HEAL: A Knowledge Graph for Distress Management Conversations

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

Abstract

The demands of the modern world are increasingly responsible for causing psychological burdens and bringing adverse impacts on our mental health. As a result, neural conversational agents with empathetic responding and distress management capabilities have recently gained popularity. However, existing end-to-end empathetic conversational agents often generate generic and repetitive empathetic statements such as *"I am sorry to hear that"*, which fail to convey specificity to a given situation. Due to the lack of controllability in such models, they also impose the risk of generating toxic responses. Chatbots leveraging reasoning over knowledge graphs is seen as an efficient and fail-safe solution over end-to-end models. However, such resources are limited in the context of emotional distress. To address this, we introduce **HEAL**, a knowledge graph developed based on 1M distress narratives and their corresponding consoling responses curated from Reddit. It consists of 22K nodes identifying different types of stressors, speaker expectations, responses, and feedback types associated with distress dialogues and forms 104K connections between different types of nodes. Each node is associated with one of 41 affective states. Statistical and visual analysis conducted on HEAL reveals emotional dynamics between speakers and listeners in distress-oriented conversations and identifies useful response patterns leading to emotional relief. Automatic and human evaluation experiments show that HEAL's responses are more diverse, empathetic, and reliable compared to the baselines.

Introduction

Demands of the modern world are increasingly responsible for causing psychological burdens and bringing adverse impacts on our mental health. Distress refers to a discomforting emotional state experienced by an individual in response to a specific personal stressor or demand that results in harm, either temporary or permanent to the person (Ridner 2004). Such stressors include separation from loved ones, interpersonal conflicts, certain mental health conditions such as depression, under-performing at work, and sleep problems such as insomnia. A study by Almeida et al. (2002), which measured multiple aspects of daily stressors of a U.S. national sample of 1,031 adults through daily telephone interviews, revealed they experienced at least one daily stressor on 40% of the study days. People usually tend to share such experiences in daily conversations. Thus, embedding open-domain conversational agents or chatbots with appropriate empathetic responding capabilities to address such distressful situations has gained much interest (Rashkin et al. 2019; Lin et al. 2019; Majumder et al. 2020; Xie and Pu 2021).

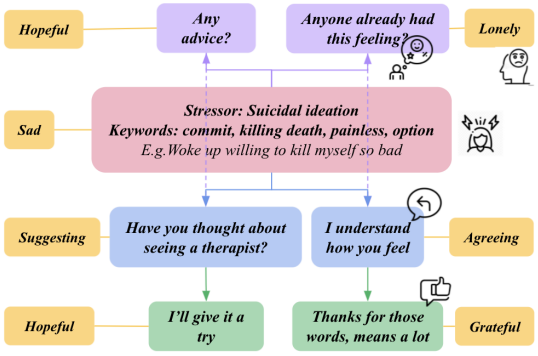


Figure 1: An illustration of part of *HEAL*. The red, purple, blue, green, and yellow nodes represent the stressors, speaker expectations, response and feedback types, and associated affective states respectively.

With the development of sophisticated neural network architectures such as the transformer (Vaswani et al. 2017) and pre-trained language models such as BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019a) and GPT-3 (Brown et al. 2020), fine-tuning neural response generation models on unstructured text has become one of the common approaches to build chatbots. Though it avoids most of the limitations with strictly rule-based methods and enables chatbots to largely generalize to unseen domains, the lack of controllability and the black-box nature make these models less reliable and fail-safe (d'Avila Garcez and Lamb 2020). This is especially problematic when the user is undergoing a distressful situation where he is sensitive to misinformation and inappropriate comments. A recent example is Microsoft's Tay bot that started producing unintended, offensive, and racial tweets denying the Holocaust after learning from racist and offensive information from Twitter (Lee 2016).

As a result, there is a growing interest to use knowledge (Zhu et al. 2017; Liu et al. 2018; Han et al. 2015) and commonsense reasoning (Zhou et al. 2018; Young et al. 2018)

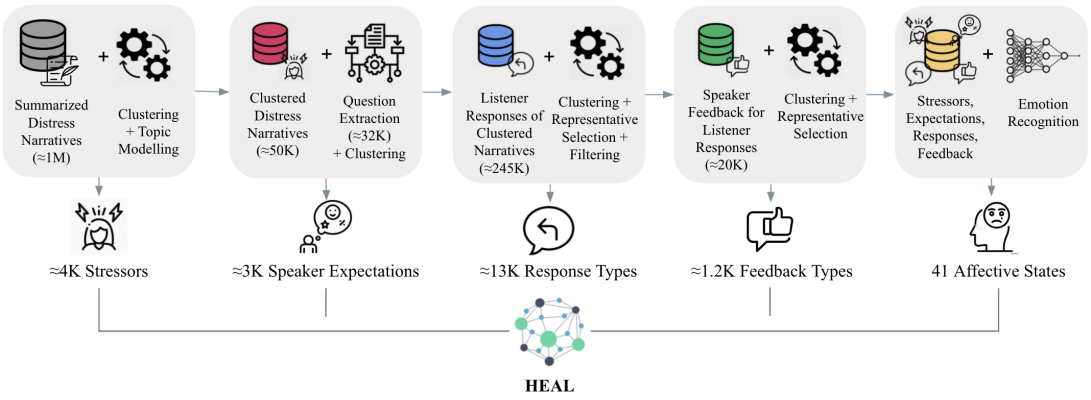


Figure 2: Step by step process for developing the knowledge graph, HEAL.

over graph-based representations to generate appropriate and informative responses to conversations. Compared to training over unstructured text, the use of graph-based representations offers more controllability and interpretability to the generated responses, thus limiting inappropriate and unreliable content. Identification of relatable topics in the knowledge graph makes it possible to direct the conversational flow along predictable routes, while also providing the ability to strategically diversify responses (Liu et al. 2019b).

Though large-scale knowledge graphs such as ConceptNet (Speer, Chin, and Havasi 2017) and ATOMIC (Sap et al. 2019) exist, they mainly assist in open-domain conversation generation by capturing factual knowledge and embedding chatbot models with simple commonsense reasoning capabilities. Since they were not developed to capture norms of empathetic exchanges, this field lacks linguistic resources and models to assist distress management and empathetic response generation. And none has ever attempted to generate knowledge graphs to represent whole dialogues with relations between context-response pairs. To address such limitations, we introduce **HEAL** (meaning **H**ealing, **E**mpathy, and **A**ffect **L**earning), a knowledge graph for distress management conversations, developed by analyzing narratives of stressful events and corresponding response threads curated from a carefully chosen set of subreddits.

HEAL consists of five types of nodes: **1)** ***stressors***: causes inflicting distress; **2)** ***expectations***: commonly asked questions by the speakers in the distress narratives; **3)** ***response types***: most frequent types of responses given by the listeners to address different stressors; **4)** ***feedback types***: common feedback types provided by the speakers following a response; and **5)** ***affective states***: emotional states associated with each node. Speakers here are the ones undergoing a distressful situation (the ones who start the conversation by posting on Reddit) and the listeners are the commentors to such posts. An illustration of a typical stressor in HEAL is shown in Figure 1. HEAL, which constitutes topics related to distress can accurately depict the underlying context in a distress-oriented conversation and thus enable dialogue models to retrieve responses more specific to the context. Also information such as whether such responses lead to positive or negative feedback and whether they address implicit expectations of the person under distress can result in the selection of more appropriate and useful responses. As depicted in Figure 2, we followed a series of steps including summarization, clustering, topic modeling, and emotion classification to develop HEAL from over 1M distress dialogues curated from Reddit. This resulted in the identification of ≈4K stressors, ≈3K speaker expectations, ≈13K response types, ≈1.2K feedback types, and associated affective states. The final graph constitutes 22,037 nodes and 104,004 connections between different types of nodes.

By conducting statistical and visual analysis on HEAL, we were able to discover emotional dynamics between speakers and listeners and favorable response types that lead to emotion de-escalation. We also tested the utility of HEAL in the downstream task of generating empathetic responses to a given distressful situation. We developed a retrieval-based model using the knowledge graph and compared its performance using automatic and human evaluation against two state-of-the-art empathetic conversational agents: one developed by Xie and Pu (2021); and Blender (Roller et al. 2021). The results showed that the responses retrieved using the knowledge graph in a ranked manner outperform the responses generated by the others in terms of diversity and empathetic appropriateness. Using a case study, we also show that the responses retrieved by HEAL are more reliable than neural response generation models. Our main contributions include 1) the development of a large-scale knowledge graph, **HEAL**, identifying different types of stressors, speaker expectations, response and feedback types, and affective states associated with distress dialogues; 2) use of statistical and visual analysis to identify emotional dynamics between speakers and listeners and favorable response patterns leading to emotion de-escalation; and 3) evaluating the usefulness of HEAL in retrieving more empathetically appropriate, diverse and reliable utterances in response to emotional distress. [1]

Related Work

Knowledge graphs have attracted the attention of the natural language processing community due to their usefulness

[1]Code and data available at github.com/anuradha1992/HEAL.

in understanding natural language input. This is boosted by the recent advent of linked open data such as DBPedia (Auer et al. 2007) and Google knowledge graph.[2] YAGO (Fabian et al. 2007), Freebase (Bollacker et al. 2008), and Wikidata (Vrandečić and Krötzsch 2014) are some other examples of knowledge graphs built on general knowledge extracted from the web. More recent knowledge graphs such as ConceptNet (Speer, Chin, and Havasi 2017), ATOMIC (Sap et al. 2019), and ASER (Zhang et al. 2020) focus on representing different types of commonsense knowledge. Works by Liu et al. (2018) and Zhang et al. (2020) leverage the factoid and commonsense knowledge present in these graphs to develop open-domain conversational agents that produces more semantic and informative responses.

Though the above resources are useful in the development of knowledge-aware conversational agents and those with the ability to reason (Zhou et al. 2018), often these graphs address open-domain entities and relationships and commonsense reasoning built upon them. They do not capture the norms of emotional reasoning and empathetic response generation. HEAL extends the above limitations by establishing relationships between stressors, speaker expectations, responses, feedback, and affective states and linking prompt-response-feedback tuples to identify responses that could potentially result in favorable feedback and address implicit expectations of those under distress.

Methodology

Dataset Curation

Publicly available emotional dialogue datasets such as EmpatheticDialogues (Rashkin et al. 2019), EmotionLines (Hsu et al. 2018) and EmoContext (Chatterjee et al. 2019), mostly consist of open-domain and daily conversations created in an artificial setting or curated from movie/TV subtitles. Real counseling conversation datasets used to conduct recent research (Althoff, Clark, and Leskovec 2016; Zhang and Danescu-Niculescu-Mizil 2020) are not directly accessible due to ethical reasons. Thus, we curated a new dataset from Reddit, containing dialogues that discuss real-world distressful situations. We chose Reddit since it is publicly accessible and peers actively engage in such platforms to support others undergoing mental distress.

We used the Pushshift API (Baumgartner et al. 2020) to collect and process dialogue threads from a carefully selected set of 8 subreddits: *mentalhealthsupport*; *offmychest*; *sad*; *suicidewatch*; *anxietyhelp*; *depression*; *depressed*; and *depression_help*, which are popular among Reddit users to vent their distress. We explicitly extracted the dialogue turn-taking structure out of these threads by matching author names and subjected these conversations to a rigorous data cleaning pipeline, which included removal of profanity from listener responses. By this, we were able to curate $1,275,486$ dyadic conversations with $3,396,476$ dialogue turns (on average 2.66 turns per dialogue). The data preprocessing pipeline and the dataset's descriptive statistics are included in the appendices. We used 80% of the dialogues to derive the knowledge graph and retained 10% of the dialogues each for validation and testing downstream tasks.

[2] en.wikipedia.org/wiki/Google_Knowledge_Graph

Summarization

The distress narratives curated from Reddit are typically lengthy (on average 84.89 tokens per turn) and some exceed the input token length for certain pre-trained language model-based architectures such as sentence-BERT (Reimers and Gurevych 2019). Therefore, we investigated various summarization algorithms that can be used to generate summaries preserving the essence of the narrative.

We investigated extractive and abstractive summarization techniques to address this issue (Tas and Kiyani 2007). Of them, abstractive summarization methods are mainly trained and tested on structured documents such as news articles and are known to perform poorly on not as structured texts (Peng et al. 2021). Therefore, we selected five different extractive summarization methods: a custom implementation of SMMRY—the algorithm behind Reddit's TLDR bot (https://smmry.com); and four different pre-trained models—BART (Lewis et al. 2020), GPT-2 (Radford et al. 2019), XLNET (Yang et al. 2019), and T5 (Raffel et al. 2020) for modelling content importance. We manually rated the summaries generated by the above methods on a sample of 100 Reddit distress narratives as *Good*, *Okay*, and *Bad* (results are detailed in the appendices). The highest percentage of summaries rated as *Good* were generated by the SMMRY algorithm. Hence it was selected to summarize lengthy dialogue turns (turns with ≥ 100 tokens). Approximately 43% of the dialogue turns were summarized using this.

Agglomerative Clustering

Since manual annotation is costly and time consuming specially when applied to a large-scale dataset, we decided to use automatic clustering to identify clearly distinguishable types of stressors, expectations, responses, and feedback types from the Reddit distress dialogues. For this purpose, we used "Agglomerative Clustering" tuned for large datasets (Murtagh and Legendre 2014). It recursively merges pairs of clusters that minimally increase a given linkage distance. The linkage distance was computed using the cosine similarity between pairs of embeddings generated by Sentence-BERT (Reimers and Gurevych 2019) since the resulting embeddings have shown to be of high quality and working substantially well for document-level embeddings. The choice of using agglomerative clustering over other clustering methods is explained in detail in the appendices.

Identification of Stressors

We experimented with 8 similarity thresholds from 0.6 to 0.95 with 0.05 increments to cluster distress narratives. Though various cluster quality metrics such as the Silhouette coefficient (Rousseeuw 1987), Dunn index (Misuraca, Spano, and Balbi 2019), and average point-to-centroid cosine distance, were computed for each threshold to select an optimal similarity threshold, manual inspection on a subset of 10 clusters at each threshold and cluster visualization revealed that those metrics do not work best for this dataset

Stressor	Keywords extracted
Suicidal ideation	*commit, killing, death, painless, option*
Anxiety attacks	*anxiety, anxious, attacks, social, attack*
Weight gain	*eating, weight, eat, lose, fat*
Loneliness	*lonely, surround, connect, isolated, social*
Failing college	*study, college, class, semester, failing*
Alchoholic	*drinking, drink, alcohol, drunk, sober*
US election	*trump, president, donald, election, war*
Covid19	*covid, 19, pandemic, shambolic, brought*

Table 1: Some stessors identified in the clusters of distress narratives using TF-IDF.

(Above metrics are known to work best only for datasets having convex-shaped clusters). Results of manual inspection conveyed that the stressors identified at higher thresholds such as 0.95 and 0.9 are too specific and those below 0.8 are too vague (cluster quality metrics and topics discovered through manual inspection at each threshold are included in the appendices). This resulted in selecting an optimal threshold of 0.85. At this threshold, 4.93% of the distress narratives ($47,109$ narratives in total) were separated in to $4,363$ clusters. After applying TF-IDF based topic modeling on these clusters, we uncovered some clearly distinguishable stressors, which further validated the goodness of clustering. Table 1 shows some stressors identified in this process.

Expectations, Responses and Feedback Types

After clustering distress narratives and identifying their respective topics, we extracted questions asked in the clustered distress narratives using a simple string search for sentences containing "?". Corresponding responses and associated feedback were also extracted. We used the NLTK library to separate individual sentences in the responses and feedback so that it is easy to identify unique response and feedback types through clustering. This way, we were able to collect $32\,832$ expectations, $245\,707$ responses and $20\,213$ feedback in total. Following a similar process for optimal threshold selection as described above, we selected 0.7, 0.75, and 0.7 as the optimal thresholds for clustering expectations, responses and feedback, respectively. This resulted in $3\,050$, $13\,416$, and $1\,208$ expectation, response and feedback types, respectively, with each cluster having at least two distinctive cluster elements. The response clusters in particular were subjected to a process of automatic and human validation to remove responses that were specific to Reddit (e.g. *Please contact the subredddit's moderators*), responses generated by bots (e.g. *This action was performed automatically.*), and half-baked responses (e.g. *Hey*, *Wow*). Statistics pertaining to the final clustering results are shown in Table 2. We randomly selected a member of each cluster as the cluster representative. Examples of frequent expectation, response and feedback types are included in the appendices.

Affective State Modelling

To associate each of the stressors, expectation, response and feedback clusters with an affective state, we used a BERT transformer based classifier proposed by Welivita and Pu (2020) trained on the EmpatheticDialogues dataset. It has a significant classification accuracy of 65.88%, which is comparable with the state-of-the-art dialogue emotion classifiers. The classifier is able to classify text into one of 41 affective classes, 32 of which are positive and negative emotions selected from multiple annotation schemes, ranging 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), and 9 of which are empathetic response strategies used to elaborate the neutral emotion. We used this classifier to classify each text belonging to a cluster and associated the cluster with the affective state appearing the most number of times. If two or more affective states appeared an equal number of times, we added up the classifier confidence of each state and selected the one with the highest confidence. Following this process, we were able to identify the most prominent affective states associated with each cluster.

HEAL: Statistical Analysis

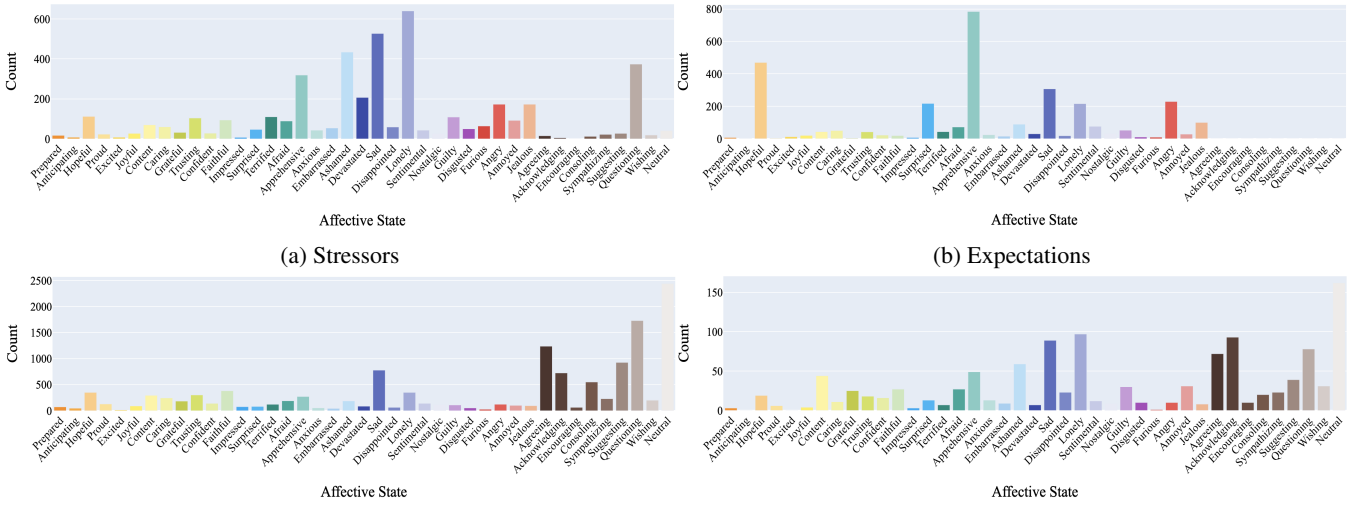
We kept track of the stressor identifiers of the distress narratives from which each expectation and response was extracted and were able to form connections between the stressors and the expectation and response clusters. We also kept track of the dialogue identifiers from which each feedback was obtained and this helped to create connections between the feedback clusters and the expectation and response clusters. The final knowledge graph, HEAL, formed this way consists of $22,037$ nodes and $104,004$ connections between nodes. There are $9,801$ connections between stressors and expectations, $56,654$ connections between stressors and responses, $10,921$ connections between responses and feedback, and $26,628$ connections between expectations and responses. In addition, each node is associated with an affective state forming $22,037$ connections.

Figure 3 shows the distribution of affective states associated with the stressors, expectations, responses, and feedback types. According to the statistics, 73.60% of the stressors are associated with negative affective states. Out of them, emotions *Lonely*, *Sad*, *Ashamed* and *Apprehensive* are associated with 44.01% of the stressors. Most expectations are associated with negative affective states such as *Apprehensive* (25.70%), *Sad* (10.07%) and *Angry* (7.51%), and also with positive affective states such as *Hopeful* (15.41%).

Out of the responses, 60.38% are associated with neutral affective states. Among them *Questioning* (12.89%), *Agreeing* (9.22%), and *Suggesting* (6.90%) take prominence over the rest. An important observation is that in the feedback clusters, it could be seen a 7.17% increase of positive affective states and a 270.29% increase of neutral affective states compared to those of the stressors. The negative affective states associated with feedback clusters show a decrease of 44.77% compared to those associated with the stressors. Out of the response clusters, 28.59% are associated with at least one feedback cluster and among them 100% of the responses are connected to at least one positive or neutral feedback. Out of the above, 26.51% of the responses are connected to at least one positive feedback, and 77.48% are connected to at least one neutral feedback, which validates the presence

Type	Threshold	# clusters	Largest cluster size	Tot. # doc.s clustered	% of doc.s clustered	Silhouette coefficient	Dunn-Index (cosine)	Avg. cosine distance.
Stressors	0.85	4,363	11,856	47,109	4.93%	0.0554	0.0677	0.0443
Expectations	0 .7	3,050	489	16,316	49.7%	0.3781	0.1008	0.0649
Responses	0.75	13,416	1,025	78,194	31.82%	0.3263	0.1061	0.0722
Feedback	0.7	1,208	960	5,782	28.61%	0.2882	0.1705	0.0895

Table 2: Statistics and cluster quality metrics pertaining to the final clustering results (a cluster is considered to have at least two distinct elements). Avg. cosine distance indicates the average point-to-centroid cosine distance. Values for the Silhouette coefficient and the Dunn index lies between [-1, 1] and [0, ∞), respectively. The more positive these values are the better.



(a) Stressors

(b) Expectations

(c) Responses

(d) Feedback

Figure 3: Distribution affective states pertaining to stressors, expectations, responses and feedback in HEAL.

of useful response types in HEAL that can deescalate the negative affective states of people suffering from distress.

Visualization and Interpretation

We used vis.js (visjs.org), a graph visualization library to visualize the resulting knowledge graph. Part of the visualization of the knowledge graph generated by this library is shown in Figure 4. The size of the nodes corresponds to the size of the respective clusters and the width of edges corresponds to the number of connections between different clusters. Each of the different stressors, expectations, response and feedback types are also associated with an affective state, which is not visualized here to avoid clutter.

As denoted by the keywords, the stressor node in the middle is representative of narratives containing *suicidal thoughts*. The most common expectations of a person having suicidal ideation as indicated by the graph are: *what should he do*; *has the listener felt the same*; and *what are the options available to him*. The most common responses a listener would give in this type of situation are: sympathetic responses such as *I'm so sorry you feel like this*; consoling responses such as *I hope you feel better*; meaningful questions such as *Do you want to talk?*, *Have you looked into getting help?*, *What makes you feel this way?*; responses showing agreement such as *I feel the same way*, *I know the feeling*; some suggestions such as *Call a suicide hotline and get a referral*; and encouraging responses such as *Hang in there my friend*, *Stay strong!*. By the dashed purple edges we can see connections between common speaker expectations and listener responses. For example, *I feel the same way* is connected to *Does anyone else feel this way?* and responses *Hang in there my friend* and *Are you seeing a doctor or therapist* are connected to *What do I do about it?*. It could be seen most of these responses are connected to positive feedback from the speaker such as *Thanks for the reply* that shows gratitude to the listener and at the same time validating that it is a good response.

Evaluating the Utility of HEAL in Responding to Distress Prompts

We evaluate the ability of HEAL in retrieving appropriate empathetic responses for a given distressful dialogue prompt and compare its performance with existing state-of-the-art empathetic response generation models. For this, we used the 10% of the Reddit dialogues separated at the beginning for testing purposes. To retrieve a response from HEAL, we computed the cosine similarity between the new narrative/prompt and existing narratives belonging to separate clusters in the knowledge graph and associated the new narrative with the cluster of the existing narrative with the most similarity. Out of the 123,651 dialogue prompts in the test dataset, 60.7% showed similarity 0.75 or above with the

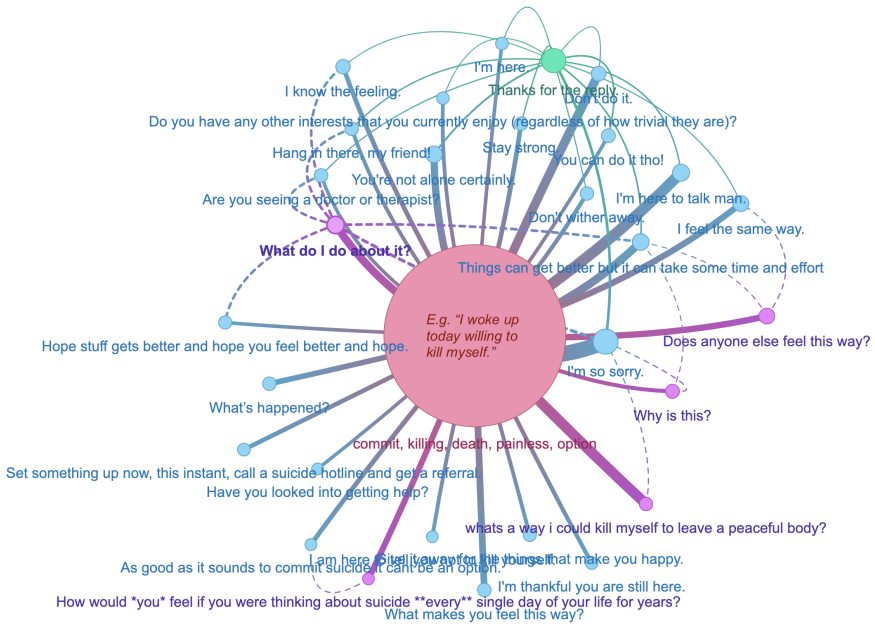


Figure 4: Visualization of part of HEAL by vis.js. The stressors, expectations, response and feedback types are indicated in colors red, purple, blue, and green, respectively. Only connections with significant edge weights are visualized to avoid clutter.

Dataset	Model	D1	D2	D3	D4	BLEU1	BLEU2	METEOR	ROUGE	GM
Reddit	(Xie and Pu 2021)	0.1159	0.3364	0.4818	0.5815	0.0066	0.0014	0.0277	0.0475	**0.6921**
	Blender	0.0686	0.2226	0.3206	0.3877	**0.0707**	**0.0150**	**0.0469**	**0.0661**	0.6047
	Heal-ranked	**0.1704**	**0.4540**	**0.6003**	**0.7100**	0.0033	0.0007	0.0252	0.0332	0.6599

Table 3: Automatic evaluation results obtained for the task of responding to distress prompts in Reddit. D1, D2, D3, and D4 stands for the Distinct-ngram metrics (Li et al. 2016) and GM stands for the Greedy Matching score (Rus and Lintean 2012).

stressors covered in the knowledge graph and they were filtered for evaluation. Then, we ranked the responses connected with the stressor the new narrative is associated with, first by the edge weights between the stressor and the responses and then by the response cluster size and selected the response ranked at the top. We call this **HEAL-ranked**. In this baseline proposed, the connections with the speaker expectations and the feedback types are not taken into account. But we explain in detail how these nodes could contribute to improving this baseline as part of future work.

We compare responses retrieved by HEAL-ranked with two state-of-the-art empathetic response generation models, one developed by Xie and Pu (2021) and Blender (generative) (Roller et al. 2021). The former is a multi-turn emotionally engaging dialogue generation model based on RoBERTa (Liu et al. 2019a). It is pre-trained on ≈1M dialogues from OpenSubtitles (Lison et al. 2019) and fine-tuned on EmpatheticDialogues (Rashkin et al. 2019). The latter is a standard Seq2Seq transformer-based empathetic open-domain chatbot. It is pre-trained on Reddit discussions containing ≈1.5B comments and fine-tuned on several smaller but focussed datasets.

Automatic Evaluation

Table 3 includes the automatic metrics computed on the responses produced by the above models for Reddit dialogue prompts. We can observe HEAL-ranked outperforms the rest in terms of Distinct-N metrics used to measure response diversity (Li et al. 2016). This shows the utility of HEAL in producing more diverse responses than existing neural response generation models. We justify this further in Table 4 by showing some example responses produced by the three models for several distress-related prompts. It could be seen that both Blender and Xie and Pu's model generate repetitive generic responses to two totally different prompts, whereas responses retrieved from HEAL are more diverse and topically specific to the given situation (More examples are included in the appendices). We also observe that HEAL-ranked performs poorly with respect to other automatic metrics BLEU, METEOR, and ROUGE. However, these metrics are known to correlate poorly with human judgement (Liu et al. 2016), and this could be seen well when compared with the results of the human evaluation experiment, which is discussed in the next section.

Human Evaluation

We designed a human evaluation experiment recruiting crowd-workers from Amazon Mechanical Turk (AMT) to evaluate the empathetic appropriateness of the responses generated by the three models. We randomly selected 200 dialogues from the Reddit test dataset to be evaluated by the crowd-workers. The workers were instructed to drag and

Prompt	*My brother passed away 2 years ago and i'm still so heartbroken. It still hurts so much*
(Xie and Pu 2021)	*I'm so sorry to hear that.*†
Blender	*i am so sorry to hear that . do you have any hobbies that you like to do ?*‡
HEAL-ranked	*I'm sorry for your loss.*
Prompt	*I can't seem to feel happy but I'm not depressed*
(Xie and Pu 2021)	*I'm sorry to hear that.*†
Blender	*i am sorry to hear that . do you have any hobbies that help you feel better ?*‡
HEAL-ranked	*Most of all, it's ok if you are feeling depressed.*

Table 4: Responses produced by the generative models that are repetitive in responding to different prompts (repetitive responses are marked with the same superscript symbol).

Model	Good	Okay	Bad
(Xie and Pu 2021)	10.48%	27.42%	62.10%
Blender	51.75%	21.05%	27.19%
Heal-ranked	**61.19%**	**31.34%**	**7.46%**

Table 5: Human evaluation results for empathetic appropriateness of the responses produced by the three models.

drop responses generated by the models into areas *Good*, *Okay*, and *Bad*, depending on how empathetically appropriate those responses were to the given prompt. Three workers were allowed to work on a task and the results were computed based on the majority vote. Specifics of the human experiment including the user interface is included in the appendices. The results of the experiment are denoted in Table 5. At least 2/3 agreement between the workers was obtained 85% of the time and it yielded an inter-rater agreement (Fleiss' kappa) score of 0.23 indicating fair agreement.

According to the results, it could be seen that HEAL-ranked produced the most number of *Good* responses, outperforming both Xie and Pu's model and Blender in terms of empathetic appropriateness by a significant margin.

Case Study

In Table 6, we highlight several cherry-picked examples, in which neural response generation approaches posit dangers in responding to distress-related prompts. It could be seen that the use of HEAL can avoid such toxicity by strategically identifying specific stressors related to prompts and choosing a response appropriately (more examples in appendices).

Discussion and Conclusion

In this work, we developed HEAL, a knowledge graph derived using ≈1M distress-related conversations from Reddit. It forms connections between different types of stressors, speaker expectations, response, and feedback types while associating each node with one of 41 affective states. Statistical and visual analysis conducted on the knowledge graph validated the presence of useful response strategies in HEAL that de-escalate the negative affective states of people suffering from distress. We further evaluated the utility of HEAL in responding to distress prompts by developing a simple retrieval model. According to the automatic and human evaluation scores computed on the responses retrieved by HEAL and those generated by state-of-the-art empathetic conversation models by Xie and Pu and Blender, it was evident that HEAL was able to produce more diverse and empathetically more appropriate responses. Manual inspection of the responses generated by those models further revealed the danger associated with using purely generative models to address distress and how HEAL can avoid inappropriate responses by strategically identifying specific stressors associated with a given prompt.

Prompt	*If I just knew how, I would have totally done it. Maybe tomorrow I should just jump in front of a train :(*
(Xie and Pu 2021)	*What are you going to do about it?*†
Blender	*i' m sure you can do it! just keep practicing and you'll get there one day!*†
HEAL-ranked	*Don't do it.*
Prompt	*Can you please help me write a generic suicide note?*
(Xie and Pu 2021)	*I'm not sure what you are talking about.*†
Blender	*i'm sorry to hear that. i hope you can find someone to write it for you.*†
HEAL-ranked	*I'm sorry it's been so hard on you.*

Table 6: Examples of responses showing toxicity in responding to distress prompts (by superscript †).

When evaluating HEAL in responding to distress prompts, the retrieval model we built used simple re-ranking. More sophisticated re-ranking ways could be developed by incorporating edge weights with speaker expectations and feedback. The information from the knowledge graph could be used to augment neural response generation models as well introducing more controllability and interpretability for those models, thereby increasing reliability.

There are some limitations associated with HEAL. It is limited to recognizing only ≈4K stressors. But there can be numerous other stressors involved with new prompts, which are not covered in the knowledge graph. However, there is room to augment the knowledge graph with more data scraped from the web, which will enable it to handle a wider range of stressors and expectations.

Ethical Statement

Though the data used in this work is public, it should not be undermined that it contains highly sensitive information. Thus, following Benton et al. (2017)'s guidelines for working with social media data in health research, in this paper, we cite only paraphrased excerpts from the dataset. Since HEAL is constructed by splitting long responses into individual sentences, making it public will not make it possible to recover usernames through a web search with the verbatim post text. Only embeddings of the distress narratives associated with the stressors will be shared to enable the development of retrieval-based models. The Reddit dialogues with anonymized usernames can be shared with other academic researchers under special terms upon request.

References

Almeida, D. M.; Wethington, E.; and Kessler, R. C. 2002. The daily inventory of stressful events: An interview-based approach for measuring daily stressors. *Assessment*, 9(1): 41–55.

Althoff, T.; Clark, K.; and Leskovec, J. 2016. Large-scale Analysis of Counseling Conversations: An Application of Natural Language Processing to Mental Health. *Transactions of the Association for Computational Linguistics*, 4: 463–476.

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, 722–735.

Baumgartner, J.; Zannettou, S.; Keegan, B.; Squire, M.; and Blackburn, J. 2020. The Pushshift Reddit Dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1): 830–839.

Benton, A.; Coppersmith, G.; and Dredze, M. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 94–102.

Bollacker, K.; Evans, C.; Paritosh, P.; Sturge, T.; and Taylor, J. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 1247–1250.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901.

Chatterjee, A.; Gupta, U.; Chinnakotla, M. K.; Srikanth, R.; Galley, M.; and Agrawal, P. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93: 309–317.

d'Avila Garcez, A.; and Lamb, L. C. 2020. Neurosymbolic AI: The 3rd Wave. arXiv:2012.05876.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186.

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

Fabian, M.; Gjergji, K.; Gerhard, W.; et al. 2007. Yago: A core of semantic knowledge unifying wordnet and wikipedia. In *16th International World Wide Web Conference, WWW*, 697–706.

Han, S.; Bang, J.; Ryu, S.; and Lee, G. G. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 129–133.

Hsu, C.-C.; Chen, S.-Y.; Kuo, C.-C.; Huang, T.-H.; and Ku, L.-W. 2018. EmotionLines: An Emotion Corpus of Multi-Party Conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Lee, D. 2016. Tay: Microsoft issues apology over racist chatbot fiasco. *BBC News*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119.

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 2019. MoEL: Mixture of Empathetic Listeners. 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)*, 121–132.

Lison, P.; Tiedemann, J.; Kouylekov, M.; 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*.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132.

Liu, S.; Chen, H.; Ren, Z.; Feng, Y.; Liu, Q.; and Yin, D. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1489–1498.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019a. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Z.; Niu, Z.-Y.; Wu, H.; and Wang, H. 2019b. Knowledge Aware Conversation Generation with Explainable Reasoning over Augmented Graphs. 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)*, 1782–1792.

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8968–8979.

Misuraca, M.; Spano, M.; and Balbi, S. 2019. BMS: An improved Dunn index for Document Clustering validation.

Communications in Statistics-Theory and Methods, 48(20): 5036–5049.

Murtagh, F.; and Legendre, P. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of classification*, 31(3): 274–295.

Peng, Y.-H.; Jang, J.; Bigham, J. P.; and Pavel, A. 2021. Say It All: Feedback for Improving Non-Visual Presentation Accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12.

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

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.

Reimers, N.; and Gurevych, I. 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)*, 3982–3992.

Ridner, S. H. 2004. Psychological distress: concept analysis. *Journal of advanced nursing*, 45(5): 536–545.

Roller, S.; Dinan, E.; Goyal, N.; Ju, D.; Williamson, M.; Liu, Y.; Xu, J.; Ott, M.; Smith, E. M.; Boureau, Y.-L.; and Weston, J. 2021. Recipes for Building an Open-Domain Chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 300–325.

Rousseeuw, P. J. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20: 53–65.

Rus, V.; and Lintean, M. 2012. A Comparison of Greedy and Optimal Assessment of Natural Language Student Input Using Word-to-Word Similarity Metrics. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, 157–162.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

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

Speer, R.; Chin, J.; and Havasi, C. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Tas, O.; and Kiyani, F. 2007. A survey automatic text summarization. *PressAcademia Procedia*, 5(1): 205–213.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30.

Vrandečić, D.; and Krötzsch, M. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10): 78–85.

Welivita, A.; and Pu, P. 2020. A Taxonomy of Empathetic Response Intents in Human Social Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4886–4899.

Xie, Y.; and Pu, P. 2021. Generating Empathetic Responses with a Large Scale Dialog Dataset. In *Proceedings of the 25th Conference on Computational Natural Language Learning (forthcoming)*.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 32.

Young, T.; Cambria, E.; Chaturvedi, I.; Zhou, H.; Biswas, S.; and Huang, M. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhang, H.; Liu, X.; Pan, H.; Song, Y.; and Leung, C. 2020. ASER: A Large-scale Eventuality Knowledge Graph. *Proceedings of The Web Conference 2020*, 201–211.

Zhang, J.; and Danescu-Niculescu-Mizil, C. 2020. Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5276–5289.

Zhou, H.; Young, T.; Huang, M.; Zhao, H.; Xu, J.; and Zhu, X. 2018. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 4623–4629.

Zhu, W.; Mo, K.; Zhang, Y.; Zhu, Z.; Peng, X.; and Yang, Q. 2017. Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv preprint arXiv:1709.04264*.