A Large-Scale Dataset for Empathetic Response Generation

Anuradha Welivita, Yubo Xie and Pearl Pu
School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne
Switzerland
{kalpani.welivita,yubo.xie,pearl.pu}@epfl.ch

Abstract

Recent development in NLP shows a strong trend towards refining pre-trained models with a domain-specific dataset. This is especially the case for response generation where emotion plays an important role. However, existing empathetic datasets remain small, delaying research efforts in this area, for example, the development of emotion-aware chatbots. One main technical challenge has been the cost of manually annotating dialogues with the right emotion labels. In this paper, we describe a large-scale silver dataset consisting of 1M dialogues annotated with 32 fine-grained emotions, eight empathetic response intents, and the *Neutral* category. To achieve this goal, we have developed a novel data curation pipeline starting with a small seed of manually annotated data and eventually scaling it to a satisfactory size. We compare its quality against a state-of-the-art gold dataset using offline experiments and visual validation methods. The resultant procedure can be used to create similar datasets in the same domain as well as in other domains.[1]

1 Introduction

Researchers are increasingly inclined towards refining pre-trained language models with domain-specific datasets to achieve certain tasks (Devlin et al., 2019; Liu et al., 2019; Rashkin et al., 2018). One such area is the development of empathetic conversational agents that can understand human emotions and respond appropriately. The aim of the empathetic response generation task is to generate syntactically correct, contextually relevant, and more importantly emotionally appropriate responses following previous dialogue turns. Such tasks require the creation and availability of large dialogue datasets, in which each utterance is annotated with the correct intents and emotions. Though many such datasets have been developed in the past (Busso et al., 2008; Poria et al., 2019; Li et al., 2017; Rashkin et al., 2018), due to the cost of manual labor, they are limited in size, thus insufficient to train robust conversational agents. Since collecting and manually annotating such gold standard data is expensive, replacing them with automatically annotated silver standard data has become a rising interest (Filannino and Di Bari, 2015). We show how such a large-scale silver standard dataset with sufficient quality can be curated and used to fine-tune pre-trained language models for the generation of empathetic responses.

Emotions revealed in social chitchat are rather complex. It has many categories of emotions to distinguish due to subtle variations present in human emotion. For example, *Sadness* and *Disappointment* are pursued and dealt with differently in human conversations even though both of them are negative emotions. Also, the listener's reaction to emotion is not always a straightforward mirroring effect of the speaker's emotion. Rather it can be more neutral and convey a specific intent, as is evident from the dialogue example in Table 1.

Speaker:	*I've been hearing some strange noises around the house at night.* **(Afraid)**
Listener:	*oh no! That's scary! What do you think it is?* **(Neutral: Acknowledging; Questioning)**
Speaker:	*I don't know, that's what's making me anxious.* **(Anxious)**
Listener:	*I'm sorry to hear that.* **(Neutral: Sympathizing)**

Table 1: An example showing the listener's reactions to emotions do not always mirror the speaker's emotions.

Welivita and Pu (2020) have analyzed listener responses in the EmpatheticDialogues dataset (Rashkin et al., 2018) and discovered eight listener specific empathetic response intents contained in emotional dialogues: *Questioning*; *Agreeing*; *Acknowledging*; *Sympathizing*; *Encouraging*; *Consoling*: *Suggesting*; and *Wishing*. They have annotated the EmpatheticDialogues dataset with 32 fine-grained emotions, eight empathetic response

[1]The datasets and the code are publicly accessible at https://github.com/anuradha1992/EDOS.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1251–1264
November 7–11, 2021.

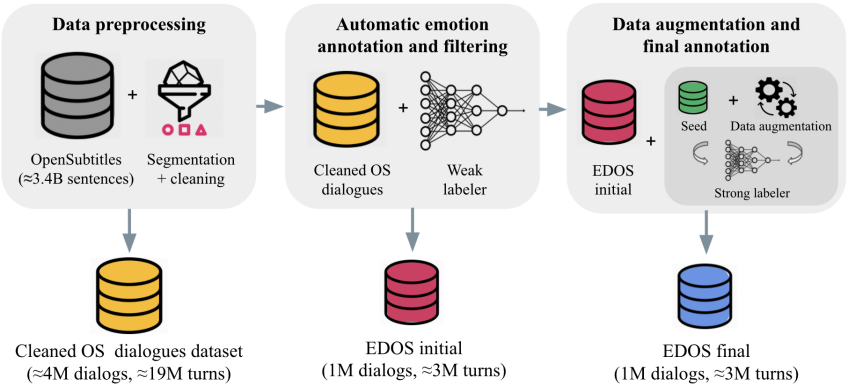


Figure 1: Steps for curating the EDOS dataset.

intents, and the *Neutral* category, and discovered frequent emotion-intent exchange patterns in empathetic conversations. They observe that this type of dataset tagged with fine-grained emotions and intents can be used to train neural chatbots to generate empathetically appropriate responses. But for this purpose, a large-scale emotion and intent labeled dataset is even more desirable. Curating such a dataset is technically challenging since 1) annotating such a large-scale dataset require costly human labor, and 2) given the fine-granularity of the emotion and intent labels, the human labeling task is more difficult and error-prone compared to the more coarse grained *Angry-Happy-Sad* emotion categories. As a result, existing manually labeled emotional dialogue datasets such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and DailyDialogue (Li et al., 2017) are smaller in scale and contain only a limited set of emotions (emotions derived from basic emotion models such as the Ekman's). Most importantly, existing datasets fail to distinguish between *Neutral* and *Questioning*, or any of the other eight empathetic response intents. They combine everything into a big label *Neutral* or *Other* when the utterance is not emotional. But *Questioning*, *Agreeing*, *Acknowledging*, *Sympathizing*, *Encouraging*, *Consoling*, *Suggesting*, and *Wishing* are important details in constructing empathetic dialogues. These eight response intents, which we call the plus categories, are novel in our work and contribute to the model's learning of important response patterns in the data.

To fill the above gap, we curate a novel large-scale silver dialogue dataset, **EDOS** (**E**motional **D**ialogues in **O**pen**S**ubtitles), containing 1M emotional dialogues from movie subtitles, in which each dialogue turn is automatically annotated with 32 fine-grained emotions, eight plus categories as well as the *Neutral* category. Movie subtitles are extensively used for emotion analysis in text in earlier and recent research (Kayhani et al., 2020; Merdivan et al., 2020; Giannakopoulos et al., 2009). The Nature article "How movies mirror our mimicry" (Ball, 2011) states *"screenwriters mine everyday discourse to make dialogues appear authentic"* and *"audiences use language devices in movies to shape their own discourse"*. Hence, it can be one of the major sources to train chatbots and learn emotional variations and corresponding response strategies in dialogues. To reduce the cost of human labeling and the complexity of labeling dialogues with fine-grained emotions and intents, we devised a semi-automated human computation task to collect fine-grained emotion and intent labels for a small set of movie dialogues (9K). We then followed automatic data augmentation techniques to expand the labeled data and trained a dialogue emotion classifier to automatically annotate 1M emotional dialogues.

The process of curating the dataset involved several stages. First, we applied automatic turn and dialogue segmentation methods, data cleaning and removal of duplicates on movie subtitles in the OpenSubtitles (OS) corpus (Lison et al., 2019) and obtained close to 4M dialogues. Then, we applied a weak labeler (a BERT-based sentence-level classifier) trained on the EmpatheticDialogues dataset (Rashkin et al., 2018), to label utterances in OS dialogues and filtered 1M emotional dialogues (EDOS initial). Thereafter, we applied data augmentation techniques on a small set of human-annotated data and used the manually annotated and extended labels to train a strong labeler that is used to annotate dialogues in EDOS initial and obtained the final 1M EDOS dataset. We evaluated the quality of the resultant dataset by comparing it against the

Dataset	Labels	No. of dialogues	No. of utterances	Publicly available
IEMOCAP (Busso et al., 2008)	*Joy*, *Sadness*, *Anger*, *Frustrated*, *Excited*, and *Neutral*	151	7,433	✓
MELD (Poria et al., 2019)	*Joy*, *Surprise*, *Sadness*, *Anger*, *Disgust*, *Fear*, and *Neutral*	1,433	13,708	✓
DailyDialogue (Li et al., 2017)	*Joy*, *Surprise*, *Sadness*, *Anger*, *Disgust*, *Fear*, and *Neutral*	12,218	103,607	✓
EmotionLines (Hsu et al., 2018)	*Joy*, *Surprise*, *Sadness*, *Anger*, *Disgust*, *Fear*, and *Neutral*	1,000	14,503	✓
EmoContext (Chatterjee et al., 2019)	*Joy*, *Sadness*, *Anger*, and *Other*	38,421	115,263	✓
Twitter customer support (Herzig et al., 2016)	Customer emotions: *Confusion*; *Frustration*; *Anger*; *Sadness*; *Happiness*; *Hopefulness*; *Disappointment*; *Gratitude*; *Politeness*; and Agent emotional techniques: *Empathy*; *Gratitude*; *Apology*; *Cheerfulness*	2,413	$\approx$ 14,078	✗
Empathetic Dialogues (Rashkin et al., 2018; Welivita and Pu, 2020)	32 fine-grained emotions (postive and negative), *Neutral*, and 8 empathetic response intents: *Questioning*; *Agreeing*; *Acknowledging*; *Sympathizing*; *Encouraging*; *Consoling*; *Suggesting*; and *Wishing*.	24,850	107,220	✓
EDOS	32 fine-grained emotions, 8 empathetic response intents, and *Neutral*.	1*M*	3,488,300	✓

Table 2: Comparison of emotion annotated dialogue datasets available in the literature against EDOS.

EmpatheticDialogues dataset by means of offline experiments and visual validation methods. Figure 1 summarizes the process of creating EDOS. The data curation pipeline we followed substantially reduced the cost of human labor while ensuring quality annotations.

Our contributions in this paper are three-fold. 1) We curate a large-scale dialogue dataset, EDOS, containing 1M emotional dialogues labeled with 32 fine-grained emotions, eight empathetic response intents (the plus categories), and *Neutral*. Compared to existing dialogue datasets tagged with emotions, EDOS is significantly larger ($\approx$ 40 times larger than EmpatheticDialogues), and contains more fine-grained emotions and empathetic response strategies. 2) We outline the complex pipeline used to derive this dataset. 3) We evaluate the quality of the dataset compared to a state-of-the-art gold standard dataset using offline experiments and visual validation methods.

2 Literature review

IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), DailyDialogue (Li et al., 2017), EmotionLines (Hsu et al., 2018), and EmoContext (Chatterjee et al., 2019) are some existing state-of-the-art dialogue datasets with emotion labels. However, these datasets are limited in size and are labeled with only a small set of emotions without any response strategies. Table 2 shows a summary of the size and the labels in these datasets. All the datasets compared here are in the English language.

Herzig et al. (2016) detected customer emotions and agent emotional techniques (e.g., *Apology*, *Empathy*) in customer support dialogues. They curated a dialogue dataset from two customer support Twitter accounts and manually annotated the customer turns with one of 9 emotions and the agent turns with one of 4 emotional techniques. But emotions expressed by customers in social media service dialogues are mainly negative (e.g. *anger*, *frustration*), and the customer service agents also respond in a restricted manner, which limits the utility of this dataset, in addition to its small size.

The EmpatheticDialogues dataset (Rashkin et al., 2018) contains 25K open-domain dialogues grounded on 32 emotions. The 32 emotions range from basic emotions derived from biological responses (Ekman, 1992; Plutchik, 1984) to larger sets of subtle emotions derived from contextual situations (Skerry and Saxe, 2015). Welivita and Pu (2020) manually analyzed a subset of the listener turns in EmpatheticDialogues and identified eight listener-specific response intents. They developed a sentence-level weak labeler using which they annotated the entire dataset with 32 emotions, eight empathetic response intents, and the *Neutral* category. However, due to the limited size of EmpatheticDialogues, it is difficult to be used for data-intensive applications. To address the above limitations, we curate EDOS containing 1M movie dialogues. We label each dialogue turn with 32 emotions, eight empathetic response intents, and *Neutral* using our own dialogue emotion and intent classifier. Table 2 compares EDOS to state-of-the-art emotion annotated dialogue datasets.

3 Methodology

This section describes the dialogue selection process, the design of the human annotation task,

the data augmentation techniques used to expand human-labeled dialogues, and the development of a strong labeler to annotate the dataset.

3.1 Dialogue curation from movie subtitles

The OpenSubtitles 2018 corpus consists of 3.7M movie and TV subtitles. It comprises 3.4B sentences and 22.2B tokens. It is an excellent source to learn emotional variations in dialogue and corresponding response mechanisms. But due to the absence of speaker markers, movie subtitles do not contain an explicit dialogue turn structure (who speaks what) and specific indicators where one dialogue ends and the next dialogue begins. To overcome the first issue, we reproduced the work by Lison and Meena (2016) to build an SVM-based classifier that determines if two consecutive sentences are part of the same dialogue turn. Our classifier achieved a segmentation accuracy of 76.69%, which is close to the accuracy of 78% that the authors claim. The set of features that gave the best turn segmentation accuracy are: 1) unigram and bi-gram features of adjacent sentences after lemmatization; 2) first and final tokens of adjacent sentences; 3) first and final bi-grams of adjacent sentences; 4) whether the two sentences belong to the same subtitle block or not (boolean); 5) genre of the movie (*Drama*, *Crime*, *Musical* etc.); 6) sentence density of the subtitles file (no. of sentences/subtitle duration); and 7) quadratic combinations of the above features with itself and the rest.

After performing turn segmentation on the OpenSubtitles corpus, we divided the turns into separate dialogues based on a simple heuristic. If the difference between the end time of the previous turn and the start time of the current turn is more than 5 seconds, we take these two turns as belonging to 2 different dialogues. An exception occurs if this timestamp information is missing in at least one of the turns. In this case, we assume that these two turns appear in the same subtitle block and consider them as belonging to the same dialogue. This way, we formed 9M dialogues from the OpenSubtitles corpus altogether. The choice of 5 sec.s to separate dialogues is explained in Appendix C.

To further clean the dialogues, we removed character names, the repetitive dialogue turns, turns that start with "previous on..." (monologue at the beginning of TV episodes), turns with character length less than 2 or greater than 100, turns with an alphabetic proportion less than 60%, and turns with a lot of repetitive tokens. When a dialogue turn was removed, all the turns following that turn were also removed from the dialogue to maintain consistency. After that, all the dialogues left with only one turn were removed from the corpus. We removed dialogues from movies of the genre 'Documentary' since they do not correspond to actual dialogues. This resulted in a cleaned OS dialogue dataset consisting of 4M dialogues.

To filter out dialogues containing emotional statements and empathetic responses from the cleaned OS dialogues dataset, we employed a weak labeler, (a BERT transformer-based sentence level classifier) trained on 25K situation descriptions from EmpatheticDialogues (Rashkin et al., 2018) tagged with 32 emotion classes, and 7K listener utterances tagged with eight empathetic response intents and the *Neutral* category (Welivita and Pu, 2020). The classifier had a high top-1 classification accuracy of 65.88%. We call it a weak labeler since it predicts emotion or intent only at the sentence level and is trained on a different dataset other than OS. We filtered the top 1M dialogues having the highest label confidence as predicted by this classifier to form the 1M EDOS (initial) dataset. The statistics of the EDOS dataset are given in Table 3. More detailed statistics including the number of dialogues per emotion are included in Appendix D.

Criteria	Statistics
Total no. of dialogues	1,000,000
Total no. of turns	2,829,426
Total no. of tokens	39,469,825
Avg. no. of turns per dialogue	2.83
Avg. no. of tokens per dialogue	39.47
Avg. no. of tokens per turn	13.95

Table 3: Statistics of the EDOS dataset.

3.2 Human computation

To train a dialogue emotion classifier that can identify both fine-grained emotions and empathetic response intents, we devised an Amazon Mechanical Turk (AMT) experiment to collect an initial set of ground truth labels for OS dialogues. But annotating dialogue turns with one of 41 labels is a daunting task. To make the task less exhaustive, we devised a semi-automated approach using our weak labeler. By applying the weak labeler on each turn of the cleaned OS dialogue dataset, we filtered out the turns having prediction confidence ≥ 0.9, along with their dialogue history. Next, we ranked these dialogues according to their readability and selected the highest readable dialogues from each

class to be labeled. This is to reduce the time spent by the workers in having to read long and complicated dialogues. The steps followed in computing dialogues' readability are included in Appendix A. Workers had to select a label from the top-3 predictions made by the weak labeler. If none of the top-3 predictions matched, they could manually specify the correct class. The main purpose of incorporating a weak labeler here was to make the task less daunting for the crowd worker. Otherwise, having to choose a label out of 41 labels may lead to even worse results due to the complicated nature of the task. The risk of reduced data reliability is avoided by taking only the labels with the majority vote. The AMT task's user interface design is included in Appendix B.

After ranking the dialogues according to readability, we selected the top 250 dialogues in each category for the AMT task. We bundled 15 dialogues in a HIT with 5 quiz questions that served as checkpoints to evaluate the crowd workers' quality. Situation descriptions from the Empathetic-Dialogues dataset for which we already knew the emotion labels were used to formulate the quiz questions. Finally, we obtained dialogues where we had 2 out of 3 worker agreements, which resulted in $8,913$ dialogues altogether. Table 4 shows the results of the AMT task.

Description	Statistics
Total no. of dialogues	$10,250$
# dialogues labeled with majority vote	$8,913(86.96\%)$
Inter-annotator agreement (Fleiss' Kappa)	0.46 (moderate agreement)
% of times workers got 3/5 quiz questions correct	77.75%
# dialogues in which the workers manually specified the label	425

Table 4: AMT task results.

3.3 Data augmentation and annotation

To scale up the training data obtained from the AMT task, we utilized a distant learning technique using dialogue embeddings (Reimers and Gurevych, 2019) and self-labeling (Triguero et al., 2015), a semi-supervised learning technique. The first approach we used is using Sentence-BERT (SBERT) proposed by Reimers and Gurevych (2019), which uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity. Using this approach, we obtained semantically similar dialogues to those annotated by crowd workers and tagged them with the same class label. Among several models the authors have proposed, we used the *roberta-base-nli-stsb-mean-tokens* model, fine-tuned on the NLI (Bowman et al., 2015) and STS benchmark (STSb) (Cer et al., 2017) datasets, since it has reported a high Spearman's rank correlation of 84.79 ± 0.38 between the cosine-similarity of the sentence embeddings and the gold labels in the STS benchmark test set outperforming the existing state-of-the-art. It is also more efficient to use than *roberta-large*. Before proceeding, we left out 20% of the crowd-annotated dialogues, balanced across all class labels, as testing data. Then, we followed the following steps in extending the rest of the dialogues using SBERT.

1) Using the SBERT model, first, we computed dialogue turn embeddings (each with a vector representation of 768 dimensionalities) for all the turns ($\approx$19M) in the cleaned OS dataset. 2) Then, we calculated dialogue embeddings for human-annotated and unlabeled dialogues from the cleaned OS dialogues dataset. For this, we applied a decaying weight starting from the last turn and took the weighted average of the turn embeddings of each dialogue. We used half decaying, i.e, if we have a dialogue with turn embeddings v_1, v_2, and v_3, the final dialogue embedding would be $(4/7)v_3 + (2/7)v_2 + (1/7)v_1$. 3) Next, we calculated the cosine similarity between annotated and unlabeled dialogue embeddings and ranked the results. 4) Finally, we applied a similarity threshold and obtained all the unlabeled dialogues with a cosine similarity that exceeds this threshold and tagged them with the same crowd annotated class label. Here, we used a threshold of 0.92 after manually inspecting a random subset of the results obtained for a range of thresholds (Examples from this stage are denoted in Appendix C).

We extended the original crowd annotated dialogue dataset by $3,196$ more dialogues with distantly annotated class labels using the above method. Thereafter, using the crowd-annotated and extended labels, we trained an initial classifier that we used to annotate the rest of the dialogues and add more labels to our dataset that had annotation confidence over 0.9. This method is termed self-labeling (Triguero et al., 2015), a semi-supervised learning technique that can be used to grow labeled data. With this, we were able to extend the labeled data by $4,100$ more dialogues. Next, we again

applied SBERT over the self-labeled data and extended them by $2,118$ more dialogues. Finally, we were able to have $\approx 14K$ labeled dialogues altogether. We used this data to train a final dialogue emotion classifier to annotate the rest of the unlabeled data. This resulted in a classifier with precision 64.11%, recall 64.59%, macro F1-score 63.86%, and accuracy 65.00%, which is comparable with the state-of-the-art dialogue emotion classifiers (as denoted in Table 5). The design of the dialogue emotion classifier we utilized to annotate the dataset is explained in section 3.3.1.

3.3.1 Design of the dialogue emotion classifier

Our dialogue emotion classifier consists of a representation network that uses the BERT architecture, an attention layer that aggregates all hidden states at each time step, a hidden layer, and a softmax layer. We used the BERT-base architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters as the representation network. It was initialized with weights from RoBERTa (Liu et al., 2019). We fed in a dialogue turn along with the preceding context in the reverse order as input to the representation network. To give more importance to the dialogue turn for which prediction has to be made and the turns that immediately precede it, we multiplied the token embeddings belonging to each turn by a decreasing weight factor. Its input representation is constructed by summing the corresponding token embedding multiplied by the weighting factor and its position embedding. More details including the hyper-parameters used are included in the Appendix C.

4 EDOS quality analysis and comparison with the state-of-the-art gold standard

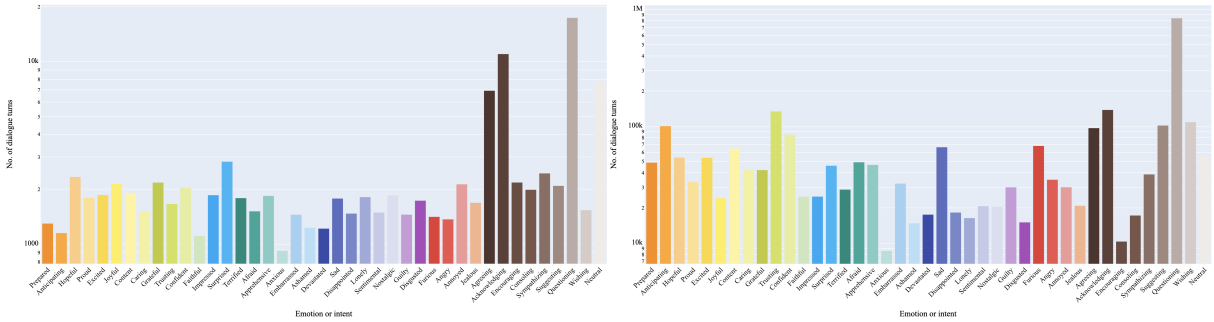
Table 6 shows some example dialogues taken from the EDOS dataset along with annotations and confidence scores. By observing the examples, it could be noticed that even for less confident predictions, the label quite accurately describes the emotion or intent of the corresponding dialogue turn.

We also conducted a qualitative comparison of the annotations in the EDOS dataset with EmpatheticDialogues (Rashkin et al., 2018; Welivita and Pu, 2020), a state-of-the-art gold standard dataset for empathetic conversations. Figure 2 compares the distributions of emotions and intents in the two datasets. It is observed that in both datasets, intent categories take prominence over individual emotion classes. This is in par with observations of Welivita and Pu (2020), where they notice that one or more intents from the taxonomy of empathetic intents are mostly utilized when responding to emotions in dialogue, rather than similar or opposite emotions. Especially, the intent *Questioning* takes the highest percentage among the annotations in EmpatheticDialogues and EDOS. We also computed the KL-divergence (≥ 0) of the emotion and intent distribution of EDOS with respect to that of EmpatheticDialogues, which measures how one probability distribution is different from a second, reference probability distribution (Kullback and Leibler, 1951). It resulted in a KL-divergence value of 0.2447, which indicates a considerable similarity between the two distributions (the lower the KL divergence, the more similar the distributions are).

Figure 3 compares the emotion-intent flow patterns in EmpatheticDialogues and EDOS. In the visualization corresponding to EmpatheticDialogues, the 1st and 3rd dialogue turns correspond to the speaker and the 2nd and 4th dialogue turns correspond to the listener. However, in EDOS, we cannot distinguish the dialogue turns as speaker and listener turns due to the absence of speaker annotations. Though this is the case, we could still observe some conversational dynamics present in EmpatheticDialogues are preserved in EDOS. For example, in both datasets, the speaker mostly starts the conversation with some emotional statement and in the subsequent turn, the response tends to be of the intent *Questioning*. In both datasets, intents *Agreeing* and *Acknowledging* follow emotions seen in the first turn irrespective of whether they are positive or negative. As the dialogues proceed, it could be seen in both datasets the emotions deescalate as more empathetic response intents emerge.

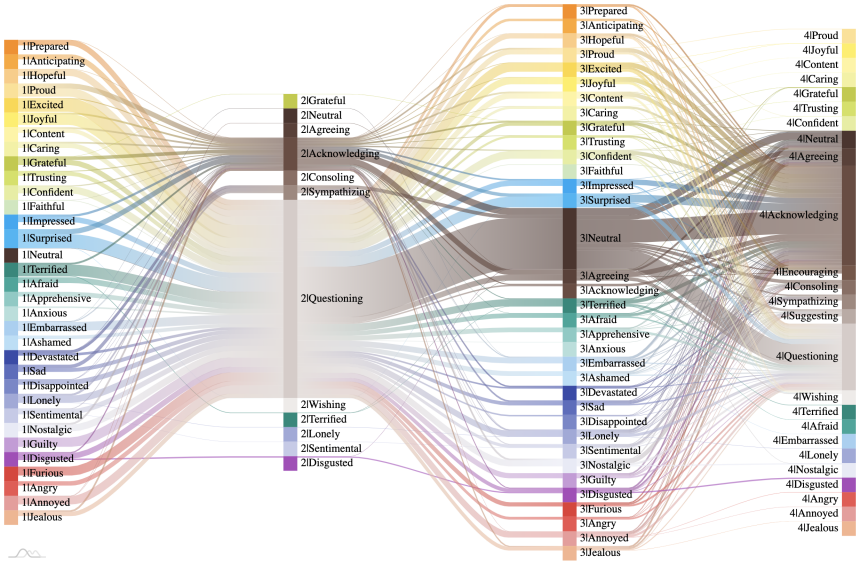
5 Experimental baselines

We propose some experimental baselines using the curated dataset for empathetic response generation and compare the performance against a dialogue model trained on the EmpatheticDialogues dataset. For this purpose, we trained a transformer (Vaswani et al., 2017) model with various training settings. Specifically, the following datasets were involved: **1) OS dialogues** (As described in Section 3.1, these dialogues were obtained by segmenting the movie subtitles. Note that for the purpose of pre-training, we excluded the EDOS dialogues, resulting in around 3M dialogues.); **2) EDOS** (1M dialogues); and **3) EmpatheticDialogues** (25K dialogues). All

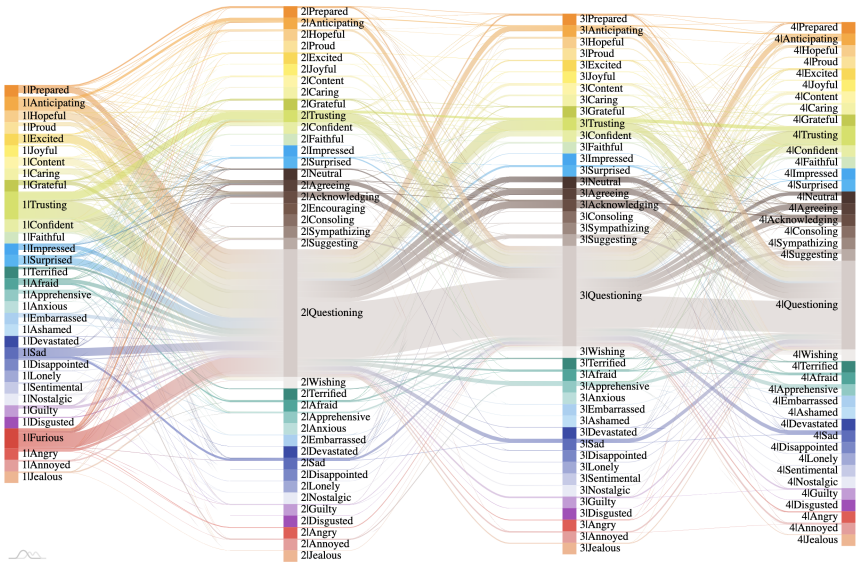


(a) EmpatheticDialogues (b) EDOS

Figure 2: Comparison of distribution of emotions and intents in the EmpatheticDialogues and EDOS datasets.



(a) EmpatheticDialogues dataset



(b) EDOS dataset

Figure 3: Comparison of emotion-intent flow patterns in the EmpatheticDialogues and EDOS datasets. For simplicity, only the first four dialogue turns are visualized.

Classifier	Dataset	No. of labels	F1	Acc.
AR (Khosla, 2018)	EmotionLines dataset (Hsu et al, 2018)	4 Emotion labels	–	Friends: 62.50 EmotionPush: 62.48
CMN (Hazarika et al., 2018b)	IEMOCAP dataset (Busso et al., 2008)	6 Emotion labels	56.13	56.56
ICON (Hazarika et al., 2018a)			57.90	58.30
IAAN (Yeh et al., 2019)			–	64.70
Dialog-RNN (Majumder et al., 2019)	IEMOCAP (Busso et al., 2008) and AVEC (Schuller et al., 2012) datasets	IEMOCAP: 4 Emotion labels; AVEC: 4 dimentional emotion labels	62.75	63.40
Dialog-GCN (Ghosal et al., 2019)	IEMOCAP (Busso et al., 2008), AVEC (Schuller et al., 2012), and MELD (Poria et al., 2019) datasets	IEMOCAP: 4 Emotion labels; AVEC: 4 dimentional emotion labels; MELD: 7 Emotion labels	64.18	65.25
Ours	OS dialogue dataset	32 Emotions + 8 Intents + Neutral	63.86	65.00

Table 5: Comparison of the performance of the dialogue emotion classifier used for annotation with performance of the state-of-the-art dialogue emotion classifiers. F1-score reported here is the macro-F1 score.

Dialogue #1:	
Turn 1	**(Excited, 0.98)** The concert will start soon.
Turn 2	**(Questioning, 0.01)** Are you excited?
Turn 3	**(Proud, 0.99)** I am. Because one of my friends made his efforts to make the concert happen. He wanted to fulfill a promise he made to his first love.
Turn 4	**(Sentimental, 0.99)** I like their story very much. I want to dedicate this concert to everyone who has truly loved someone.
Dialogue #2:	
Turn 1	**(Apprehensive, 0.89)** Staying here might not be safe.
Turn 2	**(Questioning, 0.41)** Take the earliest flight tomorrow?
Turn 3	**(Caring, 0.94)** Take Josie to mother. My home is where you are.
Turn 4	**(Faithful, 0.86)** We're not leaving.

Table 6: Example dialogues from the EDOS dataset along with annotations and confidence scores.

three datasets were split into a training (80%), validation (10%), and test (10%) sets. Based on the training strategies, we have the following models: **1) Pre-trained**—to take advantage of transfer learning, we pre-trained the transformer model on the 3M OS dialogues. The large scale of this training set is expected to provide a good starting point for fine-tuning; **2) Fine-tuned**—we took the pre-trained transformer and then fine-tuned it on EDOS and EmpatheticDialogues datasets respectively. All the models have 4 layers, 6 multi-heads, and a hidden size of 300, and were trained until the minimum validation loss was reached. For inference, we used beam search with beam size 32 and 4-gram repeats blocking.

To evaluate the performance of the dialogue models, we adopted the following metrics: **1) perplexity**; **2) distinct-1 and -2 metrics** (Li et al., 2016), which measure the diversity of the generated responses; **3) sentence embedding similarity**—we used SBERT (Reimers and Gurevych, 2019) to obtain an embedding for the generated response as well as the ground-truth and then calculated the cosine similarity between the two embeddings. The performance of the dialogue models was tested in held-out and zero-shot settings. The evaluation results are shown in Table 7.

In the held-out setting, where the model is evaluated on data from the same domain as the training data, all three models achieved good performance, and the perplexity values are much lower compared with the zero-shot setting, where the model is evaluated on data from a different domain. We also observe that the model fine-tuned on OS and EDOS dialogues achieves much higher Distinct-1 and -2 scores, even in the zero-shot setting when evaluated on EmpatheticDialogues. This indicates that by training on our curated OpenSubtitles dialogues, the model gains more diversity in the generated responses. It might be due to the larger size of the datasets containing many diverse responses. Out of the two, EDOS performs the best in terms of diversity, which reflects the quality of dialogues filtered from OpenSubtitles.

6 Discussion and conclusion

In this work, we curated a large-scale dialogue dataset, EDOS, comprising of 1M emotional dialogues from movie subtitles. This dataset is significantly larger in size and contains more fine-grained emotion categories and empathetic response intents than the existing emotional dialogue datasets. To facilitate annotation, we utilized data augmentation techniques to extend a small set of manually annotated data and trained a dialogue emotion classifier having comparable accuracy to the state-of-the-art. The data augmentation and automatic annotation procedure we employed significantly reduced the manual annotation cost and time.

Obtaining a large dataset is important only if the quality can be assured. The qualitative comparison conducted between EDOS and the state-of-the-art EmpatheticDialogues dataset by means of visual validation was one way to confirm that. The results of the comparison confirmed that most of the conversational dynamics present in EmpatheticDia-

	OS				EDOS				EmpatheticDialogues			
Model	**PPL**	**D1**	**D2**	**SES**	**PPL**	**D1**	**D2**	**SES**	**PPL**	**D1**	**D2**	**SES**
Pre-trained (OS)	**24.8**	**.046**	**.159**	.172	37.8	.046	.154	.126	564.6	**.044**	**.167**	.178
Fine-tuned (EDOS)	26.9	.044	.139	.162	**32.3**	**.056**	**.165**	**.137**	452.6	.031	.107	.176
Fine-tuned (ED)	88.9	.030	.109	**.174**	140.8	.028	.096	.130	**19.3**	.026	.091	**.316**

Table 7: Dialogue model evaluation results. Here PPL denotes perplexity, D1 and D2 denote Distinct-1 and -2, and SES denotes the sentence embedding similarity. : held-out, : zero-shot.

logues were observed in EDOS. We also proposed some experimental baselines by training a transformer model for empathetic response generation on OS, EDOS, and EmpatheticDialogues datasets and tested them in held-out and zero-shot settings. The results showed that the model fine-tuned on EDOS scored the best in terms of diversity metrics. This dataset can be readily utilized to develop empathetic conversational agents and for fine-grained emotion analysis in dialogues. The pipeline we present can be used when creating similar large-scale datasets in similar or even different domains.

As future work, we plan to utilize this dataset to further conduct experiments on empathetic response generation. Since it is annotated with emotions and intents, we will use it for experiments involving controllable and interpretable response generation. Particularly, the plus categories present in the dataset can be utilized to condition the chatbot's response generation process, making it possible to control and interpret the generated responses. The dataset can also be used to train state-of-the-art dialogue emotion classifiers.

7 Ethical considerations

EDOS contains dialogues derived from the OpenSubtitles corpus (Lison et al., 2019), which is publicly available.[2] It is part of the OPUS (Open Parallel corpUS), which is based on open source products and is delivered as an open content package. The workers annotating the dataset were compensated with $0.4 per HIT, which takes 4.12 minutes on average to complete (excluding the time taken by workers who took an unusually long time to complete the task) and a bonus of $0.1 if they completed at least 3 out of 5 quiz questions correctly. Fair compensation was determined based on the US minimum wage of $7.12 per hour. Since the dataset is in English, the annotators recruited from AMT were restricted the majority native English-speaking countries: US; UK; Canada; Australia; and New Zealand. The fact that the dataset is English-only potentially perpetuates an English bias in NLP systems.

Using this dataset to directly train end-to-end chatbot models can involve certain risks. Though we have taken steps to remove profanity from the responses in the dataset, due to the lack of controllability and interpretability in end-to-end neural response generation models, there exists the risk of generating inappropriate or biased responses for certain emotional prompts. A recent example is Microsoft's Taybot that started producing unintended and offensive tweets denying the Holocaust as a result of learning from offensive information from Twitter (Lee, 2016). To mitigate this, researchers have recently focussed on inducing controllability in these end-to-end response generation models by means of jointly modeling dialogue intent selection and response generation (Wu et al., 2018; Sankar and Ravi, 2019; Hedayatnia et al., 2020; Santhanam et al., 2020; Ke et al., 2018; Lee et al., 2020). We encourage the readers to look into these approaches when developing conversational agents using this dataset.

Though human-like chatbots with emotion recognition and empathetic responding abilities can be beneficial in a number of situations such as in the medical domain, crisis management, customer service, and elderly care, it should not be underestimated that they involve some potential harms. For example, a chatbot can be used to impersonate a real human being and used for cybercrimes such as scamming and phishing. It is also important to note that one could get emotionally attached to a bot, or even become codependent, distracting him or herself from relationships with humans and causing distress if the chatbot becomes dysfunctional. Users may tend to reveal their private and confidential information such as certain health conditions and private attributes during such interaction, which could be misused when in the hands of the wrong people. Developers should take these risks into account when deploying such chatbots in the real world to ensure safe and ethical use.

[2] https://opus.nlpl.eu/OpenSubtitles-v2018.php

References

Philip Ball. 2011. How movies mirror our mimicry.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

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

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

Michele Filannino and Marilena Di Bari. 2015. Gold standard vs. silver standard: the case of dependency parsing for italian. *CLiC it*, page 141.

Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. 2009. A dimensional approach to emotion recognition of speech from movies. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 65–68. IEEE.

Behnam Hedayatnia, Karthik Gopalakrishnan, Seokhwan Kim, Yang Liu, Mihail Eric, and Dilek Hakkani-Tur. 2020. Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Jonathan Herzig, Guy Feigenblat, Michal Shmueli-Scheuer, David Konopnicki, Anat Rafaeli, Daniel Altman, and David Spivak. 2016. Classifying emotions in customer support dialogues in social media. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 64–73.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Amir Kazem Kayhani, Farid Meziane, and Raja Chiky. 2020. Movies emotional analysis using textual contents. In *International Conference on Applications of Natural Language to Information Systems*, pages 205–212. Springer.

Pei Ke, Jian Guan, Minlie Huang, and Xiaoyan Zhu. 2018. Generating informative responses with controlled sentence function. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1499–1508, Melbourne, Australia. Association for Computational Linguistics.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Dave Lee. 2016. Tay: Microsoft issues apology over racist chatbot fiasco.

Hung-yi Lee, Cheng-Hao Ho, Chien-Fu Lin, Chiung-Chih Chang, Chih-Wei Lee, Yau-Shian Wang, Tsung-Yuan Hsu, and Kuan-Yu Chen. 2020. Investigation of sentiment controllable chatbot. *arXiv preprint arXiv:2007.07196*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016*, pages 110–119.

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.

Pierre Lison and Raveesh Meena. 2016. Automatic turn segmentation for movie & tv subtitles. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 245–252. IEEE.

Pierre Lison, Jörg Tiedemann, Milen Kouylekov, et al. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Erinc Merdivan, Deepika Singh, Sten Hanke, Johannes Kropf, Andreas Holzinger, and Matthieu Geist. 2020. Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984:197–219.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 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. Association for Computational Linguistics.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Chinnadhurai Sankar and Sujith Ravi. 2019. Deep reinforcement learning for modeling chit-chat dialog with discrete attributes. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 1–10, Stockholm, Sweden. Association for Computational Linguistics.

Sashank Santhanam, Zhuo Cheng, Brodie Mather, Bonnie Dorr, Archna Bhatia, Bryanna Hebenstreit, Alan Zemel, Adam Dalton, Tomek Strzalkowski, and Samira Shaikh. 2020. Learning to plan and realize separately for open-ended dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2736–2750, Online. Association for Computational Linguistics.

Amy E Skerry and Rebecca Saxe. 2015. Neural representations of emotion are organized around abstract event features. *Current biology*, 25(15):1945–1954.

Isaac Triguero, Salvador García, and Francisco Herrera. 2015. Self-labeled techniques for semi-supervised learning: taxonomy, software and empirical study. *Knowledge and Information systems*, 42(2):245–284.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS 2017*, pages 5998–6008.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4886–4899.

Wei Wu, Can Xu, Yu Wu, and Zhoujun Li. 2018. Towards interpretable chit-chat: Open domain dialogue generation with dialogue acts.

A Computing the readability of OS dialogues

We followed the following steps in calculating the readability of the dialogues. The dialogues that scored high in readability were preferred for the crowd-annotation task since they avoid the overhead of having to read long and complex dialogues that may exhaust the crowd-worker.

1. Build a frequency vocabulary by calculating the token count for all the dialogues in the cleaned OS dataset.

2. For each dialog, aggregate the frequencies of all tokens and take the average using the following formula, in which f_{sum} is the sum of frequencies of all tokens, n_{tokens} is the total number of tokens in the dialog, and α is a constant (set to 87 in our case). The idea behind this is that difficult to read dialogues contain less frequent words and should result in less readability.

$$f = f_{sum}/(\alpha + n_{tokens})$$

3. For each dialog, also calculate the percentage of distinct words, say d.

4. Finally, compute the readability score for each dialogue by taking the weighted sum of f and d. Experimental results showed that the combination of $f + 0.04d$ was giving the best results. We take the combination of both f and d because, if only f is considered, then dialogues that contain a lot of repetitive tokens can score high in readability, which is undesirable.

B AMT task interfaces

The user interface used to collect labels from the AMT workers is denoted in Figure 4.

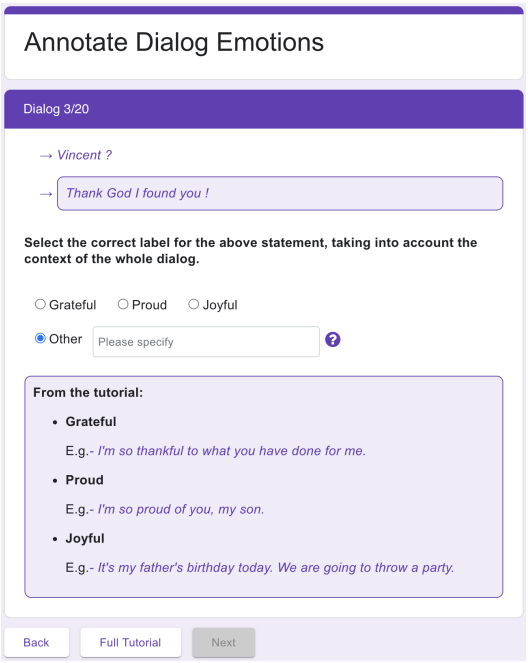


Figure 4: The user interface of the AMT crowd-annotation task.

C Choice of hyper-parameters and additional training details regarding the dialogue emotion classifier used for annotation

The choice of 5 seconds to separate dialogues is based on a histogram of time intervals between adjacent subtitle blocks in the OpenSubtitles corpus, which is denoted in Figure 5. As it can be observed in the histogram, most of the time gaps fall below 3 seconds. A clear drop in count was observed between 3-5 seconds. Therefore, we chose 5 seconds as the time interval to separate dialogues.

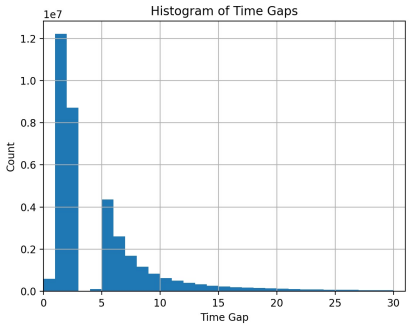


Figure 5: Histogram of time intervals between adjacent subtitle blocks in the OpenSubtitles corpus.

The choice a threshold of 0.92 to select dialogues similar to those that were already annotated was based on manually inspecting a random subset of the results obtained after using a range of similarity thresholds. Table 8 shows some example dialogues discovered at this threshold.

Using decreasing weights for context utterances is based on the intuition that in human dialogues, more attention is paid to the most recent utterances in dialogue history. This idea is backed up by time-decay functions used in neural dialogue understanding approaches (See et al., 2019). We conducted an ablation study with without using decreasing weights in the model. Performance of the unweighted models was lower than the performance of weighted models yielding final F1 scores of 63.44 and 64.86 for unweighted weighted models, respectively.

We used the same hyper-parameter setting used in RoBERTa (Liu et al., 2019) when training the dialogue emotion classifier used for annotation. We used the Adam optimizer with β_1 of 0.9, β_2 of 0.98, an ϵ value of 1×10^{-6}, and a learning rate of 2×10^{-5}. A dropout of 0.1 was used on all layers and attention weights, and a GELU activation function (Hendrycks and Gimpel, 2016). We limited the maximum number of input tokens to 100, and used a batch size of 256. All the experiments were conducted on a machine with 2x12cores@2.5GHz, 256 GB RAM, 2x240 GB SSD, and 2xGPU (NVIDIA Titan X Maxwell). 546.84 sec.s in total were taken to train the final emotion classifier. The optimal model was selected based on the average cross entropy loss calculated between the ground-truth and predicted labels of the validation set.

D EDOS statistics

Table 9 shows more descriptive statistics of the EDOS dataset: the number of dialogues; and the number of dialogues turns per emotion and intent category. A dialogue is counted under an emotion or an intent if the beginning dialogue prompt is annotated with that emotion or intent.

E Additional training details about the experiemental baselines

Here we summarize some of the parameters of the model implementation. We used the RoBERTa tokenizer to tokenize the input utterances, and the vocabulary size is 50,265. We allow a maximum number of 100 tokens as the input to the model. We used 4 sub-layers in the encoder and decoder, with 6 heads in the multi-head attention. The dimension of the hidden units is 300, and the dimension of the

Manually annotated dialogues	Dialogues discovered using similarity matching (with similarity $\geq$ 0.92)
- ***That 's beautiful !.*** **(Acknowledging)**	- *Now , let 's take a look at this beautiful piece of work* - *Oh , my God . It 's beautiful .* - ***Oh . That 's beautiful .***
- *I thought the coils were closer to me .* - *Oh , well ... It was a good one nonetheless .* - ***I 'm so happy !*** **(Joyful)**	- *Actually , I just wanted to say I love you . And I 'm sorry if I 'm a bit edgy about my book , but all that counts for me is you . You becoming my wife .* - *That 's what really matters .* - ***I 'm very happy .***
- *Hey ! Don 't eat at my house anymore .* - ***You 're disgusting .*** **(Disgusted)**	- *I thought I told you to stay the fuck away from me if you were back on that shit .* - ***You 're disgusting .***
- *Was the team mad , then ?* - *I wasn 't happy !* - ***That 's pretty bad .*** **(Acknowledging)**	- *It 's starting to hurt so bad .* - *Really ? That bad ?* - ***Really bad .***

Table 8: Examples of similar dialogues discovered above a cosine similarity threshold of 0.92. The last turn in each dialogue discovered through similarity matching was labeled with the emotion or intent of that of the last turn of the manually labeled dialogue.

Emotion or Intent	No. of dialogues	No. of turns
Prepared	21,178	48,883
Anticipating	27,256	100,433
Hopeful	21,328	54,012
Proud	13,910	33,365
Excited	22,118	53,756
Joyful	6,586	24,282
Content	20,688	64,569
Caring	13,599	42,806
Grateful	15,416	42,222
Trusting	41,650	134,197
Confident	26,199	84,918
Faithful	8,095	25,029
Impressed	12,867	25,045
Surprised	16,658	46,022
Terrified	9,449	28,730
Afraid	15,964	49,285
Apprehensive	8,634	46,727
Anxious	2,376	8,578
Embarrassed	11,541	32,338
Ashamed	3,401	14,797
Devastated	6,245	17,539
Sad	23,023	66,262
Disappointed	5,234	18,298
Lonely	3,662	16,396
Sentimental	7,104	20,715
Nostalgic	7,880	20,461
Guilty	9,632	30,043
Disgusted	5,546	15,070
Furious	54,647	169,917
Angry	13,228	34,924
Annoyed	6,637	30,072
Jealous	5,766	20,902
Agreeing	20,173	96,562
Acknowledging	39,781	138,165
Encouraging	3,024	10,329
Consoling	3,785	17,256
Sympathizing	15,557	38,774
Suggesting	42,470	101,591
Questioning	357,255	841,556
Wishing	42,789	108,668
Neutral	7,649	55,932
Total	1,000,000	2,829,426

Table 9: Descriptive statistics of the EDOS dataset pertaining to each emotion and intent category.

pointwise feed-forward layers is 1200. We use a dropout rate of 0.1, and the GELU (Hendrycks and Gimpel, 2016) activation function for the hidden layers. The loss function was optimized with the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 5 × 105 For inference, we use beam search with a beam size of 32. To prevent the models from generating repetitive tokens or n-grams, we modified the beam search algorithm so that at each time step, if any of the branches contains repetitive 4-grams, we set the log probability of this branch to infinitely negative, to stop it from being further expanded. All the models were trained with a batch size of 512, on machines with 4 Nvidia Titan X Pascal GPUs, 2 Intel Xeon E5-2680 v3 CPUs, and 256GB RAM. Table 10 lists the training details as well as the validation performance for all the models.

Model	# Parameters	# Training Epochs	Training Time	Validation PPL
Pre-trained (OS)	121M	50 epochs	171.00 hr	24.51
Fine-tuned (EDOS)	121M	5 epochs	4.23 hr	31.78
Fine-tuned (ED)	121M	9 epochs	19.50 min	21.04

Table 10: Training details and validation performance of each model configuration.