware

natural language generation (Mairesse and Walker, 2007), or even improve dialogue state tracking through multi-task learning (Heck et al., 2020a)? Existing corpora are small in size, and labels are limited to sentiment polarity, creating a bottleneck, so these questions remain largely unexplored.

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; Lin et al., 2021) to policy optimisation (He et al., 2022). 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[1] 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 an array 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 widespread. In a similar spirit, Gross and Thompson (2007) formulated the process of emotion regulation as the attention, appraisal, and response originated from various situations.

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., 2019), but some datasets have domain-specific labels. For instance, Zhou and Wang (2018) 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.

In this work, we propose a novel set of 7 emotions and motivate it using OCC model as the basis. We aim for this scheme to capture the cognitive context of emotions while retaining the simplicity of labels that facilitates large-scale crowd-sourcing of emotion annotations.

2.2. Emotion Dialogue Datasets

Early works on ERC focus on speech signals (Cowie et al., 2001; Riccardi and Hakkani-Tür, 2005; Carrión and López-Cózar, 2008). More recently, there are increasing number of text-based ERC datasets focusing on chit-chat dialogue. Chit-chat dialogue lends itself well to affective computing research due to its open-domain set-up, where conversation topics are diverse and not restricted to a particular task. One of the largest such corpora is DailyDialog (Li et al., 2017), which contains conversations between English learners on various topics ranging from relationships to money. Other similar datasets include EmoryNLP (Zahiri and Choi, 2018) and MELD (Poria et al., 2019). 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

[1] https://doi.org/10.5281/zenodo.5865437

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	**167,234**
# 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	83,617
Neutral Samples (%)	83.1%	47.0%	**30.0%**	-	88.6%	45.7%	70.1%
# 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. For corpora providing both emotion and sentiment labels, agreement metrics are measured for emotion labels. DSTC1, SentiVA, and TML refer to works by Shi and Yu (2018), Saha et al. (2020), and Wang et al. (2020), respectively.

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 et al., 2017) with sentiment for policy optimisation. Since dialogues are machine-generated, it is unclear how well these emotions match real human emotions and whether sentiment on its own sufficiently captures emotional nuances in task-oriented dialogue. 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 and application in machine learning. To summarise, existing corpora are either limited in size or not publicly available, limiting further works on emotions in task-oriented dialogue systems. Furthermore, their annotation schemes focus on sentiment polarities, overlooking the effect of goals on users' emotional states.

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). It contains over 10k dialogues. 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: We envisage emotions as learning signal for dialogue system optimisation. Since emotions in task-oriented dialogue systems can be a direct effect of the user perception of the ability of the system to fulfill their goal, the policy performance can largely influence emotion distribution. During the life span of a data-driven task-oriented dialogue system, the distributions of dialogues and emotions may change as the policy learns and improves over time. An immediate impact of such a distributional shift is the increase in the number of negative emotions due to failed dialogues during the early stages of learning. Therefore, in addition to the human wizard policy in MultiWOZ, it is important that EmoWOZ covers a variety of dialogues which represent the emotions throughout such a dialogue system life span.

We complement MultiWOZ with human-machine **dial**ogues from a **ma**chine-**ge**nerated policy (**DialMAGE**). To elicit more genuine reactions, we let subjects directly interact with a machine-generated policy instead of human wizards trying to make machine-like mistakes. We launched a dialogue interactive task on Amazon Mechanical Turk, where workers are asked to retrieve information by interacting with the learning policy. We start with a policy trained in a supervised fashion on MultiWOZ that achieved a task success rate of 55% when evaluated with the ConvLab-2 (Zhu et al., 2020) rule-based user simulator. Throughout the task, the policy learned and improved as user feedback on task success is used for further training using RL. The policy reached a final human-rated success rate of 73%. 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.

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

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, satisfaction, relief, joy	Excited, happy, anticipating	Looking forward to a good event (e.g. birthday party).
		Impolite		Not applicable to the dataset	
	Negative	Polite	Distress, resentment, hate, fears-confirmed, pity, disappointment	Fearful, sad, disappointed	Encountered a bad event (e.g. robbery).
		Impolite		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* and therefore are not modelled for now. For simplicity, emotion words in blue are used to represent each emotion category. The OCC model is illustrated in Appendix A.

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 decided 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. As suggested by Carrión and López-Cózar (2008) to improve the annotation quality, 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. Due to the high subjectivity in the emotion annotation (Devillers et al., 2005), 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 diffi-

cult 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 DialMAGE differ linguistically. As seen in Table 3, DialMAGE has longer dialogues than MultiWOZ as it takes longer for the machine-generated policy to accomplish user goals. Meanwhile, users use simpler and shorter sentences when talking to a machine. Especially when the system under-performs, users are discouraged 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	DialMAGE	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 and its subsets.

4.2. Emotion Distribution

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 DialMAGE. MultiWOZ contributes most *satisfied* utterances whereas DialMAGE 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 DialMAGE, suggesting that users tend to be more polite when talking to human operators. Dialogues from MultiWOZ also contain more event-elicited emotions (*fearful* and *excited*) than DialMAGE. 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.

Emotion	EmoWOZ		MultiWOZ		DialMAGE	
	Count	Prop.	Count	Prop.	Count	Prop.
Neutral	58,656	70.1%	51,426	71.9%	7,230	59.8%
Fearful	396	0.5%	381	0.5%	15	0.1%
Dissatisfied	5,117	6.1%	914	1.3%	4,203	34.8%
Apologetic	840	1.0%	838	1.2%	2	0.02%
Abusive	105	0.2%	44	0.1%	61	0.5%
Excited	971	1.2%	860	1.2%	111	0.9%
Satisfied	17,532	21.0%	17,061	23.8%	471	3.9%

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

4.3. Inter-annotator Agreement

We measure the inter-annotator agreement by computing Fleiss' Kappa (Fleiss, 1971). Fleiss' Kappa for EmoWOZ is 0.602, suggesting a substantial agreement. Fleiss' Kappa for MultiWOZ is 0.611, higher than 0.465 for DialMAGE. Emotions in DialMAGE are more challenging to annotate because users express emotion less explicitly when they know that they are talking to a machine that 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.

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*. A reasonable explanation is that workers adopt different valence or impoliteness thresholds when they make decisions. Note that dissatisfied is rarely confused with abusive, but rather with neutral, suggesting that the ambiguity lies in when an expression of dissatisfaction is considered

Model	Feature	Ctx.	F1 of Each Emotion in EmoWOZ							EmoWOZ		MultiWOZ		DialMAGE	
			Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mac.	Wgt.	Mac.	Wgt.	Mac.	Wgt.
BERT	BERT	No	89.8	36.2	35.1	70.4	27.5	42.9	88.8	50.1	73.5	**48.4**	**83.2**	42.7	43.8
ContextBERT	BERT	Yes	**92.1***	30.1	**61.7***	62.4	**41.7**	40.8	**89.1**	54.3	**79.7***	45.1	83.1	50.0	**73.5***
DialogueRNN	GloVe	Yes	83.5	12.7	51.4	57.7	0.0	32.7	86.4	40.1	74.6	34.1	79.2	43.2	61.2
DialogueRNN	BERT	Yes	86.9	41.3	47.5	**71.5**	25.6	39.4	87.6	52.1	75.5	44.5	81.9	51.4	60.6
COSMIC	BERT+COMET	Yes	89.8	**52.0***	50.7	70.9	31.6	**44.4**	88.4	**56.3**	77.1	46.7	82.7	**57.2**	61.7

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 **Wei**ghted F1 (excluding neutral) on EmoWOZ and its subsets. "Ctx." stands for "context". * indicates statistically significant difference with $p < 0.05$ between the best and the second best values in each column. Please refer to Appendix F.1 for more detailed results.

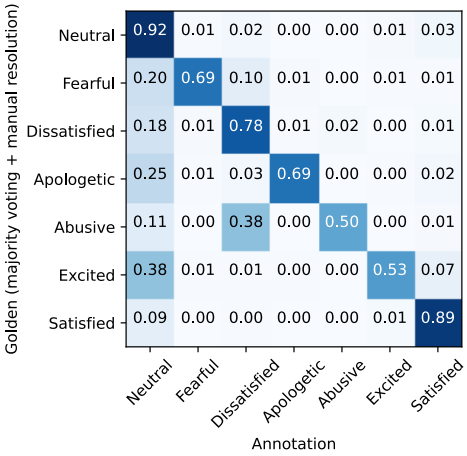


Figure 1: Confusion matrix of emotion annotations.

to be rude or abusive, and not due to the similarity between abuse and dissatisfaction.

On the other hand, confusions between *fearful* and *dissatisfied* suggest workers may also interpret elicitors differently. For example, a user may express negative emotions 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. We believe differences on interpretations are natural to a certain extent, as emotion appraisal may differ across individuals (Kuppens et al., 2007).

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. As baselines, we compare two models originally developed for chit-chat emotion recognition as well as various BERT-based models. We believe emotion recognition is the first step towards an emotion-aware task-oriented agent, as a means for a deployed agent to obtain emotion information during an interaction.

5.1.1. Baselines

BERT (Devlin et al., 2019): 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 in the reversed order to form one long sequence. We add "`User:`" and "`System:`" to mark the speaker of each turn.

DialogueRNN (Majumder et al., 2019): 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[2]. 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 hyperparameters of each model in Appendix E. To split EmoWOZ into training, validation, and testing sets, we

[2]We also performed the same experiments on 3 sentiment labels. Results can be found in Appendix F.3.

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 and the emotion prediction for the last utterance by each model.

Training Data	Test on MultiWOZ									Test on DialMAGE								
	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mac.	Wgt.	Neu.	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.	Mac.	Wgt.
MultiWOZ	**95.1***	**35.7**	**36.4***	**70.3***	**19.4**	34.1	**90.0**	**47.7**	**83.9**	80.2	**11.7**	7.7	**43.7**	11.9	**60.1**	66.3	33.6	14.5
DialMAGE	89.4	0	11.2	0	0	13.9	77.3	17.0	67.8	72.1	0	**75.7**	0	5.0	58.6	**71.7**	35.2	72.9
EmoWOZ	93.5	33.7	30.4	62.4	17.3	**37.1**	89.8	45.1	83.1	**81.6**	5.0	75.5	40.0	**52.8***	57.3	69.2	**50.0***	**73.5**

Table 6: Performance of ContextBERT in cross-dataset experiments. We report the F1 for each emotion label (**Neu**tral, **Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied), as well as **Mac**ro and **Weight**ed F1 (excluding neutral). * indicates statistically significant difference with $p < 0.05$ between the best and the second best values in each column. For detailed results, please refer to Appendix F.1.

keep the original split of MultiWOZ and further divide DialMAGE 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.

5.1.3. Results and Discussion

Recognition on emotion classes. 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 non-contextual embedding used. On the other hand, BERT scores very well on MultiWOZ dialogues but performs poorly on DialMAGE for both setups. This suggests that emotions in MultiWOZ are less context-dependent.

BERT, the only non-contextual model among our baselines, performs well for *apologetic*, *excited*, 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 emotion labels do not benefit much 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.

Complementarity between MultiWOZ and DialMAGE. Due to different linguistic features and emotion distributions in MultiWOZ and DialMAGE, 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 (5.8 in DialMAGE and 11.8 in MultiWOZ). A naive model may simply recognise the data source from word count and predict the most likely emotion from that source. Table 7 presents how ContextBERT trained on EmoWOZ predicts emotion in long DialMAGE 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.

	Dissatisfied	Satisfied
MultiWOZ Label	1.5%	24.0%
DialMAGE (#token>11.8) Label	28.8%	1.2%
DialMAGE (#token>11.8) Prediction	35.5%	1.5%
DialMAGE Label	39.3%	4.0%
MultiWOZ (#token<5.8) Label	1.2%	37.7%
MultiWOZ (#token<5.8) Prediction	3.0%	38.9%

Table 7: Emotion distribution in labels and ContextBERT prediction. See Appendix F.5 for full results.

Table 6 presents cross-data experiments with ContextBERT, examining how well the two subsets complement each other. Complementing DialMAGE with dialogues from MultiWOZ improves the macro F1 and the F1 score of *abusive* significantly. On the other

hand, while complementing MultiWOZ with DialMAGE leads to a slight improvement in the F1 score of *excited*, other F1 scores decrease to various extent.

Recall and precision on satisfied and dissatisfied for task-oriented dialogue. We further investigate the change in F1 of each emotion on MultiWOZ by looking at the change in recall and precision after complementing MultiWOZ with DialMAGE. 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. Most importantly for task-oriented dialogue system, a high recall of *dissatisfied* is desirable because the system should not miss any failure in dialogues. Failing to recognise dissatisfaction can trigger more anger from the user and therefore impair task completion (see Figure C.2). On the other hand, a high precision may be more desirable for all other emotions 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 DialMAGE is beneficial to {*dissatisfied*} for higher recall and {*fearful*, *excited*, *satisfied*} for higher precision, see Table 8. Detailed results can be found in Appendix F.4.

Metric	Fea.	Dis.	Apo.	Abu.	Exc.	Sat.
Recall	−6.7**	+29.1**	−7.8	+8.0	+0.6	-0.4
Precision	+7.5*	−22.8**	−7.9	−11.3	+7.1**	+0.1

Table 8: Change in precision and recall for each emotion label (**Fea**rful, **Dis**satisfied, **Apo**logetic, **Abu**sive, **Exc**ited, **Sat**isfied) on MultiWOZ by ContextBERT, after adding DialMAGE to training. ** and * indicate statistically significant changes with $p < 0.05$ and $p < 0.1$ respectively.

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 MultiWOZ 2.1 (Eric et al., 2020). 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: JGA of DST on MultiWOZ.

6. Conclusion

In this work, we examined emotions and their expression in the context of task-oriented dialogues, where emotions are centred around a user goal. We used the OCC model as a starting point to derive a comprehensive annotation scheme beyond sentiment polarity for emotions in relation to user goals. We designed a set of 7 emotions that differ in terms of valence, conduct and elicitor to capture the cognitive context of emotions, while maintaining labeling simplicity.

With EmoWOZ, we present a publicly available, large-scale human-annotated emotion corpus consisting of Wizard-of-Oz style as well as dialogues with a machine-generated policy. Our intention with EmoWOZ is to overcome the lack of large emotion-labelled corpora to support research towards emotion-aware task-oriented dialogue systems, for dialogues closer to human-human interactions.

We apply various emotion recognition models to EmoWOZ and examined the effect of context for different emotions. In cross-dataset experiments we analysed the complementarity of WOZ-style data and machine-generated policy data. Our results show that recognising context-dependent and implicit emotions from task-oriented dialogues is a challenging task that will benefit from further research. EmoWOZ provides an ideal test bed for that. Lastly, we leveraged emotion recognition in the dialogue state tracking task to exemplify the utility of emotion labels in dialogue modeling. 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. In future work, we plan to investigate tailored models for emotion recognition in task-oriented dialogues that take advantage of high-level features such as dialogue acts or belief states. We are also interested in using emotion as a feedback signal within reinforcement learning policy optimisation.

7. 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 and H-C. Lin are supported by funds from the European Research Council (ERC) provided under the Horizon 2020 research and innovation programme (Grant agreement No. STG2018 804636). Computing resources were provided by Google Cloud.

8. Bibliographical References

Bordes, A., Boureau, Y.-L., and Weston, J. (2017). Learning end-to-end goal-oriented dialog. In *ICLR*. OpenReview.net.

Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., and Choi, Y. (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, July. Association for Computational Linguistics.

Budzianowski, P., Wen, T.-H., Tseng, B.-H., Casanueva, I., Ultes, S., Ramadan, O., and Gašić, M. (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, October-November. Association for Computational Linguistics.

Buechel, S. and Hahn, U. (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, April. Association for Computational Linguistics.

Carrión, Z. C. and López-Cózar, R. (2008). Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Commun.*, 50:416–433.

Cercas Curry, A. and Rieser, V. (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, June. Association for Computational Linguistics.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80.

Curry, A. C. and Rieser, V. (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.

Devillers, L., Vidrascu, L., and Lamel, L. (2005). Challenges in real-life emotion annotation and machine learning based detection. *Neural networks : the official journal of the International Neural Network Society*, 18 4:407–22.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

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

Eric, M., Goel, R., Paul, S., Sethi, A., Agarwal, S., Gao, S., Kumar, A., Goyal, A., Ku, P., and Hakkani-Tur, D. (2020). MultiWOZ 2.1: A consolidated multi-domain dialogue dataset with state corrections and state tracking baselines. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 422–428, Marseille, France, May. European Language Resources Association.

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

Fontaine, J. R., Scherer, K. R., Roesch, E. B., and Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057.

Ghosal, D., Majumder, N., Gelbukh, A., Mihalcea, R., and Poria, S. (2020). COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2470–2481, Online, November. Association for Computational Linguistics.

Gopalakrishnan, K., Hedayatnia, B., Chen, Q., Gottardi, A., Kwatra, S., Venkatesh, A., Gabriel, R., and Hakkani-Tür, D. (2019). Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Gross, J. and Thompson, R. (2007). Emotion regulation: Conceptual foundations. *Handbook of Emotion Regulation*, pages 3–27, 01.

He, W., Dai, Y., Zheng, Y., Wu, Y., Cao, Z., Liu, D., Jiang, P., Yang, M., Huang, F., Si, L., et al. (2022). Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Heck, M., Geishauser, C., Lin, H.-c., Lubis, N., Moresi, M., van Niekerk, C., and Gasic, M. (2020a). Out-of-task training for dialog state tracking models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6767–6774, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Heck, M., van Niekerk, C., Lubis, N., Geishauser, C., Lin, H.-C., Moresi, M., and Gasic, M. (2020b). TripPy: A triple copy strategy for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 35–44, 1st virtual meeting, July. Association for Computational Linguistics.

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

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

Kuppens, P., Van Mechelen, I., Smits, D. J., De Boeck, P., and Ceulemans, E. (2007). Individual differences in patterns of appraisal and anger experience. *Cognition and Emotion*, 21(4):689–713.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (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, November. Asian Federation of Natural Language Processing.

Li, Z., Kiseleva, J., and de Rijke, M. (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, November. Association for Computational Linguistics.

Lin, W., Tseng, B.-H., and Byrne, B. (2021). Knowledge-aware graph-enhanced GPT-2 for dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7871–7881, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Lubis, N., Sakti, S., Neubig, G., Toda, T., and Nakamura, S. (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.

Lubis, N., Heck, M., Sakti, S., Yoshino, K., and Nakamura, S. (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.

Madotto, A., Lin, Z., Zhou, Z., Moon, S., Crook, P., Liu, B., Yu, Z., Cho, E., Fung, P., and Wang, Z. (2021). Continual learning in task-oriented dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7452–7467, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Mairesse, F. and Walker, M. (2007). PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 496–503, Prague, Czech Republic, June. Association for Computational Linguistics.

Majumder, N., Poria, S., Hazarika, D., Mihalcea, R., Gelbukh, A., and Cambria, E. (2019). Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825, Jul.

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

Ortony, A., Clore, G. L., and Collins, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, et al., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pennington, J., Socher, R., and Manning, C. (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, October. Association for Computational Linguistics.

Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. (2019). MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July. Association for Computational Linguistics.

Poria, S., Majumder, N., Hazarika, D., Ghosal, D., Bhardwaj, R., Jian, S. Y. B., Hong, P., Ghosh, R., Roy, A., Chhaya, N., Gelbukh, A., and Mihalcea, R. (2021). Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, Sep.

Preoţiuc-Pietro, D., Schwartz, H. A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., and Shulman, E. (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, June. Association for Computational Linguistics.

Riccardi, G. and Hakkani-Tür, D. Z. (2005). Grounding emotions in human-machine conversational systems. In *INTETAIN*.

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

Saha, T., Saha, S., and Bhattacharyya, P. (2020). Towards sentiment aided dialogue policy learning for multi-intent conversations using hierarchical reinforcement learning. *PLOS ONE*, 15(7):1–28, 07.

Shi, W. and Yu, Z. (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, July. Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (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, October. Association for Computational Linguistics.

Ultes, S., Budzianowski, P., Casanueva, I., Mrksic, N., Rojas-Barahona, L. M., hao Su, P., Wen, T.-H., Gašić, M., and Young, S. J. (2017). Domain-independent user satisfaction reward estimation for dialogue policy learning. In *INTERSPEECH*.

Vinyals, O. and Le, Q. (2015). A neural conversational model. *ICML Deep Learning Workshop, 2015*, 06.

Wang, J., Wang, J., Sun, C., Li, S., Liu, X., Si, L., Zhang, M., and Zhou, G. (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, Apr.

Young, S., Gašić, M., Keizer, S., Mairesse, F., Schatzmann, J., Thomson, B., and Yu, K. (2010). The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

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

Zahiri, S. and Choi, J. D. (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.

Zhou, X. and Wang, W. Y. (2018). MojiTalk: Generating emotional responses at scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1128–1137, Melbourne, Australia, July. Association for Computational Linguistics.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Zhu, Q., Zhang, Z., Fang, Y., Li, X., Takanobu, R., Li, J., Peng, B., Gao, J., Zhu, X., and Huang, M. (2020). ConvLab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems. In Asli Celikyilmaz et al., editors, *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.

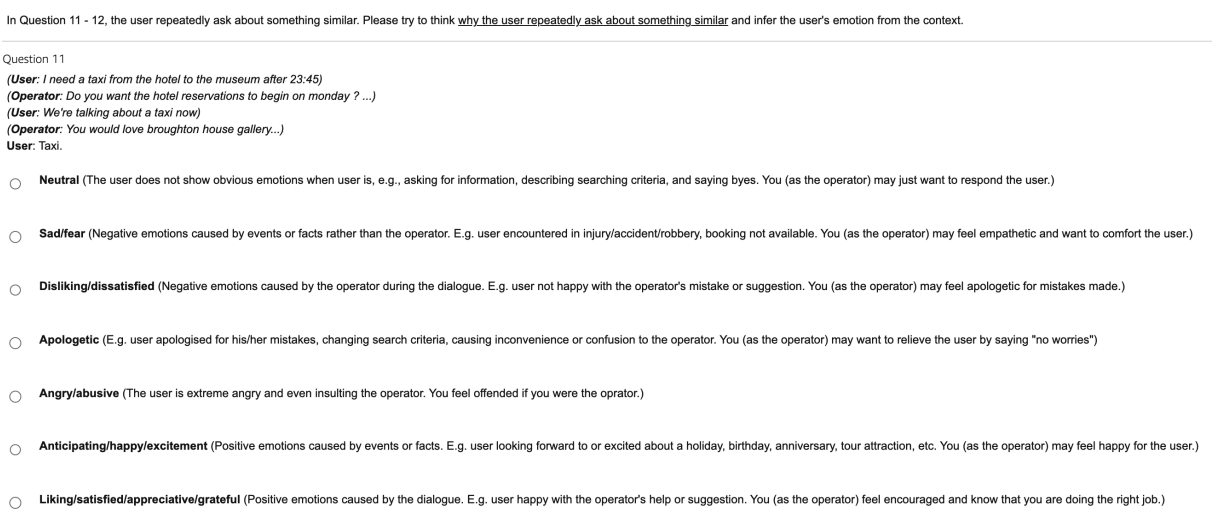
Elicitor	Aspects of events or agents			OCC Emotion
Consequences of Events	Consequence for other	Desirable for other		happy-for resentment
		Undesirable for other		gloating pity
	Consequence for self	prospects relevant	confirmed	satisfaction fears-confirmed
			disconfirmed	relief disappointment
		prospects irrelevant		joy distress
	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.


In Question 11 - 12, the user repeatedly ask about something similar. Please try to think why the user repeatedly ask about something similar and infer the user's emotion from the context.

Question 11

(**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.

- **Neutral** (The user does not show obvious emotions when user is, e.g., asking for information, describing searching criteria, and saying byes. You (as the operator) may just want to respond the user.)
- **Sad/fear** (Negative emotions caused by events or facts rather than the operator. E.g. user encountered in injury/accident/robbery, booking not available. You (as the operator) may feel empathetic and want to comfort the user.)
- **Disliking/dissatisfied** (Negative emotions caused by the operator during the dialogue. E.g. user not happy with the operator's mistake or suggestion. You (as the operator) may feel apologetic for mistakes made.)
- **Apologetic** (E.g. user apologised for his/her mistakes, changing search criteria, causing inconvenience or confusion to the operator. You (as the operator) may want to relieve the user by saying "no worries")
- **Angry/abusive** (The user is extreme angry and even insulting the operator. You feel offended if you were the oprator.)
- **Anticipating/happy/excitement** (Positive emotions caused by events or facts. E.g. user looking forward to or excited about a holiday, birthday, anniversary, tour attraction, etc. You (as the operator) may feel happy for the user.)
- **Liking/satisfied/appreciative/grateful** (Positive emotions caused by the dialogue. E.g. user happy with the operator's help or suggestion. You (as the operator) feel encouraged and know that you are doing the right job.)

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.

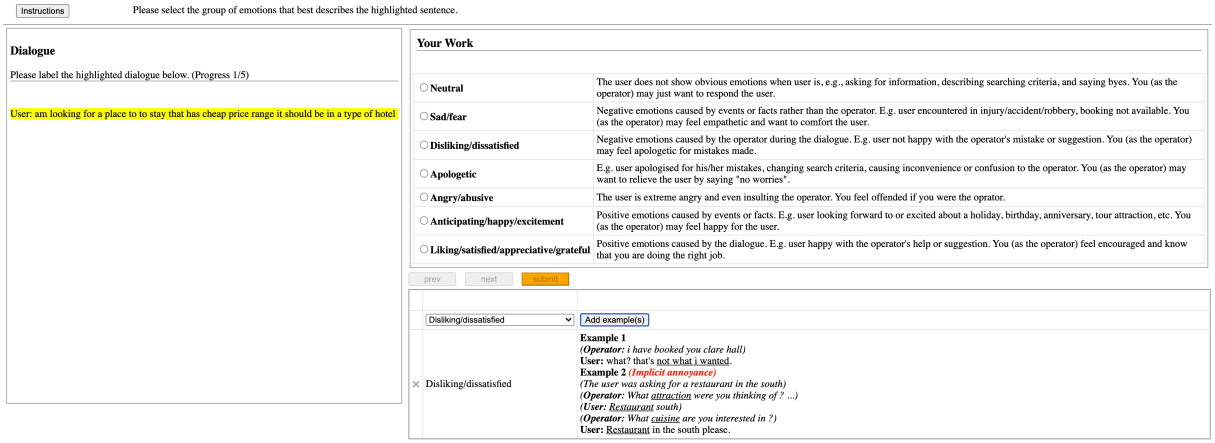


Figure B.2: Amazon Mechanical Turk main task page

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 (DialMAGE)	
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
DialMAGE	0.465	259	4,773	7,061
EmoWOZ (overall)	0.602	1,275	22,043	60,299

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	59.75	0	**50.34**	0	12.99	61.42	**72.43**	**52.50**	32.86	**51.45**	59.08	36.70	55.94
	M → D	**71.57**	**11.67**	1.36	**100**	6.15	**64.30**	68.85	16.97	42.05	9.36	56.94	46.27	43.02
	E → D	69.94	0	41.43	60.0	**29.41**	56.01	69.13	45.47	**42.66**	43.82	**61.09**	**46.56**	**57.95**
	D → M	71.09	0	6.02	0	11.11	15.60	88.07	62.63	20.13	77.16	70.58	27.41	72.77
	M → M	**95.34**	**43.00**	**40.87**	**73.03**	19.05	**40.45**	**90.39**	**85.19**	**51.13**	**84.82**	**92.57**	**57.45**	**92.43**
	E → M	92.67	41.43	27.76	70.35	**21.43**	39.98	89.44	79.79	48.4	83.19	88.88	54.72	90.05
	E → E	89.75	36.17	35.10	70.38	27.50	42.89	88.79	73.67	50.14	73.55	84.82	55.80	84.83
ContextBERT	D → D	80.16	0	**75.69**	0	5	58.58	**71.69**	**73.91**	35.16	72.85	77.19	41.59	76.81
	M → D	72.11	**11.67**	7.73	**43.71**	11.87	**60.07**	66.29	21.29	33.56	14.49	57.80	39.06	45.67
	E → D	**81.58**	5.00	75.46	40.00	**52.81**	57.31	69.23	73.71	**49.97**	**73.49**	**77.89**	**54.48**	**77.87**
	D → M	89.37	0	11.18	0	0	13.86	77.07	59.43	17.02	67.81	80.44	27.35	83.40
	M → M	**95.09**	**35.71**	**36.35**	**70.34**	**19.44**	34.05	**90.01**	**84.36**	**47.65**	**83.87**	**92.14**	**54.43**	**91.98**
	E → M	93.45	33.70	30.39	62.42	17.27	**37.06**	89.75	80.44	45.10	83.14	89.65	52.00	90.60
	E → E	92.10	30.08	61.69	62.36	41.73	40.83	89.14	78.99	54.30	79.67	87.93	59.70	88.33
DialogueRNN (GloVe)	D → D	40.13	0	**64.01**	0	0	52.05	65.59	**62.56**	30.28	**62.03**	**54.88**	31.68	**50.18**
	M → D	**67.00**	0	22.91	**100**	0	54.45	55.96	31.03	38.89	26.23	54.07	**42.90**	48.30
	E → D	23.83	0	61.75	60	0	**63.72**	**73.57**	62.06	**43.17**	61.23	50.24	40.41	40.99
	D → M	59.78	0	5.34	0	0	13.80	85.57	50.51	17.45	74.87	55.46	23.50	63.96
	M → M	87.25	**21.57**	**21.53**	52.16	0	26.21	85.51	72.78	**34.50**	78.12	82.21	**42.03**	84.72
	E → M	**88.24**	13.59	18.67	**57.56**	0	**27.92**	**86.73**	**74.04**	34.08	**79.22**	**83.41**	41.81	**85.74**
	E → E	83.46	12.71	51.38	57.67	0	32.75	86.35	70.93	40.14	74.56	78.56	46.33	80.76
DialogueRNN (BERT)	D → D	65.24	0	58.24	0	27.69	54.51	**68.35**	58.45	34.80	57.97	**61.95**	39.15	**61.90**
	M → D	**66.63**	0	4.24	43.52	2.86	41.48	53.87	17.33	24.33	9.64	50.73	30.37	40.48
	E → D	49.81	0	**61.01**	**91.67**	**28.14**	**60.92**	66.70	**60.95**	**51.41**	**60.56**	56.58	**51.18**	54.74
	D → M	85.48	0	8.17	0	6.71	20.48	87.46	65.74	20.47	76.91	78.71	29.76	83.11
	M → M	**92.11**	34.83	**34.49**	58.59	0	26.32	87.48	**79.18**	40.28	80.84	**88.13**	47.69	**88.99**
	E → M	90.54	**47.12**	18.08	**71.22**	**15.48**	**33.92**	**88.28**	76.26	**45.68**	**81.52**	86.10	**52.09**	88.04
	E → E	86.85	41.32	47.51	71.48	25.56	39.42	87.58	72.48	52.15	75.50	81.78	57.10	83.41
COSMIC	D → D	69.34	0	59.68	0	0	64.30	**72.25**	60.31	32.71	59.25	**65.07**	37.94	**64.71**
	M → D	**71.56**	33.33	2.67	100	15.38	67.04	70.80	19.98	48.21	11.03	57.16	51.54	43.79
	E → D	66.59	0	**61.47**	**100**	**43.71**	**69.74**	68.19	**62.09**	**57.18**	**61.67**	64.47	**58.53**	64.33
	D → M	86.68	0	8.78	0	0	20.91	88.90	67.85	19.77	78.19	80.32	29.32	84.33
	M → M	**94.86**	50	**40.97**	67.12	0	41.77	**89.93**	**84.22**	**48.30**	**84.27**	**91.81**	**54.95**	**91.93**
	E → M	92.61	**58.18**	24.68	**70.52**	**0**	**37.92**	89.10	79.84	46.73	82.74	88.81	53.29	89.88
	E → E	89.80	51.98	50.69	70.93	31.62	44.42	88.42	75.89	56.34	77.09	85.26	61.12	85.94

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 DialMAGE. M → D, for example, means to train on MultiWOZ and test on DialMAGE. Extreme values for "Apologetic" and "Abusive" in DialMAGE ("* → D"s) are caused by their rarity in the test set (1 and 5 occurrences respectively).

F.2. Confusion Matrix of ContextBERT

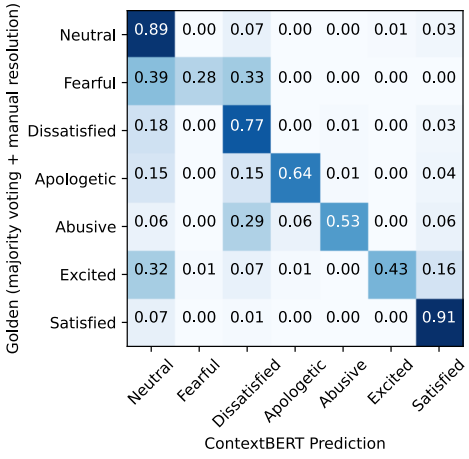


Figure F.1: Confusion matrix between Golden Labels and (the best) ContextBERT Prediction.

F.3. Sentiment Classification (3 classes)

Model	Feature	Ctx.	F1 of Each Sentiment in EmoWOZ			Average F1 w/o Neutral					
						EmoWOZ		MultiWOZ		DialMAGE	
			Neutral	Negative	Positive	Macro	Weighted	Macro	Weighted	Macro	Weighted
BERT	BERT	No	89.9	42.0	88.3	65.1	75.6	67.0	84.6	54.8	44.7
ContextBERT	BERT	Yes	**92.7**	**68.0**	87.5	**77.8**	**82.2**	**70.2**	**84.7**	**71.2**	**75.3**
DialogueRNN	GloVe	Yes	86.8	61.0	83.7	72.4	77.5	66.5	80.6	69.0	66.7
DialogueRNN	BERT	Yes	83.3	47.2	87.4	67.3	76.4	57.0	81.8	67.2	66.1
COSMIC	BERT+COMET	Yes	89.0	48.2	**88.5**	68.3	77.5	63.0	84.0	63.7	59.1

Table F2: Summarised performance of baseline models on sentiment classification. "Ctx." stands for "context".

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 → D	57.61	**61.32**	**72.48**	**62.45**	**66.90**	**62.61**	60.54	**63.80**	**59.91**
	M → D	**71.70**	3.80	68.47	17.78	36.13	11.27	57.39	47.99	43.93
	E → D	70.01	41.73	67.84	46.16	54.78	44.74	**61.68**	59.86	58.39
	D → M	71.97	7.77	84.71	56.23	46.24	76.94	65.82	54.82	73.35
	M → M	**95.43**	**56.68**	**90.05**	**87.00**	**73.37**	**86.69**	**93.11**	**80.72**	**92.99**
	E → M	92.90	44.99	89.08	81.99	67.04	84.63	89.66	75.66	90.60
	E → E	89.93	41.96	88.27	75.74	65.11	75.60	85.57	73.39	85.56
ContextBERT	D → D	80.67	**76.74**	**70.81**	**76.08**	**73.77**	**76.05**	78.53	**76.07**	78.55
	M → D	71.69	6.65	64.22	20.02	35.43	13.30	57.27	47.52	44.85
	E → D	**82.91**	76.60	65.71	75.25	71.15	75.34	**79.48**	75.07	**79.43**
	D → M	89.04	17.89	69.53	54.85	43.71	64.32	79.01	58.82	82.16
	M → M	**95.38**	**55.09**	**89.89**	**86.81**	**72.49**	**86.38**	**93.03**	**80.12**	**92.88**
	E → M	93.98	51.94	88.38	83.75	70.16	84.70	91.05	78.10	91.40
	E → E	92.69	67.98	87.55	81.95	77.76	82.19	89.36	82.74	89.49
DialogueRNN (GloVe)	D → D	58.50	**69.49**	67.55	**69.21**	68.52	69.26	**64.98**	**65.18**	**63.45**
	M → D	**68.70**	20.44	52.21	29.70	36.32	24.11	55.21	47.12	48.21
	E → D	42.57	66.00	**71.96**	66.56	**68.98**	**66.68**	58.16	60.17	53.65
	D → M	77.27	10.88	**84.93**	62.22	47.90	77.46	71.38	57.69	77.32
	M → M	**91.06**	**50.36**	84.13	**80.22**	**67.25**	**80.72**	**87.57**	**75.19**	**88.18**
	E → M	90.71	48.85	84.21	79.77	66.53	80.64	87.14	74.59	87.91
	E → E	86.75	61.03	83.74	76.42	72.39	77.53	82.91	77.18	83.94
DialogueRNN (BERT)	D → D	47.53	65.27	**69.53**	65.74	**67.40**	65.76	59.05	60.78	55.91
	M → D	**71.10**	7.04	64.41	20.25	35.72	13.67	56.77	47.52	44.70
	E → D	50.18	**65.81**	68.63	**66.10**	67.22	**66.14**	**60.04**	**61.54**	**57.52**
	D → M	59.37	8.08	86.36	51.21	47.22	78.46	55.78	51.27	64.68
	M → M	**93.75**	**52.28**	**88.13**	**84.27**	**70.20**	**84.52**	**90.94**	**78.05**	**91.18**
	E → M	86.61	25.98	88.11	73.03	57.04	81.84	81.92	66.90	85.29
	E → E	83.30	47.16	87.36	71.40	67.26	76.36	78.72	72.61	81.19
COSMIC	D → D	63.51	**68.32**	**70.52**	**68.58**	**69.42**	**68.57**	**66.45**	**67.45**	**65.83**
	M → D	**71.69**	5.22	67.57	18.79	36.39	12.42	57.32	48.16	44.45
	E → D	69.28	57.68	69.66	59.30	63.67	59.06	64.80	65.54	64.58
	D → M	80.31	12.43	86.38	62.62	49.41	78.92	73.66	59.71	79.92
	M → M	**95.01**	**57.85**	**89.48**	**86.41**	**73.67**	**86.29**	**92.58**	**80.78**	**92.58**
	E → M	91.63	36.68	89.32	79.78	63.00	84.01	87.99	72.54	89.51
	E → E	88.97	48.15	88.55	75.66	68.35	77.5	84.6	75.22	85.47

Table F3: Detailed results of baseline models on sentiment classification including cross-dataset experiments. For cross-dataset experiments, the "X → 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 DialMAGE. M → D, for example, means to train on MultiWOZ and test on DialMAGE.

F.4. Change in precision and recall on MultiWOZ after Complementing MultiWOZ with DialMAGE in Training

		Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied
ContextBERT	Recall	**95.3 → 91.5**	**34.7 → 28.0**	**31.4 → 60.4**	69.7 → 61.9	16.0 → 24.0	33.5 → 34.1	90.4 → 90.0
	Precision	**94.9 → 95.5**	**37.3 → 44.8**	**43.7 → 20.9**	71.4 → 63.5	25.0 → 13.7	**35.5 → 42.6**	89.5 → 89.6
	F1	**95.1 → 93.5**	35.7 → 33.7	36.4 → 30.4	**70.3 → 62.4**	19.4 → 17.3	34.0 → 37.1	90.0 → 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 DialMAGE in training. A is the value when trained on MultiWOZ and B is the value when trained on EmoWOZ. Values with statistical significance ($p < 0.1$) 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*, *fearful*, *apologetic*, *abusive*, *excited*, and *satisfied* where as a high recall is more desirable for *dissatisfied*.

F.5. Emotion Distribution in Model Predictions

Test Set	Model	Neutral	Fearful	Dissatisfied	Apologetic	Abusive	Excited	Satisfied
MultiWOZ Label		72.31	0.2	1.47	0.98	0.07	1.0	23.97
DialMAGE (#token>11.8) Label		61.96	0.61	28.83	0.61	0.0	6.75	1.23
DialMAGE (#token > 11.8) Prediction	BERT	89.57	0.0	2.45	0.0	0.0	6.75	1.23
	ContextBERT	58.13	0.0	35.47	0.0	0.0	4.93	1.48
	DialogueRNN-GloVe	6.52	0.0	74.97	1.28	0.0	6.05	11.18
	DialogueRNN-BERT	55.65	0.12	23.05	8.61	0.0	5.01	7.57
	COSMIC	61.93	0.23	18.63	7.57	0.0	3.84	7.8
DialMAGE Label		54.12	0.24	39.3	0.08	0.95	1.35	3.96
MultiWOZ (#token<5.8) Label		60.76	0.0	1.21	0.0	0.0	0.3	37.73
MultiWOZ (#token < 5.8) Prediction	BERT	60.91	0.15	0.45	0.3	0.0	0.45	37.73
	ContextBERT	57.43	0.0	2.97	0.0	0.0	0.66	38.94
	DialogueRNN-GloVe	46.54	0.0	2.1	0.7	0.0	0.88	49.78
	DialogueRNN-BERT	46.28	0.0	8.94	0.26	0.09	1.67	42.77
	COSMIC	49.43	0.09	5.0	0.18	0.09	1.58	43.65

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