Distilling Knowledge for Empathy Detection

Mahshid Hosseini and **Cornelia Caragea**
Computer Science
University of Illinois at Chicago
mhosse4@uic.edu, cornelia@uic.edu

Abstract

Empathy is the link between self and others. Detecting and understanding empathy is a key element for improving human-machine interaction. However, annotating data for detecting empathy at a large scale is a challenging task. This paper employs multi-task training with knowledge distillation to incorporate knowledge from available resources (emotion and sentiment) to detect empathy from the natural language in different domains. This approach yields better results on an existing news-related empathy dataset compared to strong baselines. In addition, we build a new dataset for empathy prediction with fine-grained empathy direction, seeking or providing empathy, from Twitter. We release our dataset for research purposes.

1 Introduction

Empathy is the ability to feel, understand, and correlate with the thoughts and feelings of another person (Decety and Jackson, 2004). Empathy enables us to build rapport with other people by acknowledging their cognitive state and making them feel that they are being heard and understood. Applications of analyzing and detecting empathy have been examined from numerous perspectives, including medical and healthcare (Decety and Fotopoulou, 2015; Williams et al., 2015; Raab, 2014), human-computer interaction (De Vicente and Pain, 2002; Buechel et al., 2018), neuroscience (Decety and Ickes, 2011), philosophy and psychology (Yan and Tan, 2014; Coplan and Goldie, 2011; Batson, 2009), and education (Virvou and Katsionis, 2003).

Social platforms facilitate expressing empathy and sharing of thoughts and information through natural language and text-based communication. Consequently, many people turn to social networks to share their experiences and feelings in different situations. Several psychological and social science studies have recently examined the relationship between users' empathetic ability in a social network and their behavioral patterns (Kardos et al., 2017; Morelli et al., 2017; Medeiros and Bosse, 2016; Reis et al., 2004). For example, Kardos et al. (2017) examined social networks and observed that more empathetic capabilities in users lead to a larger group of close friends. Morelli et al. (2017) and Medeiros and Bosse (2016) also showed that empathy as an individual's personality influences their ability to attract social ties.

To analyze and understand empathy at scale, it is important to devise models to detect empathy from the natural language. Effectively training such models depends on the presence of quality labeled data. However, annotating such data at scale is challenging due to the subjective nature of empathy (Decety and Jackson, 2004) and the high annotation costs. Consequently, existing datasets on the task of text-based empathy classification are small in size. To address the small data issue, we study the use of data-rich tasks related to empathy and utilize their correlation in a multi-task learning setup. Multi-task learning delivers an efficient means of using supervised data from multiple related tasks. It is beneficial for various relevant tasks to be learned jointly so that each task can benefit from the knowledge learned in other tasks (Fukuda et al., 2017; Zagoruyko and Komodakis, 2016; Ma et al., 2018).

To put forward the relevant tasks, we follow the notion of the correlation between empathy and emotion discussed by Szanto and Krueger (2019) and Hein and Singer (2008). Szanto and Krueger (2019) showed that empathy is correlated with affective and emotional expression. Hein and Singer (2008) also characterized empathy as *"an affective state, caused by sharing of the emotions of another person."* Therefore, we can expect that empathetic sentences are rich in emotion and sentiment. It can be seen from Table 1 that when expressing empathy, people often show emotional behavior. For exam-

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3713–3724
November 7–11, 2021. ©2021 Association for Computational Linguistics

NewsEmp	I'm sorry to hear that about Dakota's parents. Even when you are adult it must be hard to see your parents splitting up. No one wants that to happen and it's unfortunate that her parents couldn't work it out. I hope they are able to still remain civil around the kids and family. Just because it didn't work romantically doesn't mean it won't work at all. **Emotion**: sadness , **polarity**: negative	*empathetic*
	It's a shame that air pollution has potentially been linked to increased mental damage with young children. We often don't take into account all the damage that the fossil fuel companies have done to our society. We only praise them for creating the fuels we use but never tax them appropriately for all the damage that they cause us. **Emotion**: anger , disgust , **polarity**: negative	*none-empathetic*
TwittEmp	My granddaughter has Wilms Cancer stage 4, she has been fighting since January. I cry everyday. There is not much to say, anyone who outlives a child suffers heartache , and the grandparents suffers both for their child and their grandchild. **Emotion**: sadness , fear , **polarity**: negative	*seek*
	God Bless! The heartache!! My condolences! Lost my dad and brother to cancer also! I know the pain! Last respects are so important! So sorry! Comfort in Jesus **Emotion**: sadness , **polarity**: negative	*provide*
	Joanie Shawhan's just released book, In Her Shoes: Dancing in the Shadow of Cancer , is a collection of vignettes, highlighting the stories of everyday women with everyday lives interrupted by cancer—their challenges, heartbreaks, questions and... **Emotion**: anticipation , **polarity**: positive	*none*

Table 1: Samples from NewsEmp and TwittEmp datasets. Emotion-associated text is highlighted in respective colors of each emotion.

ple, sentences like *"I'm sorry to hear that about Dakota's parents"* or *"I cry everyday"* are rich in *sadness* emotion and *negative* sentiment polarity.

In this paper, we show that better performance can be obtained by leveraging external knowledge related to empathy: emotion and sentiment. To this end, we use multi-task training with knowledge distillation technique (Clark et al., 2019) to incorporate knowledge into empathetic content from emotion and sentiment. In particular, we utilize two available resources as the external knowledge to improve empathy prediction: (1) EmoNet (Abdul-Mageed and Ungar, 2017), an emotion detection dataset; and (2) SST (Socher et al., 2013) a sentiment classification dataset. We employ EmoNet and SST as single-task models to teach a multi-task model to detect empathy. We show that the multi-task training with knowledge distillation outperforms strong baselines on two empathy datasets, each collected from different platforms on different domains: news and health. Table 1 shows examples from these datasets—NewsEmp by Buechel et al. (2018) and TwittEmp our dataset created from Twitter on health posts.

We explore empathy at higher granularity of empathy versus non-empathy and lower granularity of seeking empathy versus providing empathy. Results of our experiments show that with the higher granularity, detecting empathy from the news context is more challenging than detecting empathy from the health domain. However, detecting empathy (from the health domain) at the seeking and providing granularity makes it more difficult for the models to detect empathy. This may imply that empathy detection in a fine-grained granularity requires more implicit reasoning, which is not present as surface-level lexical information.

Our contributions are as follows: (1) We propose to use multi-task training with knowledge distillation for empathy classification to incorporate emotion and sentiment knowledge into empathetic content (§3); (2) We achieve better performance on the news empathy reactions dataset (NewsEmp) (Buechel et al., 2018) culminating (on average) in +4% F1 score (§5.2). Moreover, we bridge the domain gap between the existing empathy datasets (e.g., NewsEmp (Buechel et al., 2018)) and our TwittEmp dataset by employing unsupervised domain adaptation, from news to health (§6). To our knowledge, we are the first to explore unsupervised domain adaptation for empathy detection; (3) We introduce TwittEmp (§4), a Twitter dataset of perceived empathy annotated with fine-grained empathy direction. We release our dataset[1] as a step towards to facilitate research in social domains.

2 Related Work

Numerous studies have discussed the importance of empathy and its impacts on individuals' physiological condition and medical health. The appli-

[1] https://github.com/Mahhos/KDempathy

cations of empathy and its benefit have been examined from numerous perspectives, including human-computer interaction (De Vicente and Pain, 2002; Virvou and Katsionis, 2003; Kort and Reilly, 2002), healthcare (Raab, 2014; Williams et al., 2015), psychology (Batson, 2009; Davis, 1983), cognitive science (Wakabayashi et al., 2006; Launay et al., 2015), and neuroscience (Carr et al., 2003; Singer and Lamm, 2009; Keysers et al., 2004). Empathy is shown to have correlation with gender and language, as well as behavior and culture (Chung and Bemak, 2002; Chung et al., 2010; Gungordu, 2017). Gungordu (2017) analyzed the impacts of gender and cultural orientations on individuals' empathetic expression and observed that women are more empathetic compare to men, and people from different cultures express empathy in diverse ways.

However, only recently, computational studies have been conducted on analyzing empathy from text (Sharma et al., 2020; Yang et al., 2019; Sedoc et al., 2019; Buechel et al., 2018; Abdul-Mageed et al., 2017; Khanpour et al., 2017) and from spoken dialogues (Alam et al., 2018; Pérez-Rosas et al., 2017; Fung et al., 2016). For example, Khanpour et al. (2017) proposed a neural network model to detect empathetic messages in health-related posts from lung and breast discussion boards in a cancer support network. Their work is different from ours as they only focus on high-level empathy presented in the text and do not detect the direction of empathy at a fine-grained level.

Abdul-Mageed et al. (2017) identified a pathogenic type of empathy by collecting ≈ 1.8M Facebook posts. Unlike our study, Abdul-Mageed et al. (2017) modeled the detection of empathy in a regression setup. Xiao et al. (2012) employed an n-gram language model based maximum likelihood strategy to detect empathetic utterances from clinical trial studies. Yang et al. (2019) recognized eleven functional roles for users participating in cancer support communities such as story sharer, welcomer, and support provider. Inspired by the Gaussian mixture model (McLachlan and Basford, 1988), Yang et al. (2019) defined a statistical model that clusters different session representations into a set of roles. Unlike our work, Yang et al. (2019) analyzed the behavioral features of users in online health communities. Wang et al. (2014) used engineered features through machine learning techniques to detect types of social support in an online health community and analyzed empathy as part of emotional support, not detecting empathy or the fine-grained empathy direction expressed in text.

For the text-based empathy prediction, to date, only three contributions (Hosseini and Caragea, 2021; Sharma et al., 2020; Buechel et al., 2018) previously built publicly available datasets, to our knowledge. Hosseini and Caragea (2021) used BERT to detect the direction of empathetic support from an online cancer network. Unlike our work, Hosseini and Caragea (2021) modeled the empathy direction at the sentence level, not considering the whole message expressing empathy (which usually contains more than one sentence; see Table 1). Sharma et al. (2020) employed a RoBERTa-based bi-encoder model to detect empathy in conversations in online mental health platforms. In contrast to our work, Sharma et al. (2020) focused on the level of communication (weak, strong, or no communication) in a response post and developed a framework of expressed empathy consisting of three communication mechanisms, emotional reactions, interpretations, and explorations. Buechel et al. (2018) also built a corpus of messages from people's written reactions to news articles. Other publicly available datasets addressed other tasks on empathy, such as empathetic dialogue generation (Rashkin et al., 2018), and learning word ratings for empathy (Sedoc et al., 2019).

3 Detecting Empathy

Detecting the empathy from textual input is challenging due to the scarcity of labeled training data. Manually annotating a corpus at a large scale is not a feasible solution either due to the task's difficulty and the high cost of the annotation process. Here, we propose to use multi-task learning with knowledge distillation and teacher annealing to leverage knowledge from available resources of sentiment and emotion to detect empathy.

3.1 Multi-Task Learning

In multi-task learning (MTL) (Liu et al., 2019; Caruana, 1997), a target task is learned by employing knowledge from related auxiliary tasks so that knowledge learned in one task is shared across all tasks. In our setting, the target task is empathy detection and the auxiliary tasks are emotion and sentiment classification. As in Liu et al. (2019), we build all the models on top of the pre-trained BERT language model (Devlin et al., 2018). In MTL, the bottom layers (corresponding to BERT) are shared across all three tasks, and the top layers are spe-

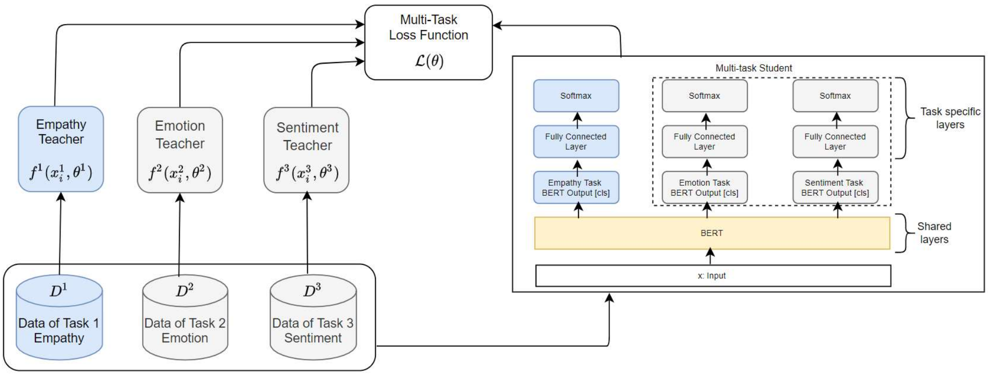

Figure 1: Multi-Task Training with Knowledge Distillation.

cific for each task as shown in Figure 1 (right side). Specifically, we use a fully connected layer for each task followed by softmax for classification.

During MTL training, examples from the three tasks are shuffled together (within minibatches) and the sum of the losses of all three tasks is minimized using backprop. That is, let $\mathcal{D}^\tau = \{(x_i^\tau, y_i^\tau)\}_i$ be the training set for task τ, where τ could be any of the three tasks (empathy, emotion, or sentiment). The loss of the MTL model with parameters θ is:

$$\mathcal{L}(\theta) = \sum_{\tau=1}^{3} \sum_{(x_i^\tau, y_i^\tau) \in \mathcal{D}^\tau} \ell(y_i^\tau, f^\tau(x_i^\tau, \theta)) \quad (1)$$

where $f^\tau(x_i^\tau, \theta)$ is the output of model θ on the input x_i^τ and ℓ is the cross-entropy loss. That is, the MTL model is optimized based on one-hot labels.

3.2 Multi-Task Learning with Knowledge Distillation and Teacher Annealing

Rather than optimizing the model based on one-hot labels, better training signal can be obtained from the data when distilling knowledge using a teacher-student framework, in which the student model learns the knowledge offered by the teachers' output. Thus, we propose to use the MTL model that distills knowledge from the auxiliary tasks into the target task, proposed by (Clark et al., 2019), which employs the idea of applying knowledge distillation (Ba and Caruana, 2014; Buciluă et al., 2006; Hinton et al., 2015) with the purpose that single-task models (teachers) teach a multi-task model (student) so that the student becomes better than the teachers. During training, as before, various tasks' examples are mixed jointly and the aggregated loss over all three tasks is minimized.

Formally, let $\mathcal{D}^\tau = \{(x_i^\tau, y_i^\tau)\}_i$ be the training set for task τ (empathy, emotion, or sentiment), as before. A single-task (teacher) model, denoted θ^τ, is trained on each task τ ($\tau = 1, 2, 3$), which produces output $f^\tau(x_i^\tau, \theta^\tau)$ on the input x_i^τ (see Figure 1). Then, a multi-task shared (student) model with parameters θ (right side of Figure 1) learns to imitate the output of the single-task (teacher) models θ^τ (left side of Figure 1). The loss of the multi-task (student) model becomes:

$$\mathcal{L}(\theta) = \sum_{\tau=1}^{3} \sum_{(x_i^\tau, y_i^\tau) \in \mathcal{D}^\tau} \ell(f^\tau(x_i^\tau, \theta^\tau), f^\tau(x_i^\tau, \theta)) \quad (2)$$

That is, the MTL model with knowledge distillation is optimized based on teachers' predictions.

Emulating the teacher model in knowledge distillation may limit the student model to transcend the teacher model. Clark et al. (2019) uses a training strategy called teacher annealing. That is, the MTL with knowledge distillation and teacher annealing combines gold-standard with predictions:

$$\mathcal{L}(\theta) = \sum_{\tau=1}^{3} \sum_{(x_i^\tau, y_i^\tau) \in \mathcal{D}^\tau} \ell(\lambda y_i^\tau + (1 - \lambda) f^\tau(x_i^\tau, \theta^\tau), \\ f^\tau(x_i^\tau, \theta)) \quad (3)$$

where λ is linearly increased from 0 to 1 over the course of training. This approach benefits the student model to outperform its teachers. We adopt this approach in our experiments.

4 Data

We incorporate knowledge from data-rich tasks of emotion and sentiment to detect empathy. We specifically use SST-2 (Socher et al., 2013) and EmoNet (Abdul-Mageed and Ungar, 2017). SST-2 is a binary dataset for sentiment analysis consisting of sentences from movie reviews and their sentiment (positive and negative). It has $67,349$ samples in the training set, 872 samples in the validation set, and $1,821$ samples in the test set. EmoNet is a dataset for emotion detection from Twitter. We use the EmoNet version that contains tweets annotated with Plutchik-8 emotions (joy, trust, fear, surprise, sadness, disgust, anger, anticipation). It has $41,669$ samples in training, $5,166$ samples in validation, and $5,214$ samples in test.

We incorporate knowledge from related tasks of emotion and sentiment to detect empathy using two datasets. These datasets are chosen from (1) different domains: news and health; and (2) different platforms: online news platforms and Twitter. Despite the significance of empathy in improving patients' positive feelings, only a few datasets are publicly available. We model empathy on the recent dataset by Buechel et al. (2018), leveraging available resources. We refer to this dataset as NewsEmp dataset. In addition, to experiment with a data from a different domain, we introduce TwittEmp, a new dataset of perceived empathy collected from Twitter. We describe the datasets below.

4.1 NewsEmp Dataset

NewsEmp is a dataset of empathic reactions to news stories released by Buechel et al. (2018). The dataset contains $1,860$ messages written in reaction to news articles rated with a numeric level of empathy and distress on a 7-point scale. Buechel et al. (2018) provided empathy binary labels, indicating if a message contains empathetic content or not. We leverage these labels to model empathy in a binary setting. We split the dataset into three sets of train, validation, and test with 80% of data used for training, 10% for validation and, 10% for test.

4.2 TwittEmp Dataset

We present our dataset of perceived empathy annotated by fine-grained empathy direction (seeking vs. providing). TwittEmp contains $3,000$ English tweets, which will be publicly available for further research in social domains.

Definitions of Seeking and Providing Empathy. Empathy needs one to embrace the subjective standpoint of the others (Decety and Jackson, 2004). We characterize seeking empathy as a need to be heard and understood. When people experience challenging situations, they need their feelings to be recognized and acknowledged. Providing empathy can be defined as the psychological perception of the individuals' feelings, thoughts, or attitudes who are enduring challenging experiences. Our definitions are derived in consultation with a psychologist and follow (Decety and Jackson, 2004) and online definitions of empathy.

4.2.1 Data Collection and Annotation

We collect a dataset of $3,000$ tweets from Twitter, which are annotated with three categories: *seeking-empathy*, *providing-empathy*, or *none*. We collect data related to the *cancer* topic using the Twitter streaming API, starting from July 2015 to August 2020. We employ filtering techniques to ensure that the collected tweets are likely to contain empathetic content. We specifically use the empathy and distress lexicon[2] by Sedoc et al. (2019), which consists of $9,356$ word types, each with associated empathy and distress ratings. The lexicon is context-independent; therefore, there are several words in the lexicon with high empathy ratings, such as gaza, zambia, myanmar, that do not correlate with our topic of interest (i.e., health). Consequently, we select 200 words with the highest empathy rating that are relevant to the health topic. The selected words and their corresponding empathy rating are presented in Appendix A.

We require that empathetic tweets contain at least one of the 200 high-rating empathy words plus "cancer". As part of the preprocessing, we remove duplicate tweets and replace links and usernames with $<$*URL*$>$ and $<$*USER*$>$, respectively.

To ensure the quality of annotations and reliability of the labels, we trained two graduate students through multiple iterations with a psychologist-in-the-loop for the initial round of labeling. Following prior studies (D'Mello, 2015; Fort, 2016), the annotation task was done iteratively. In each round, the annotators were asked to annotate 200 tweets and discussed the disagreements with researchers. 100% inter-annotator agreement (IAA) was obtained, measured by Cohen's kappa coefficient, after each round of discussions. After three initial rounds of annotations, the annotation continued until we get $1,000$ annotated samples per class of *seeking-empathy*, *providing-empathy*, and *none*. Finally, the last round of annotations was reviewed and finalized by one of the authors of this paper.

4.3 Characteristics of Datasets

Characteristics of TwittEmp compared with NewsEmp dataset are outlined in Table 2. As shown in Table 2, Buechel et al. (2018) modeled "intended" empathy as they obtained empathy scores from the *writer* of a text. In contrast, we study "perceived" fine-grained empathy from the *reader's* perspective. This allows us to examine

[2] http://www.wwbp.org/lexica.html

Dataset	Labels	Empathy Type	Domain	Platform	Size
NewsEmp	Empathy Distress	Intended	News	Online news platforms	1860
TwittEmp	Seek Provide	Perceived	Health	Twitter	3000

Table 2: Characteristics of TwittEmp compared with NewsEmp dataset by Buechel et al. (2018).

TwittEmp	NewsEmp
sorry for your loss	just read an article
passed away from cancer	I feel bad for
prayers are with you	Did you hear about
heart goes out to	in the middle east

Table 3: Most frequent noun phrases.

and model empathy from different perspectives.

Table 3 presents top frequent noun phrases (4-grams) in TwittEmp and NewsEmp datasets. Analyzing top noun phrases denotes a distinct theme and a storyline of each of these datasets. Unlike, NewsEmp which is collected from reactions to news stories, TwittEmp covers health-related content. For instance, *"Sorry for your loss, cancer has robbed our lives of some wonderful people."* represents the user's intention to provide empathetic support for others. In contrast, sentences like *"So I just read an article where 2 friends went diving to a place they shouldnt have and ended up dying. While they were using brand new equipment, I feel like idiots who take stupid risks and go to places where no humans should be, kind of deserve what ends up happening to them. If you dont sky dive, you never have to worry about going splat when your chute doesnt open"*, from NewsEmp, describes a reaction to a heartbreaking news story. Table 1 in §1 shows samples from NewsEmp and TwittEmp, along with their Plutchik-8 emotions and sentiment polarity.

The average length of a tweet in TwittEmp is around 37 words (max=62 words), while NewsEmp has an average message length of 82 words (max=163 words). TwittEmp also holds an average number of 3 sentences per tweet, while NewsEmp has an average number of 5 sentences per message. Figures 2a and 2c compare the tweet and message length distribution across TwittEmp and NewsEmp datasets, respectively. Figures 2b and 2d show the length distribution in the datasets per class. Comparing the two results suggests that NewsEmp often carries longer sentences.

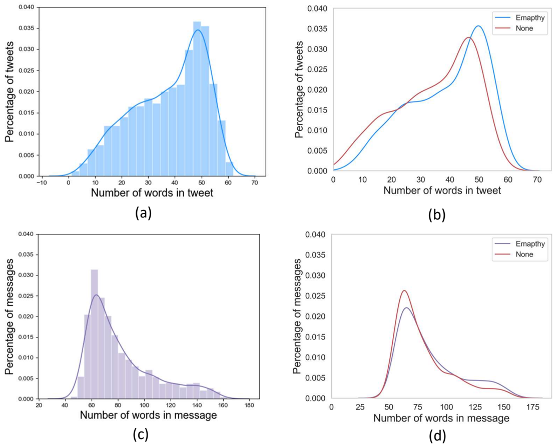

Figure 2: (a) Tweet length distribution across TwittEmp; (b) Tweet length distribution across each class. (c) Message length distribution across NewsEmp; (d) Message length distribution across each class.

5 Experiments

We model empathy in a binary setting in both datasets, detecting if a message contains empathetic content or not. For modeling empathy in the TwittEmp dataset, we keep tweets with labels *seeking-empathy*, and *providing-empathy* as positive samples and tweets in *none* class as negative samples. We then split the dataset into three sets of train, validation, and test with 80% of data used for training and the remaining 20% split equally for validation and test.

Detecting Fine-grained Empathy. Given a tweet, our goal is to classify it into one of the two categories of *seeking-empathy*, and *providing-empathy*. We create two classifiers in a binary setting, one to detect tweets seeking empathy and one to detect tweets providing empathy. For the seeking-classifier, we keep *seeking-empathy* as positive samples and combine the two classes of *none* and *providing-empathy* as negative samples. Similarly, to create the providing-classifier, we keep *providing-empathy* as positive samples and combine the two classes of *none* and *seeking-empathy* as negative samples. We then split the datasets, keeping 60% of data for the training set, 20% for validation, and 20% for the test set.

Model	Pr	Re	F-1
CNN	46.54	**85.06**	60.16
LSTM	38.89	28.05	43.33
BiLSTM	54.55	41.38	47.06
ConvLSTM	48.91	51.72	50.28
BERT	73.97	54.55	63.79
MT$_{SST}$	64.29	63.64	63.96
MT$_{EmoNet}$	61.70	58.21	62.12
MT$_{SST+EmoNet}$	62.37	60.59	64.42
KD$_{SST}$	66.98	71.72	68.27
KD$_{EmoNet}$	63.06	70.69	66.67
KD$_{SST+EmoNet}$	**67.68**	66.54	**68.41**

Table 4: Results on NewsEmp dataset.

Model	Pr	Re	F-1
CNN	84.71	83.50	83.91
LSTM	86.09	80.65	82.28
BiLSTM	**88.61**	72.91	79.74
ConvLSTM	73.97	**90.52**	83.38
BERT	85.60	84.55	84.07
MT$_{SST}$	80.65	84.40	83.06
MT$_{EmoNet}$	81.58	85.21	84.77
MT$_{SST+EmoNet}$	80.53	84.73	83.11
KD$_{SST}$	83.13	84.15	84.64
KD$_{EmoNet}$	86.13	83.33	**85.71**
KD$_{SST+EmoNet}$	82.42	85.77	84.56

Table 5: Results on TwittEmp empathy prediction.

5.1 Models

The details of the experiments are as follows. We contrast the multi-task learning with knowledge distillation and teacher annealing (§3.2) that learns from teachers' outputs and one-hot labels (denoted KD) (Eq. 3) with the multi-task learning (§3.1) that uses one-hot labels (denoted MT) (Eq. 1) and with the following baselines.

Standard Neural Methods. We experiment with **(1) CNN** (Kim, 2014), **(2) LSTM** (Hochreiter and Schmidhuber, 1997), **(3) ConvLSTM** a combination of the two previous models used in prior work on the empathetic message identification task (Khanpour et al., 2017), and **BiLSTM** (Hochreiter and Schmidhuber, 1997). All the neural models were trained with pre-trained 100d GloVe (Pennington et al., 2014) word embeddings. The best hyper-parameters reported by (Kim, 2014) are used for CNN. For the LSTM-based models, we used 128 hidden units and a dropout rate of 0.5 with a softmax layer on top to obtain the final predictions.

Pre-Trained Language Models. We fine-tune BERT (Devlin et al., 2018), in particular `bert-base-uncased`, with an added single linear layer on top of the `[CLS]` token.

5.2 Results

Our main results for the NewsEmp dataset (Buechel et al., 2018) are shown in Table 4. We observe that multi-task training with knowledge distillation and teacher annealing achieve clear improvements over the best BERT model and multi-task training. Starting with the single task BERT baseline with an F1 score of 63.79, distilling knowledge from SST in 'KD_{SST}' improves the F1 score to 68.27 (+4.48). Distilling knowledge from EmoNet in 'KD_{EmoNet}' also results in improvements: 66.67 (+2.88). When both teachers are used simultaneously in '$KD_{SST+EmoNet}$', the F1 score further improves to 68.41 (+4.62), suggesting that the two tasks provide a complementary signal that is beneficial for the empathy prediction task. The results also suggest that using teachers' output distribution over classes (i.e., 'KD_*') instead of one-hot labels (i.e., 'MT_*') positively improves the performance. The results indicate that teachers' outputs help to gain further information on training examples.

Table 5 shows the main results on TwittEmp dataset empathy detection, where we see that leveraging knowledge from EmoNet improves the performance over 'KD_{SST}' and '$KD_{SST+EmoNet}$' on this dataset. The observed performance could be attributed to the EmoNet's content, which contains general tweets, resembling the TwittEmp dataset's content. The results also suggest that MT+KD outperforms MT with one-hot labels. Comparing the results with Table 4 suggests that modeling empathy in NewsEmp is more challenging compared to TwittEmp. This may be due to the longer sentences in NewsEmp, which are harder to classify.

Table 6 shows the main results on TwittEmp dataset fine-grained empathy direction. We see similar patterns for both the seek and provide classifiers. Each multi-task training improves model performance. 'KD_{EmoNet}' is more effective on the performance showing that leveraging knowledge from more related tasks helps to enhance the performance to a greater extent. We can also observe that detecting empathy at a finer granularity is more challenging compared to coarse-grained empathy detection. This may denote that modeling empathy at the fine-grained level requires more implicit reasoning, making modeling empathy more challenging. Similar to previous tasks, we can see that leveraging knowledge distillation provides more information than solely employing one-hot labels resulting in improved performance.

	seek			*provide*		
	Pr	Re	F-1	Pr	Re	F-1
CNN	76.80	60.40	60.87	68.93	77.20	70.83
LSTM	50.47	62.43	56.59	56.32	78.40	65.58
BiLSTM	54.49	38.81	58.33	67.58	77.48	70.21
ConvLSTM	50.16	**63.60**	56.08	53.09	76.00	65.65
BERT	78.07	61.60	66.51	76.94	75.40	77.16
MT$_{SST}$	78.20	60.39	67.69	75.70	81.60	78.04
MT$_{EmoNet}$	77.37	60.05	67.34	76.51	79.20	78.36
MT$_{SST+EmoNet}$	77.16	59.81	67.01	76.19	82.20	78.49
KD$_{SST}$	79.58	60.80	67.73	76.14	81.74	78.56
KD$_{EmoNet}$	77.32	61.09	**68.57**	**77.21**	**82.57**	**79.48**
KD$_{SST+EmoNet}$	**79.89**	59.60	67.97	76.68	81.15	79.06

Table 6: Empathy direction identification on TwittEmp.

6 Unsupervised Domain Adaptation

Empathy annotations are not always available. Nevertheless, from a psychological perspective, these annotations would be valuable to understand users' empathetic profile during hard situations. In this section, we examine methods to leverage supervision from existing empathy datasets (i.e., NewsEmp (Buechel et al., 2018)) in providing labels for the TwittEmp empathy dataset. We set up this task as *unsupervised* domain adaptation; NewsEmp is considered as the labeled source domain (SRC), and our TwittEmp dataset is considered as the unlabeled target domain (TRG). Below, we provide details on the adaptation method.

We employ BERT as the classifier. As Han and Eisenstein (2019), we mainly focus on using pre-training techniques that facilitate effective transfer between different domains. We experiment with pre-training on dynamic masked language modeling by leveraging unsupervised data from different domains and platforms: (1) **Unsupervised EMPATHETICDIALOGUES** (Rashkin et al., 2018) is a dataset of crowdsourced conversations from emotional situations; (2) **Unsupervised Twitter:** we collect a large amount of unsupervised data from Twitter in the health domain using the words as before from the lexicon by Sedoc et al. (2019); (3) **Unsupervised GoEmotions.** GoEmotions (Demszky et al., 2020) is a large-scale emotion detection dataset from Reddit comments; (4) **Unsupervised ISEAR** ISEAR (Scherer and Wallbott, 1994) is a survey on emotion antecedents and reactions to emotional situations; (5) **Unsupervised DailyDialog**. DailyDialog (Li et al., 2017) comprises dialogues from educational websites.

For comparison, we experiment with different systems: (1) SOURCE-ONLY: the source domain is used for fine-tuning BERT (the training portion) and the target domain is used for the evaluation (the test portion); (2) TARGET-ONLY: the target domain is used for both training and evaluation of BERT. These results are adopted from Table 5 to show the performance in-domain; (3) PRETRAIN-$*$: BERT undertakes dynamic masked language modeling (MLM) pre-training by leveraging a large set of unsupervised data from task/dataset $*$, i.e., EMPATHETICDIALOGUES, GoEmotions, ISEAR, Twitter, and DailyDialog (one at a time) and then BERT is trained (fine-tuned) on the source domain (the training portion), and ultimately evaluated on the target domain (the test portion) (Han and Eisenstein, 2019).

	Pr	Re	F-1
SOURCE-ONLY	51.12	94.71	67.24
PRETRAIN-$*$			
EMPATHETICDIALOGUES	**52.71**	94.80	67.73
DailyDialog	52.70	**98.78**	**68.72**
GoEmotions	51.85	96.74	67.51
ISEAR	52.63	97.56	68.37
Twitter	51.29	96.34	67.94
TARGET-ONLY	85.60	84.55	84.07

Table 7: Unsupervised domain adaptation.

6.1 Results

Table 7 presents the results of the unsupervised domain adaptation. Generally, we do not observe a noticeable improvement in performance over the SOURCE-ONLY baseline using EMPATHETICDIALOGUES, GoEmotions, and Twitter. Leveraging unsupervised data from DailyDialog improves performance by 1.48%. The results also suggest that incorporating ISEAR yields 1.13% improvement in performance. It can also be seen from Table 7 that pre-training adds a small improvement in recall in most of the settings. We can posit that incorporating knowledge from a different domain can be beneficial to get most of the relevant results (less false negatives). But still, we can see a big gap between PRETRAIN-$*$ and TARGET-ONLY. The results suggest that more explicit strategies may be needed for empathy to enable domain adaptation.

7 Conclusion

In this study, we show that distilling knowledge from available related resources on emotion and sentiment can be effectively used to inform empathy classification. We use multi-task training with knowledge distillation technique to incorporate knowledge into empathetic content from EmoNet and SST. This approach achieves better results on two datasets from different domains. We also show promising results on unsupervised domain adaptation for empathy detection which represents an interesting future direction.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 718–728.

Muhammad M Abdul-Mageed, Anneke Buffone, Hao Peng, Johannes Eichstaedt, and Lyle Ungar. 2017. Recognizing pathogenic empathy in social media. In *Eleventh International AAAI Conference on Web and Social Media*.

Firoj Alam, Morena Danieli, and Giuseppe Riccardi. 2018. Annotating and modeling empathy in spoken conversations. *Computer Speech & Language*, 50:40–61.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? In *Advances in neural information processing systems*, pages 2654–2662.

C Daniel Batson. 2009. These things called empathy: eight related but distinct phenomena.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

Sven Buechel, Anneke Buffone, Barry Slaff, Lyle H. Ungar, and João Sedoc. 2018. Modeling empathy and distress in reaction to news stories. *CoRR*, abs/1808.10399.

Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C Mazziotta, and Gian Luigi Lenzi. 2003. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the national Academy of Sciences*, 100(9):5497–5502.

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

Rita Chi-Ying Chung and Fred Bemak. 2002. The relationship of culture and empathy in cross-cultural counseling. *Journal of Counseling & Development*, 80(2):154–159.

Winnie Chung, Sherilynn Chan, and Tracy G Cassels. 2010. The role of culture in affective empathy: Cultural and bicultural differences. *Journal of Cognition and Culture*, 10(3-4):309–326.

Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc V Le. 2019. Bam! born-again multi-task networks for natural language understanding. *arXiv preprint arXiv:1907.04829*.

Amy Coplan and Peter Goldie. 2011. *Empathy: Philosophical and psychological perspectives*. Oxford University Press.

Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Angel De Vicente and Helen Pain. 2002. Informing the detection of the students' motivational state: an empirical study. In *International Conference on Intelligent Tutoring Systems*, pages 933–943. Springer.

Jean Decety and Aikaterini Fotopoulou. 2015. Why empathy has a beneficial impact on others in medicine: unifying theories. *Frontiers in behavioral neuroscience*, 8:457.

Jean Decety and William John Ickes. 2011. *The social neuroscience of empathy*. Social Neuroscience.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*, 3(2):71–100.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sidney K D'Mello. 2015. On the influence of an iterative affect annotation approach on inter-observer and self-observer reliability. *IEEE Transactions on Affective Computing*, 7(2):136–149.

Karën Fort. 2016. *Collaborative Annotation for Reliable Natural Language Processing: Technical and Sociological Aspects*. John Wiley & Sons.

Takashi Fukuda, Masayuki Suzuki, Gakuto Kurata, Samuel Thomas, Jia Cui, and Bhuvana Ramabhadran. 2017. Efficient knowledge distillation from an ensemble of teachers. In *Interspeech*, pages 3697–3701.

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Yan Wan, and Ho Yin Ricky Chan. 2016. Zara the supergirl: An empathetic personality recognition system. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 87–91.

Nahide Gungordu. 2017. *Are empathy traits associated with cultural orientation?: a cross-cultural comparison of young adults*. Ph.D. thesis, University of Alabama Libraries.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4229–4239.

Grit Hein and Tania Singer. 2008. I feel how you feel but not always: the empathic brain and its modulation. *Current opinion in neurobiology*, 18(2):153–158.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Lstm can solve hard long time lag problems. In *Advances in neural information processing systems*, pages 473–479.

Mahshid Hosseini and Cornelia Caragea. 2021. It takes two to empathize: One to seek and one to provide.

Peter Kardos, Bernhard Leidner, Csaba Pléh, Péter Soltész, and Zsolt Unoka. 2017. Empathic people have more friends: Empathic abilities predict social network size and position in social network predicts empathic efforts. *Social Networks*, 50:1–5.

Christian Keysers, Bruno Wicker, Valeria Gazzola, Jean-Luc Anton, Leonardo Fogassi, and Vittorio Gallese. 2004. A touching sight: Sii/pv activation during the observation and experience of touch. *Neuron*, 42(2):335–346.

Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. 2017. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Barry Kort and Rob Reilly. 2002. An affective module for an intelligent tutoring system. In *International Conference on Intelligent Tutoring Systems*, pages 955–962. Springer.

Jacques Launay, Eiluned Pearce, Rafael Wlodarski, Max van Duijn, James Carney, and Robin IM Dunbar. 2015. Higher-order mentalising and executive functioning. *Personality and individual differences*, 86:6–14.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1930–1939.

Geoffrey J McLachlan and Kaye E Basford. 1988. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York.

Lenin Medeiros and Tibor Bosse. 2016. Empirical analysis of social support provided via social media. In *International Conference on Social Informatics*, pages 439–453. Springer.

Sylvia A Morelli, Desmond C Ong, Rucha Makati, Matthew O Jackson, and Jamil Zaki. 2017. Empathy and well-being correlate with centrality in different social networks. *Proceedings of the National Academy of Sciences*, 114(37):9843–9847.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435.

Kelley Raab. 2014. Mindfulness, self-compassion, and empathy among health care professionals: a review of the literature. *Journal of health care chaplaincy*, 20(3):95–108.

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*.

Harry T Reis, Margaret S Clark, and John G Holmes. 2004. Perceived partner responsiveness as an organizing construct in the study of intimacy and closeness.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

João Sedoc, Sven Buechel, Yehonathan Nachmany, Anneke Buffone, and Lyle Ungar. 2019. Learning word ratings for empathy and distress from document-level user responses. *arXiv preprint arXiv:1912.01079*.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Tania Singer and Claus Lamm. 2009. The social neuroscience of empathy. *Annals of the New York Academy of Sciences*, 1156(1):81–96.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Thomas Szanto and Joel Krueger. 2019. Introduction: empathy, shared emotions, and social identity. *Topoi*, 38(1):153–162.

Maria Virvou and George Katsionis. 2003. Relating error diagnosis and performance characteristics for affect perception and empathy in an educational software application. In *Proceedings of the 10th International Conference on Human Computer Interaction (HCII) 2003*, pages 22–27.

Akio Wakabayashi, Simon Baron-Cohen, Sally Wheelwright, Nigel Goldenfeld, Joe Delaney, Debra Fine, Richard Smith, and Leonora Weil. 2006. Development of short forms of the empathy quotient (eq-short) and the systemizing quotient (sq-short). *Personality and individual differences*, 41(5):929–940.

Xi Wang, Kang Zhao, and Nick Street. 2014. Social support and user engagement in online health communities. In *International Conference on Smart Health*, pages 97–110. Springer.

Brett Williams, Ted Brown, Lisa McKenna, Claire Palermo, Prue Morgan, Debra Nestel, Richard Brightwell, Susan Gilbert-Hunt, Karen Stagnitti, Alexander Olaussen, et al. 2015. Student empathy levels across 12 medical and health professions: an interventional study. *Journal of Compassionate Health Care*, 2(1):4.

Bo Xiao, Dogan Can, Panayiotis G Georgiou, David Atkins, and Shrikanth S Narayanan. 2012. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4. IEEE.

Lu Yan and Yong Tan. 2014. Feeling blue? go online: an empirical study of social support among patients. *Information Systems Research*, 25(4):690–709.

Diyi Yang, Robert E Kraut, Tenbroeck Smith, Elijah Mayfield, and Dan Jurafsky. 2019. Seekers, providers, welcomers, and storytellers: Modeling social roles in online health communities. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*.

A Words and Empathy Ratings

Tables 8 presents the selected words and their corresponding empathy rating chosen from (Sedoc et al., 2019).

Word	Empathy Rating	Word	Empathy Rating	Word	Empathy Rating	Word	Empathy Rating
healing	6.596821	counselling	5.797798	died	5.406086	family	5.229086
grieve	6.501427	heart	5.796642	surive	5.395749	help	5.218461
heartbreaks	6.432908	families	5.794646	impacted	5.395483	incredible	5.21781
empathize	6.358549	disease	5.765653	happiness	5.394868	surrendered	5.21266
heartbreaking	6.353829	souls	5.764528	devastating	5.38774	aid	5.208723
sorrow	6.299757	illness	5.760072	healthcare	5.386095	adorable	5.204847
heartache	6.268606	lungs	5.749102	guilt	5.383146	fathers	5.202694
wounds	6.268003	hearts	5.742606	dear	5.379856	survivors	5.201591
wounded	6.265173	heals	5.72467	livelihoods	5.377989	revengeful	5.19882
empathic	6.25796	tumor	5.718558	diagnosis	5.376528	melted	5.197083
scars	6.248851	sacrifices	5.716784	emotions	5.37484	cry	5.189401
grieving	6.243041	losses	5.714543	rehabilitation	5.370863	sympathize	5.188904
healer	6.208337	spiritual	5.709272	compassionate	5.370323	ambulances	5.182411
heal	6.195529	depression	5.704418	die	5.369835	patients	5.180013
cancer	6.190857	mistreated	5.702716	warmth	5.364146	disabilities	5.17669
suffering	6.150593	poverty	5.691481	scarred	5.363813	regains	5.173569
hardships	6.131669	bleed	5.676499	overcome	5.359576	hurting	5.169916
blisters	6.116005	childhood	5.671929	darkness	5.35705	hurt	5.168229
journey	6.099534	nutrients	5.655005	burned	5.350904	lung	5.163016
trauma	6.072249	liver	5.649638	learnings	5.350009	experiences	5.159558
loss	6.069548	breast	5.649456	saddened	5.349804	grandmother	5.151736
prayers	6.069082	depression	5.648896	comfort	5.348674	honey	5.148947
empathy	6.059864	loved	5.613978	oppressors	5.348175	extinction	5.146607
soul	6.059769	tragedy	5.6093	emotionally	5.335611	survivor	5.139349
distressed	6.012959	graves	5.598697	crushing	5.331666	children	5.136825
grief	6.004203	sacrifice	5.592741	wonderful	5.328334	deaths	5.136596
decease	5.990951	memories	5.58466	lifes	5.325804	childrens	5.13552
sadness	5.979282	rescued	5.583999	destruction	5.302191	birth	5.134894
compassion	5.970341	rough	5.583473	amazing	5.301863	struggle	5.130493
ilness	5.953731	cried	5.574442	tragic	5.296545	damaged	5.128968
sicknesses	5.917092	rescuers	5.569765	damage	5.295643	tear	5.127956
extinctions	5.912342	innocents	5.567511	memorial	5.292897	weight	5.123735
heartbroken	5.910799	struggles	5.563548	felt	5.288165	humans	5.123192
fellings	5.909871	mortality	5.530533	gorgeous	5.285042	understanding	5.121832
journeys	5.908955	braves	5.517905	hospitalized	5.277785	energy	5.119498
wound	5.90535	diseases	5.517528	scarring	5.276179	walk	5.117874
emotional	5.905242	burns	5.496034	cries	5.275641	needy	5.117693
hardship	5.896923	suicide	5.478393	sickness	5.269866	understand	5.111647
pain	5.864645	health	5.475095	painful	5.268242	killed	5.110344
heathrow	5.859716	resilience	5.475025	home	5.259529	behavioral	5.107367
painless	5.85916	prayer	5.469543	death	5.255763	transforming	5.106391
tears	5.847304	dying	5.465345	positivity	5.253156	hurdles	5.106323
familes	5.845559	love	5.459695	feelings	5.250849	enemy	5.106023
keepsakes	5.827441	illnes	5.45155	helpless	5.246301	hospital	5.100527
gratitude	5.826963	lost	5.447177	contact	5.24313	biology	5.098369
belongings	5.824559	regain	5.429861	beloved	5.243001	protect	5.095969
devastated	5.822644	bless	5.427919	mothers	5.242946	crushed	5.093634
tragedies	5.820653	bleeding	5.417132	emotion	5.237587	despair	5.091182
traumatic	5.817914	hospitals	5.415503	feel	5.236701	guardianship	5.083399
relatable	5.812753	dies	5.41015	hits	5.233051	burning	5.083041

Table 8: Selected Words and Empathy Ratings.