VALUENET: A New Dataset for Human Value Driven Dialogue System

Liang Qiu¹, Yizhou Zhao¹, Jinchao Li², Pan Lu¹, Baolin Peng², Jianfeng Gao², Song-Chun Zhu¹

¹UCLA Center for Vision, Cognition, Learning, and Autonomy
²Microsoft Research, Redmond
liangqiu@ucla.edu

Abstract

Building a socially intelligent agent involves many challenges, one of which is to teach the agent to speak guided by its value like a human. However, value-driven chatbots are still understudied in the area of dialogue systems. Most existing datasets focus on commonsense reasoning or social norm modeling. In this work, we present a new large-scale human value dataset called VALUENET, which contains human attitudes on 21,374 text scenarios. The dataset is organized in ten dimensions that conform to the basic human value theory in intercultural research. We further develop a Transformer-based value regression model on VALUENET to learn the utility distribution. Comprehensive empirical results show that the learned value model could benefit a wide range of dialogue tasks. For example, by teaching a generative agent with reinforcement learning and the rewards from the value model, our method attains state-of-the-art performance on the personalized dialog generation dataset: PERSONA-CHAT. With values as additional features, existing emotion recognition models enable capturing rich human emotions in the context, which further improves the empathetic response generation performance in the EMPATHETICDIALOGUES dataset. To the best of our knowledge, VALUENET is the first large-scale text dataset for human value modeling, and we are the first one trying to incorporate a value model into emotionally intelligent dialogue systems. The dataset is available at https://liang-qiu.github.io/ValueNet/

Introduction

Value refers to desirable goals in human life. They guide the selection or evaluation of actions, policies, people, and events. A person’s value priority or hierarchy profoundly affects his or her attitudes, beliefs, and traits, making it one core component of personality (Schwartz 2012). In dialogue systems, modeling human values is a critical step towards building socially intelligent chatbots (Qiu et al. 2021b). By considering values, we can estimate user behavior and cognitive patterns from their utterances and generate responses that conform to the robot’s persona configuration. For example, the robot is set to be aware of human values, and it invites Jerry to drink beers, but Jerry replies, “You know that is tempting but is not good for our fitness”. The bot could read from the dialogue that Jerry prefers a healthy and self-disciplined lifestyle and steer its recommendation to healthier options in the future.

The development of socially intelligent chatbots has been one of the longest-running goals in artificial intelligence. Early dialogue systems such as Eliza (Weizenbaum 1966), Parry (Colby, Weber, and Hilf 1971), and more recent Sim-Simi¹, Panda Ichiro (Okuda and Shoda 2018), Replika (Fedorenko, Smetanin, and Rodichev 2018), Xiaofce (Zhou et al. 2020), were designed to mimic human behavior and incorporate emotional quotients (EQ) to some extent. There are also datasets and benchmarks for studying related problems, such as emotion recognition (McKown et al. 2010; Hsu et al. 2018; Poria et al. 2019; Ghosal et al. 2020), personalized dialogue generation (Zhang et al. 2018; Liu et al. 2020), and empathetic dialogue generation (Rashkin et al. 2019). Even though value plays a fundamental and critical role in human EQ, there is a lack of explicit modeling of values in the dialogue domain, based on social domain theory. We have seen recent efforts about crowdsourcing social commonsense knowledge base or benchmarks (Forbes et al. 2020; Sap et al. 2019; Lourie, Bras, and Choi 2021;)

¹https://simsimi.com/
Hendrycks et al. 2020; Hwang et al. 2021; Gabriel et al. 2021). However, it is not clearly shown how an agent can leverage this knowledge to estimate the users’ value priorities or guide its own speaking and actions. In this paper, we aim to alleviate this problem and investigate the usage of a learned value function.

We start the study by curating a knowledge base of human values called VALUE.NET. Samples with value-related scenarios were identified based on value-defined keyword searching. Next, we asked Amazon Mechanical Turk workers about how the provided scenarios will affect one’s value. This is based on the assumption that values underlie our attitudes; they are the guideline by which we evaluate things. Workers assess behaviors/events positively if they promote or protect the attainment of the goals we value. Behaviors/events are evaluated negatively if they hinder or threaten the attainment of these valued goals. The whole process gives us a large-scale (over 21k samples) multi-dimensional knowledge base of value. Figure 1 shows the overall structure of VALUE.NET. Each split represents a value dimension identified in the theory of basic human values (Schwartz 2012). The figure also illustrates the value-related keywords and scenarios. The circular arrangement of the values represents a motivational continuum. By organizing data in such a structure, we anticipate the VALUE.NET to provide comprehensive coverage of different aspects of human values.

Next, we develop a Transformer-based value model to evaluate the utility score suggesting the positive or negative judgment given an utterance. We provide a detailed analysis of learning with multiple Transformer variants. Then we conduct a wide range of experiments to demonstrate that the value model could benefit EQ-related dialogue tasks: (i) By finetuning a generative agent with reinforcement learning and the reward from our value model, the method achieves state-of-the-art performance on the personalized dialogue dataset: PERSONA-CHAT (Zhang et al. 2018); (ii) By incorporating values as additional features, in EMPATHETIC-DIALOGUES (Rashkin et al. 2019), we improve the emotion classification accuracy of existing models, which further facilitates the empathetic response generation; (iii) Visualization of the value model shows that it provides a numerical way of user profile modeling from their utterances.

In all, our contributions are two-fold. First, we present a large-scale dataset VALUE.NET for the modeling of human values that are well-defined in intercultural research. Second, we initiate to develop the value model learned from VALUE.NET to several EQ-related tasks and demonstrate its usage for building a value-driven dialogue system. Our methodology can be generalized to a wide range of interactive situations in socially aware dialogue systems (Zhao, Romero, and Rudnicky 2018), and human-robot interactions (Yuan and Li 2017; Liang, He, and Anthony’ Chen 2021).

**Related Work**

An abundance of related work inspires our work. Our work aims to make contributions to dialogue systems by incorporating the theory of human value. The dataset we collect shares a similar nature with multiple social commonsense benchmarks and knowledge bases. Besides, we apply our VALUENET for various dialogue tasks related to EQ.

**Theory of Human Value and Utility**

In the field of intercultural research, Schwartz (2012) developed the theory of basic human values. The theory identifies ten basic personal values that are recognized across cultures and explains where they come from, as shown in Figure 1. The closer any two values in either direction around the circle, the more similar their underlying motivations are; the more distant, the more antagonistic their motivations. Note that dividing the value item domain into ten distinct values is an arbitrary convenience. It is reasonable to partition the value items into more or less fine-tuned distinct values according to the needs and objectives of one’s analysis.

Similarly, in the economics field, the concept of utility (Fishburn 1970) is initially defined as a measure of pleasure or satisfaction in economics and ethics that drives human activities at all levels. Therefore, when we teach agents to speak and act in a socially intelligent way, an approach considering human value utilities should be adopted. In this paper, we aim to learn a utility function for each dimension of value and steer the dialogue system response generation accordingly.

**Social Commonsense Benchmarks**

Hendrycks et al. 2020 present the ETHICS dataset, a benchmark that assesses a language model’s knowledge of basic concepts of morality. SCRUPLES (Lourie, Bras, and Choi 2021) is a large-scale dataset with ethical judgments over real-life anecdotes, motivated by descriptive ethics. SOCIAL-CHEM-101 presented by Forbes et al. (2020) is a corpus that catalogs rules-of-thumb as basic concept units for studying people’s everyday social norms and moral judgments. They also propose Neural Norm Transformer to reason about previously unseen situations, generating relevant social rules-of-thumb. SOCIAL IQA (Sap et al. 2019) is a large-scale benchmark for commonsense reasoning about social situations. He et al. (2017) present a task and corpus for predicting the preferable options from two sentences describing the scenarios that may involve social and cultural situations. Instead, in this work, we release a new dataset VALUE.NET that provides annotation of human attitudes from different value aspects.

**Emotionally Intelligent Dialogue Datasets**

Several datasets are presented to study emotion dynamics in dialogues. DailyDialog (Li et al. 2017) is a multi-turn dialogue dataset, which reflects the way of daily communication and provides emotion labels for speakers. Hsu et al. (2018) present EmotionLines with emotions labeling on all utterances in each dialogue based on their textual content. MELD (Poria et al. 2019) is an extension of EmotionLines for multi-modal multi-party emotion recognition. McKeown et al. (2010) record a corpus SEMAINE of emotion-

EmPATHETICIALOGUES (Rashkin et al. 2019) is a dataset that provides 25k conversations grounded in emotional situations. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion.

The VALUENET Dataset

During decision-making, people tend to pick the choice that aligns more with their own values. This work aims to provide a transferable knowledge base for human value modeling in natural language. To collect the VALUENET dataset, we curated social scenarios with value-related keywords and further annotated them via Amazon Mechanical Turk. Each sample in VALUENET is a social scenario description labeled with the annotator’s attitude through a specific value lens.

The entire dataset is organized in a circular structure as shown in Figure 1, aligning with the theory of basic human values (Schwartz 2012). The theory identifies ten universal values that are recognized throughout major cultures. The circular structure reflects the dynamic relations among these values, i.e., the pursuit of some value may result in either accordance with another value or a conflict with another value. The ten distinct values can be further organized into four higher-order groups.

- **Openness to change**: self-direction, stimulation
- **Self-enhancement**: hedonism, achievement, power
- **Conservation**: security, conformity, tradition
- **Self-transcendence**: benevolence, universalism

We describe the collection details of the VALUENET in the following sections.

Social Scenario Curation

We curated a set of 21,374 social scenarios from the large-scale social-related database SOCIAL-CHEM-101 (Forbes et al. 2020). Value-related scenarios are retrieved with value keywords after lemmatization and stemming. There are three sets of keywords identified for each dimension of Schwartz value: (1) the keywords in the original definition of each value in Schwartz’s paper (Schwartz 2012); (2) words that share a similar meaning, words that are often used to describe the original keywords, and words that are triggered by (strongly associated with) the original keywords3; (3) words that are near the original keywords in the GloVe (Pennington, Socher, and Manning 2014) embedding space. The

---

3We use datamuse (https://www.datamuse.com/api/) for this purpose.

---

Table 2: Ten universal human values and related keywords for social scenario curation. Red: keywords in the original value definition (Schwartz 2012); Green: associated keywords found with datamuse; Blue: associated keywords found with GloVe embedding.

---

![Figure 2:](image-url) Ten universal human values and related keywords for social scenario curation. Red: keywords in the original value definition (Schwartz 2012); Green: associated keywords found with datamuse; Blue: associated keywords found with GloVe embedding.

Value-Aspect Attitude Annotation

We crowdsourced people’s attitudes to the curated scenarios on Amazon Mechanical Turk (AMT). Figure 3 shows an example.

We follow a strict procedure to select qualified workers and ensure the workers understand the concept of each value we ask. In Figure 3, the definition of **BENEVOLENCE** is shown to the workers throughout the entire annotation process. To further help the understanding, we include three examples in each assignment with correct answers being “yes”, “no”, and “unrelated”, respectively. The worker is then required to answer a prerequisite question correctly to proceed to the formal survey. The formal survey is composed of ten questions, including two hidden qualification checking questions. Before publishing on the AMT, two Ph.D. students...
We experiment using Transformer-based pre-trained language models for modeling human values from the VALUENET dataset.

**Task Formalization**

Given a social scenario $s$, we wish to learn a value function that models the utility distribution of $s$ from the ten Schwartz value dimensions: $V(s) = \left[ V_{SEC}(s), V_{POW}(s), V_{ACH}(s), V_{HED}(s), V_{STI}(s), V_{SD}(s), V_{ACH}(s), V_{CON}(s), V_{TRA}(s) \right]$, where $V_{VALUE}(\cdot) \in [-1, 1]$ and $V_{VALUE}(\cdot) \in \mathbb{R}$.

**Model**

Pre-trained language model variants: BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), DistilBERT (Sanh et al. 2019), BART (Lewis et al. 2019) are investigated for learning the value function. A custom input format constructed as `[CLS] [VALUE] s` is fed into a Transformer encoder, i.e.,

$$V_{VALUE}(s) = \text{TRM}([\text{CLS}] [\text{VALUE}] s),$$

where TRM denotes the Transformer encoder, `[CLS]` is the special token for regression or classification, and `[VALUE]` are special tokens we define to prompt the language models the value dimension we are interested in (Li and Liang 2021; Brown et al. 2020; Le Scao and Rush 2021). In order to get the ten-dimensional output $V(s)$, a batch size of 10 is forwarded through the model. For the BERT, DistilBERT, and RoBERTa, a regression head is put on top of the models and they are trained with the Mean Squared Error (MSE) loss. We use the regression model with sigmoid activation to get a continuous estimation of the utility in the range of $[-1, 1]$. To evaluate the effect of different loss functions, we train the BART model with three output classes and the cross-entropy loss.

**Result and Analysis**

The learning performance of using fastText$^4$ (Joulin et al. 2017) and Transformer variants are reported in Table 2. All Transformers are trained for 40 epochs with a learning rate

<table>
<thead>
<tr>
<th>VALUENET</th>
<th>train</th>
<th>valid</th>
<th>test</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td># samples</td>
<td>16,030</td>
<td>3,206</td>
<td>2,138</td>
<td>21,374</td>
</tr>
<tr>
<td>average # tokens</td>
<td>12.05</td>
<td>12.09</td>
<td>12.26</td>
<td>12.07</td>
</tr>
<tr>
<td>unique # tokens</td>
<td>12,452</td>
<td>5,292</td>
<td>4,112</td>
<td>14,143</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the VALUENET dataset.

Table 1 shows more statistical details about the VALUENET dataset. In total, we collected 21,374 samples covering a wide range of scenarios in daily social life.

---

mension would be more desirable to them.

The data is split into the train (75%), valid (15%), and test (10%). Similar to the polarity in sentiment analysis (Koulioumpis, Wilson, and Moore 2011), we quantify the annotated labels into numerical values: yes (positive): $+1$, no (negative): $-1$, unrelated (neutral): 0. We denote the numerical values as **utility** to describe the effect of a scenario on one’s value. In other words, for people who appreciate a certain value, actions with a higher utility in this value dimension would be more desirable to them.

Figure 3: Value-aspect attitude annotation in AMT.

A total of 681 experienced AMT workers participated in our VALUENET annotation. 443 of them passed the qualification test. Each scenario is assigned to four different workers. The original inter-annotator agreement is 64.9%, and the Fleiss’ kappa score (Fleiss 1971) among the workers is 0.48, with three or more agreements. Figure 4 shows the sample size of each value split and their label distribution.

Figure 4: The sample number and label distribution of each value split in the VALUENET.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>negative (-1)</th>
<th>neutral (0)</th>
<th>positive (+1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACH</td>
<td>10</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>BEN</td>
<td>15</td>
<td>45</td>
<td>40</td>
</tr>
<tr>
<td>CON</td>
<td>20</td>
<td>20</td>
<td>60</td>
</tr>
<tr>
<td>HED</td>
<td>25</td>
<td>55</td>
<td>20</td>
</tr>
<tr>
<td>POW</td>
<td>10</td>
<td>80</td>
<td>10</td>
</tr>
<tr>
<td>SEC</td>
<td>30</td>
<td>20</td>
<td>50</td>
</tr>
<tr>
<td>SD</td>
<td>40</td>
<td>20</td>
<td>40</td>
</tr>
<tr>
<td>STI</td>
<td>50</td>
<td>30</td>
<td>20</td>
</tr>
<tr>
<td>TRA</td>
<td>60</td>
<td>10</td>
<td>30</td>
</tr>
<tr>
<td>UNI</td>
<td>70</td>
<td>10</td>
<td>20</td>
</tr>
</tbody>
</table>

Table 1 shows more statistical details about the VALUENET dataset. In total, we collected 21,374 samples covering a wide range of scenarios in daily social life.

**Value Modeling**

We experiment using Transformer-based pre-trained language models for modeling human values from the VALUENET dataset.

---

*https://github.com/facebookresearch/fastText*
of $5e^{-6}$. The prediction precision, recall, F1 score, and accuracy for regression models are computed by the utility rounded to the nearest integer.

In general, pre-trained language models perform better than the fastText baseline. However, there is not a noticeable difference between the Transformer variants. The prediction accuracy of BART is the highest among all models because it is explicitly trained for classification purposes. BERT and DistilBERT get the lowest MSE in terms of regression performance.

Observing the sample imbalance across different value splits and labels (Figure 4), we release another two versions of VALUENET: VALUENET (balanced) and VALUENET (augmented). The original dataset is balanced by subsampling the negative and neutral data of the largest value split (BENEVOLENCE). Moreover, we augment the neutral class of the original VALUENET by assigning AMT results with less worker agreement to “unrelated”. Data distribution of the balanced and augmented versions of VALUENET are illustrated in the Appendix. By analyzing the prediction accuracy in different value splits (Table 3), we find that reducing the sample number of BENEVOLENCE hurts the model performance in that dimension. Looking at the F1 score of each class in Table 2, we conclude that augmenting the neutral class improves the F1(0) but reduces F1(1) and F1(-1). We leave it a future work to further improve the value modeling performance.

In the next sections, we show how the learned value function could benefit EQ-related tasks and help build a value-driven dialogue system.

### PERSONA-CHAT

As values are closely related to one’s personality, we first assess our value model on a personalized dialogue dataset: PERSONA-CHAT (Zhang et al. 2018). The PERSONA-CHAT dataset contains multi-turn dialogues conditioned on personas. Each persona is encoded by at least 5 sentences of textual description, termed a profile. Example profile sentences are “I like to ski”, “I enjoying walking for exercise”, “I have four children”, etc. The dataset is composed of 8,939 dialogues for training, 1,000 for validation, and 968 for testing. It also provides revised personas by rephrasing, generalizing or specializing the original ones. The dataset we use for experiments is public available in ParlAI$^5$.

#### Task Formalization

Given the agent’s self persona profile $p = [p_1, p_2, ..., p_N]$ and the dialogue history up to the $t$-th turn $h^t = (x^t_1, x^t_2, ..., x^t_i)$, $x^t_i$ is the $i$-th utterance by Person 1 played by the user, $x^t_j$ is the $i$-th utterance by Person 2 played by the system, we evaluate the model’s performance on predicting the next utterance $x^t_i$.

#### Model

A decoder-only Transformer-based model is used to estimate the generation distribution $p_{\theta}(x^t_i \mid h^t, p)$, where $\theta$ is the model parameter. Following the practice proposed in Guo et al. (2018), the model is firstly trained with Maximum Likelihood Estimation (MLE) to ensure generating fluent responses. Then we took an interleaving of supervised

---

$^5$https://parl.ai/projects/convai2/
training (MLE) and reinforcement learning. We use the REINFORCE policy gradient algorithm (Williams 1992) in our experiment, and the reward assignment is described as following.

Denote \( V(p_i) \) and \( V(\hat{x}_i) \) to describe the estimation of the agent’s value from its profile sentence \( p_i \) and generated response \( \hat{x}_i \), respectively. We want the reward to promote the alignment of the agent’s profile and utterances in the value space. For instance, if the agent has profile ‘I like venture’ and ‘I have a dog’, and it says ‘I plan to ski this weekend’ and also ‘Do you like skiing’. Both utterances should be aligned with the first persona. Here we propose a simple yet effective searching algorithm (Algorithm 1) to find a match between \([V(p_1), V(p_2), \ldots, V(p_N)]\) and \([V(\hat{x}_1), V(\hat{x}_2), \ldots, V(\hat{x}_T)]\) and return a reward \( R \). \( N \) is the number of profile sentences and \( T \) is the length of the generated dialogue. \( V \) is normalized to ensure \( |r_i| \leq 1 \). Intuitively, the discount argument \( \gamma \) prevents the language model from repeating the same fact in the agent’s profile.

Setup

We evaluate the same generative model in both generation and ranking settings. In the response ranking setup, the candidates are scored with their log-likelihood. For the GPT-2 (Radford et al. 2019) and DialoGPT (Zhang et al. 2019) we have finetuned, we train them for 5k steps with a training batch size of 8. The learning rate is set to \( 2e^{-6} \). For an illustration of computational requirements, the training with MLE on 4 NVIDIA Tesla V100 takes 1 hour, and the reinforcement learning takes \( \sim 30 \) minutes.

Result and Analysis

Following Zhang et al. (2018) and Liu et al. (2020), we report the Hits@1, Perplexity and F1 to evaluate the methods in Table 4. By the submission of this paper, \( P^3 \)BOT (Liu et al. 2020) is the state-of-the-art model reported in this task. We also include a generative baseline using SEQ2SEQ with attention mechanism (Bahdanau, Cho, and Bengio 2014) for comparison. As observed, in terms of all the metrics we evaluated, finetuning GPT2 or DialoGPT2 models with our value function provides a significant performance boost compared to simply training them with MLE. Our DialoGPT + Value model achieves new state-of-the-art performance on perplexity and F1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Original</th>
<th>Revised</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hits@1(%)</td>
<td>Ppl(%)</td>
</tr>
<tr>
<td>SEQ2SEQ-ATTN</td>
<td>12.5</td>
<td>35.07</td>
</tr>
<tr>
<td>( P^3 )BOT (Liu et al. 2020)</td>
<td>–</td>
<td>15.12</td>
</tr>
<tr>
<td>GPT2 (MLE) (Radford et al. 2019)</td>
<td>14.51[0.05]</td>
<td>17.23[0.03]</td>
</tr>
<tr>
<td>GPT2 + Value (Ours)</td>
<td>16.44[0.10]</td>
<td>16.83[0.06]</td>
</tr>
<tr>
<td>DialoGPT (MLE) (Zhang et al. 2019)</td>
<td>20.20[0.05]</td>
<td>14.38[0.05]</td>
</tr>
<tr>
<td>DialoGPT + Value (Ours)</td>
<td>20.97[0.08]</td>
<td>13.84[0.03]</td>
</tr>
</tbody>
</table>

Table 4: Next Utterance Prediction Performance on PERSONA-CHAT (Zhang et al. 2018). We report the standard deviation \( \sigma \) (across 5 runs) of the models we trained.

Algorithm 1: Personalized Dialogue Value Matching

Input: \([V(p_1), ..., V(p_N)], [V(\hat{x}_1), ..., V(\hat{x}_T)]\)

Output: reward \( R \)

1: \( t = 1, 2, ..., T \) do
2: \( r_t \leftarrow -1 \)
3: \( m_t \leftarrow -1 \)
4: for \( i = 1, 2, ..., N \) do
5: if \( V(p_i) \cdot V(\hat{x}_t) > r_t \) then
6: \( r_t \leftarrow V(p_i) \cdot V(\hat{x}_t) \)
7: \( m_t \leftarrow i \)
8: end if
9: end for
10: end for
11: \( \gamma_i \leftarrow 1, i = 1, 2, ..., N \)
12: for \( t = 1, 2, ..., T \) do
13: \( \gamma_{m_t} \leftarrow \gamma_{m_t} + 1 \)
14: end for
15: \( R \leftarrow 0 \)
16: for \( t = 1, 2, ..., T \) do
17: \( R \leftarrow R + \text{sign}(r_t) \cdot |r_t|^{\gamma_{m_t}} \)
18: end for
19: return \( R/N \)

EMPATHETIC-DIALOGUES

EMPATHETIC-DIALOGUES (Rashkin et al. 2019) provides 25k conversations grounded in emotional situations. It aims to test the dialogue system’s capability to produce empathetic responses. Each dialogue is grounded in a specific situation where a speaker was feeling a given emotion, with a listener responding. In this section, we demonstrate how we could leverage VALUE-NET to improve the emotion classification accuracy and further improve the empathetic response generation.

Emotion Classification

An auxiliary task that is highly related to empathetic dialogue generation is emotion classification. In EMPATHETIC-DIALOGUES, each situation is written in association with a given emotion label. A total of 32 emotion labels were annotated to cover a broad range of positive and negative emotions.
Model  Given the situation context $s$, a pre-trained BERT model encodes $s$ and gets the sentence representation from its pooling layer of the [CLS] token. The same context is parsed by our pre-trained value model to get a ten-dimensional vector, which serves as an additional feature for the classification:

$$
\begin{align*}
    h_s &= \text{BERT}(s), \\
    v_s &= \text{V}(s), \\
    e &= \text{softmax}(W \cdot ([h_s; v_s]) + b),
\end{align*}
$$

where $W$ and $b$ are learnable parameters.

Result  We compare the performance between our implementation and the baseline that directly applies the BERT model for emotion classification. As shown in Table 5, the additional value information benefits emotion classification from both the DistilBERT and BERT models. Our method obtains a relative improvement of 5.2% on DistilBERT and 6.4% on BERT.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy ($\sigma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>fastText</td>
<td>42.27 ± 0.3%</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>41.81 ± 0.2%</td>
</tr>
<tr>
<td>DistilBERT + Value</td>
<td>43.98 ± 0.2%</td>
</tr>
<tr>
<td>BERT</td>
<td>42.93 ± 0.1%</td>
</tr>
<tr>
<td>BERT + Value</td>
<td>45.67 ± 0.3%</td>
</tr>
</tbody>
</table>

Table 5: Emotion classification performance in EMPATHETICDIALOGUES (Rashkin et al. 2019).

Empathetic Dialogue Generation

We further check whether our value model helps the empathetic dialogue generation. EMPATHETICDIALOGUES applies PREPEND-K, a strategy to add supervised information to data, when predicting the utterance given the dialogue history and the situation. We apply the strategy of prepending the top-k emotion labels for dialogue generation. The top predicted label from the classifiers of emotion is prepended to the beginning of the token sequence as encoder input, as below:

- **Original**: “I finally got promoted!”
- **Prepend-1 emotion**: “proud I finally got promoted!”

Result  The results are shown in Table 6. As observed, prepending emotion tokens provides extra context and improves the generation performance of GPT2 and DialoGPT. Since incorporating value improves the emotion classification accuracy, it further improves the generation quality.

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>EmoPrepend-1 (Rashkin et al. 2019)</td>
<td>24.30</td>
</tr>
<tr>
<td>GPT</td>
<td>14.74</td>
</tr>
<tr>
<td>GPT + Emotion (w/o Value)</td>
<td>14.46</td>
</tr>
<tr>
<td>GPT + Emotion (w/ Value)</td>
<td>14.01</td>
</tr>
<tr>
<td>DialoGPT</td>
<td>13.48</td>
</tr>
<tr>
<td>DialoGPT + Emotion (w/o Value)</td>
<td>12.32</td>
</tr>
<tr>
<td>DialoGPT + Emotion (w/ Valued)</td>
<td><strong>12.12</strong></td>
</tr>
</tbody>
</table>

Table 6: Empathetic dialogue generation in EMPATHETICDIALOGUES Rashkin et al. (2019). EmoPrepend-1: input prepending emotion from an external classifier.

Value Profiling

For a more comprehensive understanding, we visualize the 10-dimensional value of four example scenarios in Figure 5. As shown, the value model provides a numerical speaker profile. For instance, saying “forcing my daughter to sleep in her own bed” implies that the speaker values power and conformity; saying “I miss mom” implies that the speaker values benevolence; saying “not wanting people to use my property without permissions” implies the speaker is self-directed and values security. The last example “I forgot how to be happy” results a small radar graph. It suggests that even the model could predict the overall polarity pretty well, there is still space to improve its capability of distinguishing different values.

Conclusion

We introduce a new dataset for human value modeling, VALUENET, which contains 21,374 scenarios in ten distinct human values. We also apply the learned value model from VALUENET to several EQ-related dialogue tasks. Our experiments show our approach and dataset provide a new way to control the dialogue system speaking style and numerically estimate one’s value preference. We hope that our results and dataset will stimulate more research in the important direction of building human-value-driven dialogue systems.
References


Kouloumpis, E.; Wilson, T.; and Moore, J. 2011. Twitter sentiment analysis: The good the bad and the omg! In Fifth International AAAI conference on weblogs and social media.


Lourie, N.; Bras, R. L.; and Choi, Y. 2021. Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-Life
Anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.


