Thèse n° 9810

# Empathetic Conversational Agents for Distress Support

Présentée le 29 septembre 2023

Faculté informatique et communications
Groupe SCI IC PFP
Programme doctoral en informatique et communications

pour l'obtention du grade de Docteur ès Sciences

par

**Kalpani Anuradha WELIVITA**

Acceptée sur proposition du jury

Dr R. Boulic, président du jury
Dr. P. Pu Faltings, directrice de thèse
Dr M. Sachan, rapporteur
Dr Q. V. Liao, rapporteuse
Dr M. Rajman, rapporteur

École
polytechnique
fédérale
de Lausanne

2023

*The greatest glory in living lies not in never falling, but in rising every time we fall.*
— Confucius

To all who have supported me.

# Acknowledgements

I would like to thank everyone who helped me throughout my P.hD. journey, which had not been an easy one. First of all, I would like to pay my sincere gratitude to my supervisor Dr. Pearl Pu for the continuous guidance and support that she had been offering me throughout these years. She has been an amazing, hands-on supervisor and I consider myself to be very fortunate to be under her guidance. She has been very enthusiastic about my research and has been continuously giving me feedback and suggestions to improve. Without her, this thesis would not have been possible and I sincerely thank her for her enormous effort in moulding me into the researcher that I am today. Next, I would like to thank my jury members, Dr. Ronan Boulic, Dr. Martin Rajman, Dr. Mirinmaya Sachan, and Dr. Vera Liao for dedicating their valuable time to attend my P.hD. defense. I would also like to take this opportunity to thank Prof. Marcel Salathé who gave me the opportunity to work in his lab during the first semester of my P.hD. study and for his guidance throughout that time. I would also like to thank my colleagues in the HCI group, Yubo Xie, Ekaterina Svikhnushina, Kavous Salehzadeh, Igor Kulev, and Chenkai Wang for their friendship, support and the valuable time spent with me, which left me with wonderful memories to carry forward. I would also thank our lab's administrative assistants Karin Getaz, and Defferrard Patricia for their kind support in a number of administrative tasks and also for their wonderful friendship. My thanks also extend to all the Bachelors' and the Masters' students who worked with me on various projects, especially Chun-Hung Yeh, Mustapha Zied, Wanhao Zhou, Wang Zheng, Sena Cetin Necla, Joshua Swanson, Changyang He, and Furrer Stanislas Alexandre Titus.

In addition, I would like to thank all the professors and lecturers at the Department of Computer Science, University of Moratuwa, Sri Lanka where I completed my bachelor's degree, especially Prof. Chandana Gamage, Prof. Indika Perera, and Prof. Dulani Meedeniya who supported and advised me immensely when I laid my first steps into research and identified my strengths and recommended me to the P.hD. program at EPFL. I would also like to thank all the teachers at Southlands College, Galle, Sri Lanka, who have been instrumental in shaping my academic and personal growth.

I would also like to extend my gratitude to my parents who provided me with all the necessary resources to pursue my dreams and sacrificed so much to ensure that I had the best possible education and opportunities. I also thank my husband's family, my mother-in-law, father-in-law, brother Udaranga, and sister Heshani for listening to my concerns and offering me continuous encouragement. I would like to thank my extended family members, all my aunts, uncles, and cousins for their support and beloved presence in my life. I also thank all my friends in Switzerland, especially, Dewmini, Yasara, Wenuka, Buddhima, and Pasindu for being my family away from home. Last but not least, I would like to express my love and sincere gratitude towards my husband, Ruchiranga Wickramasinghe, who stood behind me in all my ups and downs, and has been providing continuous support throughout my P.hD. journey. His emotional support and belief in my abilities have been essential in helping me overcome challenges and persevere towards achieving my P.hD. and I am grateful for his presence in my life.

*Lausanne, July 21, 2023* - Anuradha

# Abstract

Due to the increasing demands of today's fast-paced world, mental health concerns are on the rise, which necessitates innovative approaches to provide support to those in need. Open-domain conversational agents known as chatbots, offer a unique opportunity to provide empathetic support to individuals struggling with psychological distress. By combining the advancements in natural language processing, such as the advent of large language models and machine learning techniques that can understand human emotions, empathetic chatbots can establish meaningful connections, provide support in distress, and promote mental well-being. This thesis aims to develop empathetic conversational agents that are capable of providing emotional support to people undergoing distress. They are designed in a way such that they offer a reliable space for individuals to express their feelings and motivate them to navigate their emotional challenges and cope with them, ultimately leading to enhanced mental well-being. However, developing such chatbots poses significant challenges such as understanding subtle variations in human emotion, overcoming limitations in training data, ensuring interpretability and reliability of responses, and adhering to established psychological norms and professional tone when responding to distressing situations. In this thesis, we develop resources and methods to address the above challenges and attempt to pave the way for a more compassionate and accessible approach to emotional well-being. To achieve this goal, first, we look at subtle emotional variations present in human conversations and communication strategies humans use to convey empathy, which form the foundation for developing more controllable and interpretable chatbot models that can respond to a wide range of emotions. Then we narrow our attention toward the more challenging task of responding empathetically to extremely negative emotions in psychologically distressing situations. Analyzing dialogues from online peer support forums, we build a knowledge graph that identifies a multitude of distress-related topics and emotionally relieving responses associated with them, facilitating the development of more reliable and topically appropriate chatbot models for distress support. Moving a step further, we analyze the differences in language used by laypersons and professionals when responding to distress and guided by these observations, develop methods to enhance chatbots' professional tone and adherence to therapeutic norms. Overall, this thesis contributes to the advancement of empathetic chatbots that can provide safe, dependable, and professional assistance to users in need.

**Keywords:** Conversational agents, Emotion recognition, Empathy, Empathetic response generation, Psychological distress, Crowdsourcing, Knowledge graphs, Reliability, Dependability.

# Résumé

En raison des demandes croissantes du monde moderne et rapide d'aujourd'hui, les problèmes de santé mentale sont en hausse, ce qui nécessite des approches innovantes pour apporter du soutien à ceux qui en ont besoin. Les agents conversationnels à domaine ouvert connus sous le nom de chatbots offrent une opportunité unique de fournir un soutien empathique aux personnes qui luttent contre la détresse psychologique. En combinant les avancées dans le traitement du langage naturel, telles que l'avènement de grands modèles de langage et de techniques d'apprentissage automatique capables de comprendre les émotions humaines, les chatbots empathiques peuvent établir des connexions significatives, fournir du soutien en cas de détresse et favoriser le bien-être mental. Cette thèse vise à développer des agents conversationnels empathiques capables de fournir un soutien émotionnel aux personnes en détresse. Ils sont conçus de manière à offrir un espace fiable aux individus pour exprimer leurs sentiments et les motiver à surmonter leurs défis émotionnels, favorisant ainsi un bien-être mental accru. Cependant, le développement de tels chatbots présente des défis importants tels que la compréhension des variations subtiles des émotions humaines, le dépassement des limitations des données d'entraînement, la garantie de l'interprétabilité et de la fiabilité des réponses, ainsi que le respect des normes psychologiques établies et du ton professionnel lors de la réponse aux situations de détresse. Dans cette thèse, nous développons des ressources et des méthodes pour relever les défis mentionnés ci-dessus et nous nous efforçons de jeter les bases d'une approche plus compatissante et accessible au bien-être émotionnel. Pour atteindre cet objectif, nous examinons d'abord les variations émotionnelles subtiles présentes dans les conversations humaines et les stratégies de communication que les humains utilisent pour exprimer de l'empathie, ce qui constitue la base du développement de modèles de chatbot plus contrôlables et interprétables capables de répondre à une large gamme d'émotions. Ensuite, nous concentrons notre attention sur la tâche plus difficile de répondre de manière empathique à des émotions extrêmement négatives dans des situations de détresse psychologique. En analysant des dialogues de forums d'entraide en ligne, nous construisons un graphe de connaissances qui identifie de multiples sujets liés à la détresse et des réponses émotionnellement apaisantes associées à ceux-ci, facilitant ainsi le développement de modèles de chatbot plus fiables et appropriés pour le soutien en cas de détresse. En allant encore plus loin, nous analysons les différences de langage utilisées par les profanes et les professionnels lorsqu'ils répondent à la détresse, et en nous basant sur ces observations, nous développons des méthodes pour améliorer le ton professionnel des chatbots et leur respect des normes thérapeutiques. Dans l'ensemble, cette thèse contribue à l'avancement des chatbots empathiques capables de fournir une assistance sûre, fiable et professionnelle aux utilisateurs dans le besoin.

**Mots-clés :** Agents conversationnels, Reconnaissance des émotions, Empathie, Génération de réponses empathiques, Détresse psychologique, Crowdsourcing, Graphes de connaissances, Fiabilité, Fiabilité.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

In today's fast-paced and interconnected world, psychological distress has emerged as a pressing concern, impacting the well-being and mental health of people worldwide. According to the statistics of the World Health Organization (WHO), nearly a billion people worldwide are suffering from psychological distress. The COVID-19 pandemic aggravated this condition triggering 25% increase in the prevalence of anxiety and depression worldwide (WHO, 2022b). Today, depression has become one of the leading causes of disability. Suicide has become the fourth leading cause of death among youngsters (WHO, 2022a). Despite the increase in mental health issues, only a small fraction of people in need have access to effective and affordable mental health care. The limited availability of mental health professionals, the stigma associated with seeking help, and the overwhelming demand for support create barriers that prevent many individuals from accessing the assistance they need. As a result, there is an urgent need for innovative solutions that can bridge this support gap and provide empathetic assistance to those suffering from distress-related issues.

In recent years, open-domain conversational agents or chatbots have emerged as a popular tool that is used across a variety of domains (Zhou et al., 2020a; Adiwardana et al., 2020; Roller et al., 2021; Zhang et al., 2020b). Early chatbots such as ELIZA (Weizenbaum, 1966) and PARRY (Colby et al., 1971), which relied on manually defined patterns laid the foundation for the development of conversational agents. In the early 2000s, the advent of deep learning and neural networks enabled the development of more sophisticated conversational agents capable of understanding and generating natural language (Vinyals and Le, 2015a; Shang et al., 2015; Serban et al., 2016; Xing et al., 2018). In recent years, the introduction of large language models, such as GPT-4 (OpenAI, 2023) and LLaMA (Touvron et al., 2023) has revolutionized the capabilities of conversational agents, allowing for more context-aware and human-like interactions. Notably, there has been a growing focus on incorporating empathy and emotional intelligence into these conversational agents (Asghar et al., 2018; Zhong et al., 2019; Rashkin et al., 2019). This emerging development allows chatbots to be effective means to provide emotional support to those struggling with psychological distress and intervene during crisis situations (Vaidyam et al., 2019). Their availability 24/7 and the ability to maintain the user's anonymity create an efficient and safe space for individuals to seek support without fear of judgment or identification.

However, large language models and other neural models trained in an end-to-end fashion, while exhibiting remarkable capabilities in various natural language processing tasks, currently fall short in providing effective empathetic distress support. Despite their fluency and the ability to generate contextually relevant responses, these models often lack understanding of subtle variations in human emotion and the complex dynamics of human empathy. Empathy, being a deeply nuanced and multifaceted human experience, requires not only linguistic proficiency but also a deep understanding of human psychology, emotions, and social context. The current state of large language models, which primarily relies on statistical patterns and surface-level correlations, often results in generic or even inappropriate responses to individuals experiencing distress. The black-box nature of these models adds to the lack of controllability and interpretability of the responses generated, making them quite unreliable. Also, they raise concerns about how professionally sounding those responses are. To address these challenges and limitations that exist with the current state-of-the-art, in this thesis, we aim to build empathetic conversational agents that can understand a wide range of nuanced human emotions and provide emotional support to users undergoing psychological distress. We particularly aim at developing resources and methods that can enable these chatbot models to be more interpretable, reliable, and sound more professional.

Empathy is the ability to understand or feel what another person is experiencing by placing oneself in his or her situation. It refers to the multidimensional capacity to recognize, feel, and/or react compassionately to others' emotional states (Ekman, 1992; Bellet and Maloney, 1991). It is an important component of interpersonal relationships that contributes to healthy social and emotional functioning (Eisenberg and Eggum, 2009; Decety, 2010). Empathetic conversation modeling has been one of the longest-running goals in artificial intelligence (Zhou et al., 2020a). Modeling empathy in chatbots enables chatbots to build trust and rapport with the users (Liu-Thompkins et al., 2022). Empathetic responding is also a key component in making chatbots more human-like, which can help to increase the adoption of this technology (Goetz et al., 2003; Stroessner and Benitez, 2019). This task is challenging since it requires a deep understanding of human emotions, social dynamics, and empathetic communication patterns. It is often difficult to accurately detect and interpret the nuances of human emotions. The communication patterns also fluctuate depending on these emotions. Even subtle emotional variations such as being *Sad* vs being *Disappointed* can make the response vary. However, most existing work in this field is limited to identifying and responding to coarse-grained emotions. Coarse-grained emotions refer to broad categories of emotions, such as happiness, sadness, anger, and fear, which are mostly adapted from traditional emotion models such as the Ekman's basic emotions (Ekman, 1992) and the Plutchik's emotion wheel (Plutchik, 1984). While detecting these emotions is an important step in empathetic conversation modeling, it does not capture the full complexity of human emotions. Thus, it is important to pay attention to complex human emotions at a finer granularity and specific empathetic response patterns associated with them when developing empathetic chatbots.

It is also found that chatbots find it more challenging to respond to negative emotions than positive ones. In a study conducted by Bickmore (2003), participants interacting with a social chatbot rated the chatbot as more likable and trustworthy when it was responding to positive emotions than negative ones. The authors suggest that chatbots may struggle to respond to negative emotions because they lack the ability to show empathy and provide the necessary emotional support. Hence, the development of empathetic chatbots for applications such as distress support requires more sophisticated natural language processing techniques and a deeper understanding of the intricacies leading to distress.

According to the emotion theorist, Paul Ekman (1992), empathy is operationalized in terms of affective, cognitive, and compassionate correlates. *Affective empathy*, also known as "emotional contagion" defines the subjective mirroring of others' feeling states. This involves actually feeling the emotions of another person as if they were your own. Though this type of empathy can help us build emotional connections with others, it can sometimes lead to emotional distress as a result of sharing negative emotions. *Cognitive empathy* describes the understanding and accurate identification of others' feelings and states, without necessarily sharing their emotions. Cognitive empathy is seen as a foundation upon which affective and compassionate empathy is built. It is also observed to be more adaptive than affective empathy in minimizing the distress associated with negative emotions. *Compassionate empathy* is feeling concerned for the well-being of another person and taking action to reduce their suffering. It goes beyond mere understanding and makes it actionable. It is linked to positive social outcomes such as charitable behavior. It is identified as the most socially desirable type of empathy out of the three types of empathetic responses. However, most works involving the development of empathetic chatbots pay attention only to the affective component of empathy, which is most often implemented using emotion mimicry. They ignore the cognitive and compassionate aspects, which are deemed to be more desirable specially in a distress support setting.

Figure 1.1 summarizes the existing work on conversational agents. The bottom-left quadrant shows conversational agents capable of carrying out general chitchat, which are trained in an end-to-end fashion. The end-to-end transformer-based neural conversational models such as those introduced by Vinyals et al. (2015a) and Shang et al. (2015), and the recently proposed large-language models such as OPT (Zhang et al., 2022), GPT-4 (OpenAI, 2023), and LLaMA (Touvron et al., 2023) fall under this quadrant. However, empathetic responding is different from general chitchat. As elaborated above, it requires explicit understanding of human emotions and how they are addressed. The most common approach to making chatbots respond in an empathetic manner is fine-tuning dialogue models on datasets that contain a variety of emotions and empathetic responses. Blender (Roller et al., 2021) and XiaoIce (Zhou et al., 2020a) are some chatbots that are developed in this manner. However, due to the black-box nature of these models and the fact that they are trained in an end-to-end fashion, they often lack interpretability, which makes it difficult to rely on them. As a remedy, some works focus on making these models more controllable and interpretable by incorporating simple rules or heuristics into the training process that can help to steer the response generation. These works are denoted under the top two quadrants. In the top-left are the works that incorporate rules and heuristics when responding to day-to-day chitchat. These often include simple dialogue acts such as *making a statement*, *question*, and *answer*, which do not reflect empathy. In the top-right are the works that incorporate rules and heuristics to explicitly induce empathetic responding capabilities to dialogue models. The works included here incorporate mechanisms such as affective word embeddings and affect-based objective functions (Asghar et al., 2018; Zhong et al., 2019; Xie et al., 2020), and affect matching by conditioning the response on a predefined emotion label (Ghosh et al., 2017; Zhou et al., 2018a; Zhou and Wang, 2018; Song et al., 2019; Colombo et al., 2019) to control the response generation. However, human empathy cannot be operationalized in terms of simple rules or heuristics and such rules do not exist in the psychological literature. Hence, operationalizing empathy in terms of its cognitive, affective, and compassionate counterparts by identifying specific empathetic response strategies is a gap that remains to be filled.

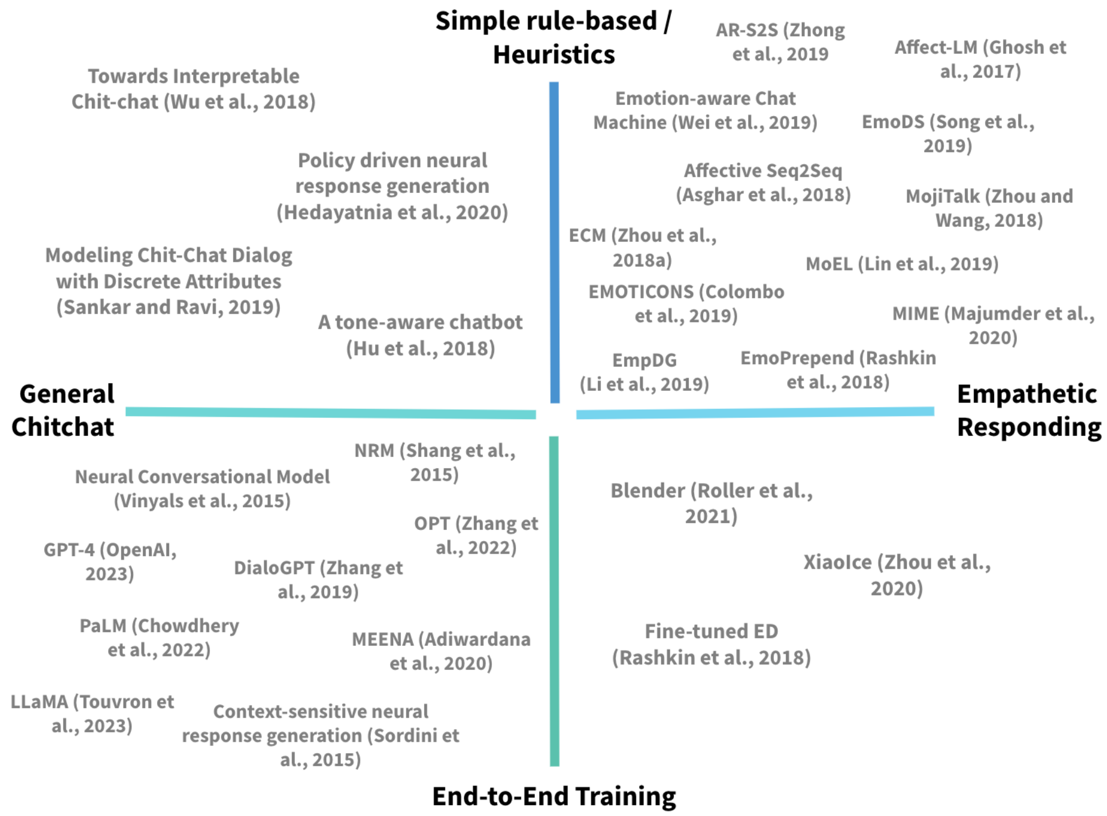



Figure 1.1: Analysis of existing work on conversational agents. The two quadrants on the left denote general chitchat-oriented dialogue agents that do not give a particular emphasis on empathetic responding. On the right are existing conversational agents that can respond empathetically to dialogue prompts. They either rely on end-to-end training on dialogue datasets containing empathetic responses (bottom-right quadrant) or use simple rules or heuristics (top-right quadrant) to explicitly induce empathetic responding capabilities to the dialogue models.

While significant progress has been made, there is still limited understanding of and mitigation against the risk of insensitivity in chatbots. Chatbots entail the risk of providing inappropriate or insensitive responses to users' emotional concerns. They can generate generic or repetitive responses that may be perceived as uncaring and unsupportive. This is especially true with neural network-based chatbot models that are trained in an end-to-end fashion. The black-box nature of these models makes the response generation process unpredictable and uninterpretable. This in turn makes them difficult to trust and rely on. Hence, it is vital to investigate more controllable and interpretable methods that can steer the chatbots' response generation process along predictable routes avoiding any inappropriate or misaligned responses. The requirement for guaranteeing safe, reliable, and appropriate responses is higher in applications of distress support. In a study conducted by Liu and Sundar (2018), the users' expectations of empathy from a chatbot giving health-related advice were observed to be higher when they were expressing negative emotions. While the chatbot's use of empathy was positively associated with users' perceived satisfaction and trust in the chatbot, inappropriate expressions of empathy were negatively associated with users' perceived competence of the chatbot. Thus, extreme caution should be taken to avoid the delivery of inappropriate responses when responding particularly to distress. It is not limited to avoiding profane or judgemental responses. As pointed out by R. Tatman (2022) a response such as *"You're not alone"* may be comforting to someone with depression, however, can bring detrimental effects to someone suffering from paranoia. So, it is vital to ensure the appropriateness as well as the topical specificity of the generated text when responding to distressful situations.

Moving further, empathy displayed by a professional or therapist trained in the practice of counseling or mental health support is different from that expressed by a layperson. Empathy displayed in a mental health support setting involves being present and attuned to the client's emotional state, and responding with compassion and support. This type of empathy is essential to build trust and rapport with the client and is one of several factors crucial to a strong therapeutic alliance (Jeremy Sutton, 2022). Therapists are trained to use a specific language that is designed to be empathetic, supportive, and non-judgmental in order to help clients feel heard, understood, and validated. They may also use specialized therapeutic techniques that involve specific language and terminology (Raskin and Rogers, 2005; Hayes and Pierson, 2005; Gilbert, 2009; Hettema et al., 2005). In contrast, laypersons may not have the same level of training or experience in responding to distress, which may be reflected in the responses generated by chatbots trained on such data. Thus, it is important to investigate these different approaches in eliciting empathy in chatbot models used for applications of distress support and enhance chatbots' responses to reflect the language used by professionals and make them conform to therapeutic norms that elicit empathy in the therapeutic alliance. This could help to make the chatbots used for distress support more dependable and supportive for the users.

In this thesis, we attempt to address the above concerns that pose significant challenges when developing empathetic conversational agents capable of providing emotional support to people in distress. Specifically, we aim to express empathetic interaction that is beyond emotion mimicry, with a particular emphasis on enhancing their reliability and professional tone. In this regard, we first focus on the chatbot's ability to respond to a wide range of human emotions encountered in everyday situations, both positive and negative. We then narrow our focus to the challenging scenario of responding to extremely negative emotions, particularly when the user is experiencing psychological distress. In addition to ensuring the reliability and fail-safe operation of chatbot responses, we investigate how they can be made more professional and adhere to established norms in psychology.

## 1.1 Research Motivations

Our research towards developing empathetic conversational agents for distress support was primarily driven by several key motivations that can be listed as follows:

**Studies that attempt to gain control over chatbots' response generation process using dialogue acts, emotion labels, or rules are not suitable for generating empathetic responses.** Studies have attempted to make response generation more controllable by conditioning on dialogue acts or intentions (Xu et al., 2018; Zhao et al., 2017; Serban et al., 2017b). But these approaches are limited to general chitchat and cannot be applied to empathetic response generation. Studies that attempt to gain control over emotionally appropriate responses often rely on manually specified sentiment or emotion labels or rules such as emotion mimicry (Zhou et al., 2018a; Zhou and Wang, 2018; Hu et al., 2018; Asghar et al., 2018; Song et al., 2019). However, empathy is a much more complex psychological construct than mere mimicry of emotion and to the best of our knowledge, there are no definitive rules in the psychological literature that describe how empathy is elicited. Thus, the lack of understanding of the means through which humans elicit empathy in response to different emotional situations has become a major obstacle in attempting to gain control over chatbots' empathetic response generation process.

**Dialogue datasets that accurately capture the empathetic communication strategies used to address a variety of emotions are limited.** Another obstacle to progress in this field is the limited availability of dialogue datasets that accurately capture the complex communication strategies that humans use when responding empathetically to a wide variety of emotions. Existing datasets are limited in scale, capturing only coarse-grained emotions and lacking any empathetic response intents (Busso et al., 2008; Poria et al., 2019; Li et al., 2017a; Hsu et al., 2018; Chatterjee et al., 2019a). Annotating datasets with fine-grained emotions and intents is a daunting task because of its inherent complexity. With a vast array of potential labels, human annotation becomes more time-consuming and prone to error, which makes curating such datasets technically challenging.

**The lack of linguistic resources capturing knowledge related to psychological distress presents significant challenges to generating distress-related responses that are both topically specific and reliable.** Topical specificity and reliability are two important factors when generating responses to address distress-related situations. To achieve this objective, there have been research efforts directed at incorporating knowledge and commonsense reasoning over graph-based representations when developing conversational agents (Zhou et al., 2018b; Young et al., 2018; Sap et al., 2019; Speer et al., 2017). The use of knowledge graphs is expected to provide more control over the generated response than training over unstructured text and generate topically relevant responses. However, the scarcity of linguistic resources capturing knowledge related to psychological distress poses a significant challenge to the application of these methods. The lack of high-quality large-scale conversational data between people who are psychologically distressed and psychologists, counselors, or social workers trained in providing them with the required support adds to this difficulty.

**The use of dialogues from online peer support platforms in place of mental health-related dialogues between clients and professionals has become an obstacle in developing chatbots with a professional tone and adherence to psychological norms.** Due to the scarcity of mental health-related datasets, existing work utilizes dialogue data scraped from online peer support platforms such as Reddit. However, the extent to which they align with conversations with professionals and how it affects the response generation process is largely understudied. While peer supporters on these platforms are motivated and well-intentioned to help others seeking support, they are untrained and typically unaware of best practices in therapy (Sharma et al., 2020b). Thus, the possibility for this data to contain inappropriate responses such as judgments, confrontations, and advices without permission is high. Chatbots trained on such data can also hold a tone that is unprofessional. This area of research can largely benefit from therapeutic annotations of the peers' and counselors' utterances. However, prior work has found that it is very challenging to annotate therapeutic constructs (Pérez-Rosas et al., 2016; Klonek et al., 2015). It requires trained professionals, clear guidelines, and quality checks to obtain high-quality annotations. This makes it difficult to gain an understanding of the differences between laypersons' and professionals' responses to distress. This has in turn become an obstacle in the path to enhancing chatbots' professional tone and adherence to psychological norms, which has slowed the progress of research in this area.

In summary, the lack of understanding of how human empathy is elicited in day-to-day conversations as well as when supporting someone experiencing psychological distress, the scarcity of linguistic resources such as large-scale empathetic dialogues and distress-support datasets and knowledge graphs built on top of them, and the high costs of human labor for understanding differences between professional and online available dialogues make it challenging to develop methods to steer the chatbots' response generation process to provide safe, reliable, and professional empathetic support to the users. In this thesis, we direct our research efforts to overcome these obstacles and build empathetic chatbots that can respond to a variety of emotions and provide emotional support in psychologically distressing situations in a more reliable, trustworthy, and dependable manner.

## 1.2 Research Questions

The primary objective of this thesis is to study the generation of empathetic responses to dialogue prompts encompassing a wide spectrum of emotions, starting from day-to-day positive and negative emotions to exceedingly negative and distressing emotions. We give a particular emphasis on making these responses more reliable, interpretable, and sound more professional. There are several research questions that we try to address in this regard. Figure 1.2 shows an overview of the research questions we attempt to answer and how they relate to the key components in the empathetic chatbot development process. The figure also denotes the corresponding chapters in the thesis that investigate these research problems. We elaborate on them further as follows.

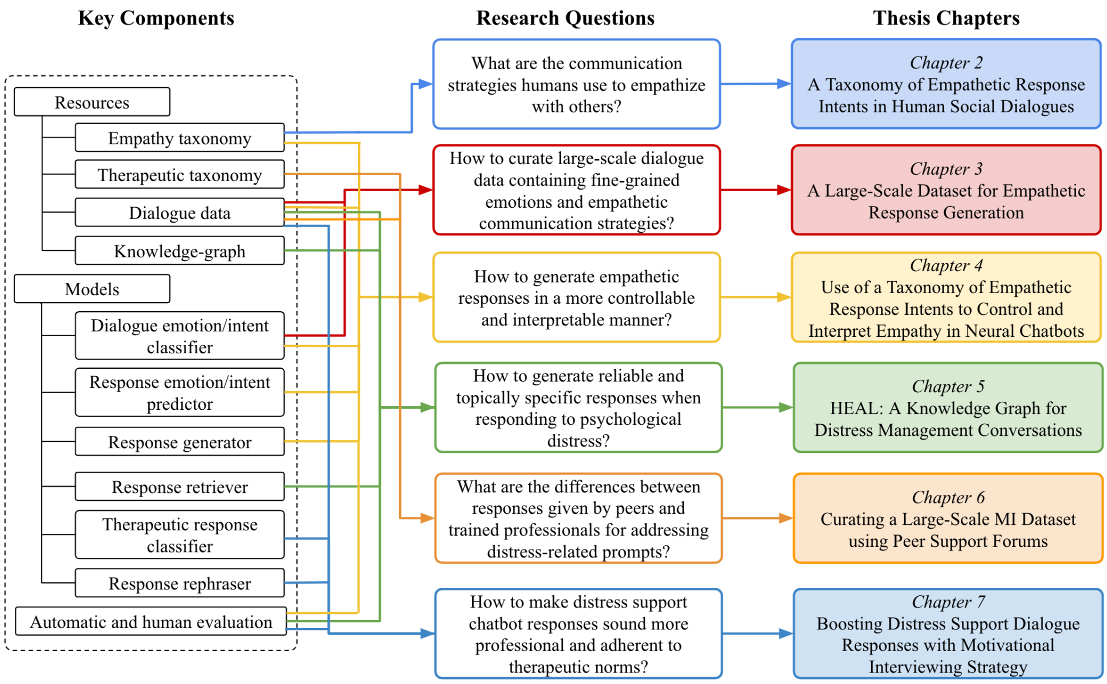



Figure 1.2: The main research questions addressed in this thesis and how they relate to the key components in the empathetic chatbot development process. The diagram also denotes the corresponding chapters of the thesis that investigate these research problems.

1. **What are the communication strategies humans use to empathize with others? (Chapter 2)**

   Humans are born with core *affect* neural circuitry, and thus implicitly develop the ability to empathize with others. Chatbots can acquire these empathetic capabilities through data-driven approaches that involve training neural models on human conversations to learn patterns of empathetic responses. However, due to the uncontrollability and the unpredictable nature of fully data-driven methods, these chatbots still encompass the risk of generating generic, repetitive, and inappropriate responses. Explicitly identifying empathetic communication strategies can help to guide the conversation flow in predictable directions and help chatbots more appropriately respond to users' emotional needs. However, there are no definitive rules or established strategies in the psychological literature that helps us in this process. So, it is vital to understand and establish specific communication means that humans use to empathize with others when they encounter different types of positive and negative emotions.

We attempt to answer the following research questions in this regard: What defines empathy? What common strategies do humans use to empathize with others? Do these always involve emotion? How can we establish a taxonomy of empathetic response intents used in human social conversations to be universally used by conversational agents? What empathetic response patterns can we explicitly identify in human-human dialogues when responding to different types of positive and negative emotions? How do they vary as the conversation progress?

2. **How to curate large-scale dialogue data containing fine-grained emotions and empathetic communication strategies? (Chapter 3)**

   To better understand human emotion and generate empathetically appropriate responses, it requires large-scale dialogue datasets covering different types of emotions and empathetic response strategies. But due to the cost of human labor, existing datasets are small in size and contain only coarse-grained emotion annotations (Busso et al., 2008; Poria et al., 2019; Li et al., 2017a; Hsu et al., 2018; Chatterjee et al., 2019a; Rashkin et al., 2019). The response strategies mostly remain unannotated in these datasets. Since collecting and manually annotating such gold standard data is expensive, there is rising interest in curating automatically annotated silver standard data (Filannino and Di Bari, 2015). Inspired by this movement, we investigate how such a large-scale silver standard dataset containing fine-grained emotions and empathetic response strategies can be curated.

   We break this problem down to the following research questions: What are the sources out there that are easily accessible to curate a large number of dialogues? What methods can we use to identify dialogues that are emotionally colored and empathetic without the use of human labor? How do dialogues curated and annotated using automatic methods (silver-standard dataset) compare to dialogues in gold-standard datasets? Do they preserve the same quality as those of the gold standard? How do empathetic conversational agents trained based on large-scale silver-standard data compare to those trained on small-scale gold-standard data?

3. **How to generate empathetic responses in a more controllable and interpretable manner? (Chapter 4)**

   With the introduction of neural network architectures such as the transformer (Vaswani et al., 2017a) and large pre-trained language models (Devlin et al., 2019; Zhuang et al., 2021; Brown et al., 2020b), developing open-domain conversational agents in a fully data-driven end-to-end fashion has become mainstream. But the black-box nature of these models offers the developer less controllability and the responses generated in this manner are less interpretable. This makes chatbots developed in such a manner less reliable and trustworthy. Thus, discovering means by which we can control and interpret the empathetic response generation process holds importance for the advancement of reliable and secure empathetic conversational agents.

We break this down into the following questions and attempt to solve each of them: What methods can accurately predict the emotion or intent of the response that should be generated? How do these methods compare to each other? Does incorporating the dialogue history help this prediction process? How do responses generated by end-to-end neural models compare to responses generated by conditioning on a predicted emotion or intent? Does it add more controllability and interpretability to the response generation process?

4. **How to generate reliable and topically specific responses when responding to psychological distress? (Chapter 5)**

   Due to the black-box nature and lack of controllability in end-to-end chatbot models, the use of these chatbots to address sensitive situations such as when the user is undergoing psychological distress is less reliable and not fail-safe (Garcez and Lamb, 2023). Also, since distress can be multi-faceted, it requires responses that are topically specific and tailored to the particular situation. Generic and repetitive responses generated by end-to-end models do not suit this purpose. So, we try to find solutions to how we can generate reliable and topically specific responses to distress-related dialogue prompts. As an attempt to solve this, we investigate the ability to utilize knowledge and reasoning over graph-based representations to generate empathetically appropriate, reliable, and topically specific responses by identifying relatable topics in the graph and directing the flow of the conversation along predictable routes (Liu et al., 2019).

   We break this down to the following research questions and attempt to develop solutions to each of them: How to curate dialogues that represent a wide range of distress-related topics? How to preprocess this data to filter out responses that are unreliable and contain profanity? How to recognize different types of distress-related topics present in this dialogue data and the common types of responses associated with them leading to emotional relief? How to explicitly contain this knowledge in a structure that is easy to access and be referred to by conversational agents when generating responses to address distress? How do agents developed based on this knowledge compare to the state-of-the-art empathetic conversational agents?

5. **What are the differences between responses given by peers and trained professionals for addressing distress-related prompts? (Chapter 6)**

   Due to the lack of large-scale psychotherapeutic datasets, researchers are seen to resort to less ideal means such as scraping dialogues from online peer-support forums to enable chatbot models to respond to distress-related prompts (Sharma et al., 2020b; Roller et al., 2021; Brown et al., 2020b). Prior work has observed peers' responses to be rich in information and contain high levels of empathy (Nambisan, 2011; De Choudhury and De, 2014; Sharma et al., 2020a,b). But since peers are not professionals, there can exist significant flaws in the responses given by peers compared to those from professionals, which can thus impose severe risks on chatbots developed using this data. So, it is vital to investigate what differences exist between the responses given by peers and trained professionals when addressing psychological distress and use these observations to make chatbot responses more aligned with responses from professionals.

In this regard, we attempt to find solutions to the following questions: How to curate a dataset that represents responses formulated by both peers and professional counselors to distress-related dialogue prompts? What are the schemes available in the psychological literature to assess the quality of responses addressing psychological distress? How to design a human computation experiment to assess dialogue responses formulated by peers and professional counselors using such a scheme? How can we control and evaluate the quality of the human computation experiment? How do peers' responses differ from counselors' responses? Can these observations support the process of enabling chatbots trained on peer support dialogues to sound more professional and adherent to therapeutic norms?

6. **How to make distress support chatbot responses sound more professional and adherent to therapeutic norms that elicit empathy in therapeutic settings? (Chapter 7)**

   After recognizing the differences between distress-support responses given by peers and professionals, it is important to investigate how we can use these observations to make chatbot responses trained on abundantly available peer-support dialogues reflect therapists' communication style and be more empathetic and compliant with therapeutic norms. This could be similar to the task of Text Style Transfer (TST), in which the goal is to automatically control the style attributes of text while preserving the content (Jin et al., 2022). Though attempts have been made to rewrite text by transferring attributes such as sentiment (Shen et al., 2017), formality (Rao and Tetreault, 2018), politeness (Madaan et al., 2020), and humor (Gan et al., 2017), there have been limited attempts at text style transfer between the responses given by peers and therapists (Sharma et al., 2021). The non-existence of parallel datasets composed by professional counselors rewriting peers' responses into more professional ones makes this task even more difficult and worth exploring.

   We break this down to the following research questions and attempt to solve each of them: How to recognize response types that conform and non-conform to therapeutic norms? In the event of responses non-conforming to therapeutic norms, can we rephrase them into a more conforming form? How can we build such rephrasers in the absence of parallel datasets containing human rephrasings? How can we develop effective rephrasers in the presence of less training data? Using these, how can we build an end-to-end pipeline to boost chatbot responses to sound professional and adherent to therapeutic response strategies?

# 2 A Taxonomy of Empathetic Response Intents in Human Social Dialogues

*This chapter is based on the work of Anuradha Welivita and Pearl Pu (Welivita and Pu, 2020). The author of this thesis (Anuradha Welivita) was mainly responsible for the development of the taxonomy of empathetic response intents, its expansion to the EmpatheticDialogues dataset, and visual analysis of empathetic response patterns.*

## 2.1 Introduction

Inspired by the recent success of deep neural networks for natural language processing (NLP) tasks such as language modeling (Mikolov et al., 2010) and machine translation (Sutskever et al., 2014), neural response generation is at the forefront of research in the NLP community. Recent advances in this field have proven the efficacy of deep neural networks in modelling both task-oriented and open-domain dialogue systems (Wen et al., 2015; Sutskever et al., 2014; Vinyals and Le, 2015b). Most of the existing neuralconversation models are capable of generating syntactically and contextually well-formed responses. Some of the work also focuses on enabling chatbots to generate emotionally colored and affect-rich responses (Asghar et al., 2018; Zhou et al., 2018a; Xie et al., 2020). Despite the efforts in modeling affect in natural language, work that focuses specifically on modeling empathy in chatbots is relatively limited and remains an open research area (Spring et al., 2019).

Empathy plays a vital role in human psychological processes for smooth social interaction (Decety, 2010). Empathy-related responding includes caring and sympathetic concerns for other people. Humans are born with core *affect* neural circuitry, and they gradually develop the ability to apprehend the emotional states of others and respond in an empathetic manner. Empathy motivates pro-social behavior and increases the sense of social bonding (Eisenberg and Eggum, 2009). Therefore, in the context of social interaction, a chatbot needs to be empathetic to maintain healthy interaction with humans and develop trust. The task of augmenting social chatbots with empathy is challenging because the generated responses have to be appropriate in terms of both content and emotion information (Spring et al., 2019).

Several neural response generation models have attempted to address this challenge in a fully data-driven manner. For example, Rashkin et al. (2019), use the full transformer architecture (Vinyals and Le, 2015b) pre-trained on 1.7 billion Reddit conversations and fine-tuned on the EmpatheticDialogues dataset (Rashkin et al., 2019) to generate empathetic responses. Lin et al. (2020) adapt the Generative Pretrained Transformer (GPT) (Radford et al., 2018) to empathetic response generation task by fine-tuning it on the PersonaChat (Zhang et al., 2018) and EmpatheticDialogues datasets. Even though these models are capable of mimicking human empathetic conversation patterns in some ways, it is often unpredictable what the chatbots might generate, for example, they may generate inconsiderate remarks, redundant responses, asking the same questions repeatedly, or any combinations of them. Since it is really important to respond to humans' emotions appropriately, we believe controllability of response generation is essential.

Several other neural response generation approaches attempt to gain control over the generated response by conditioning it on a manually specified emotion label (Zhou et al., 2018a; Zhou and Wang, 2018; Hu et al., 2018; Song et al., 2019) or using affective loss functions based on heuristics such as minimizing or maximizing affective dissonance between prompts and responses (Asghar et al., 2018). These models claim to generate emotionally more appropriate responses than those generated from purely data-driven models. However, the primary concern of these handcrafted rules is their practicality. Despite the various theories related to emotion and empathy that exist in the literature (Zillmann, 2008; Gordon, 1992; Gopnik and Wellman, 1992; Singer and Klimecki, 2014), no prior work has shown normative associations between the speaker's emotions and the corresponding listener's emotions. As our work would reveal, listeners are much more likely to respond to sad or angry emotions with questioning than expressing similar or opposite emotions in the first turn. Xu et al. (2018), however, has shown the benefit of incorporating dialogue acts as policies in designing a social chatbot. They were able to avoid the need to manually condition the next response with a label by jointly modeling dialogue act selection and response generation. Their framework first selects a dialogue act from a policy network according to the dialogue history. The generation network then generates a response based on both dialogue history and the selected dialogue act. It is thus possible to explicitly learn human-human conversational patterns in social chitchat and generate more controlled and interpretable responses. Unfortunately, they did not study empathetic response generation.

To fill this gap, we have developed a taxonomy of empathetic listener intents by manually annotating around 500 utterances of the EmpatheticDialogues dataset (Rashkin et al., 2019), covering 32 types of emotion categories. In the following, we first describe in detail how this taxonomy was derived (Figure 2.1) and how we chose the dataset to support this annotation work. To extend this subset, we employ automatic techniques to label all speaker and listener utterances, covering 25k empathetic human-human conversations. To be able to explain the patterns and trends of the conversation flow, we employ visualization methods to illustrate the most frequent exchanges and reveal how they temporally vary as dialogues proceed. Finally, we discuss how these results can be used to derive more informed heuristics for controlling the neural response generation process.

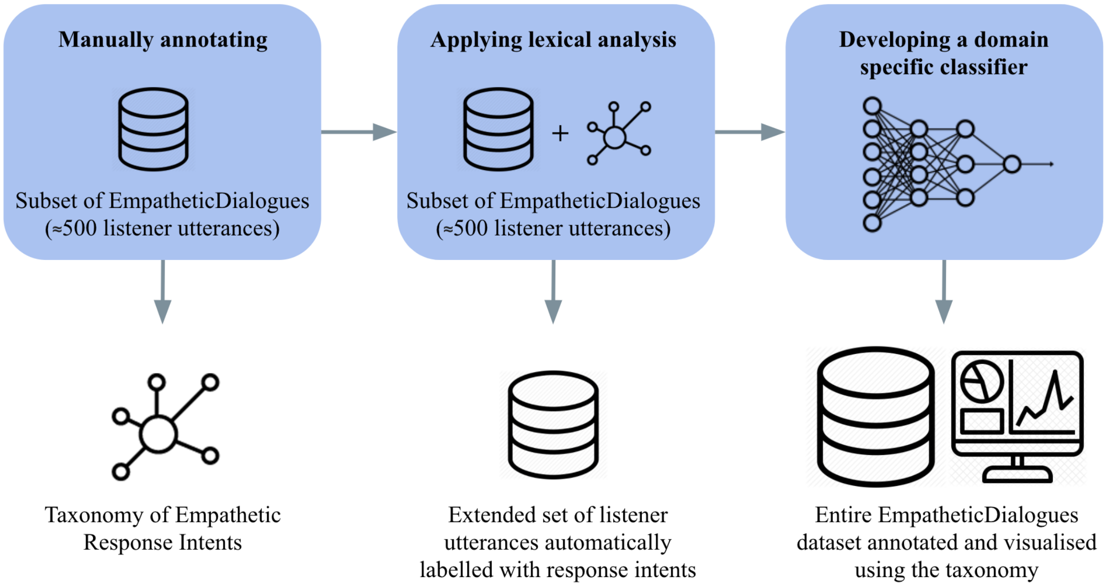



Figure 2.1: Three development steps for constructing the taxonomy of empathetic response intents.

## 2.2 Related Work

To provide a background for this research, we begin by describing some of the existing theories related to empathy in other fields such as psychology and neuroscience and their limitations in incorporating them into the design of social chatbots. Then we describe seminal work on existing neural-based open-domain response generation systems and means by which they control the generated response. These studies serve as the motivation and inspiration for our work. Next, we discuss some existing dialogue-act/intent taxonomies and their limitations in modeling empathy in human social conversations.

### 2.2.1 Theories of Empathy in Psychology and Neuroscience

We can find many theories in psychology and neuroscience that describe empathy. Zillmann (2008) defines empathy as a social emotion in response to the emotions of others. Further, he states that the evoked empathetic reaction itself constitutes an emotional experience, primarily because it is associated with increased excitement and awareness. It tends to be a *feeling with* or *feeling for* the observed party. Two of the most famous theories in the psychological literature to explain the phenomenon of empathy are the “simulation theory" (Gordon, 1992) and the “theory-theory" (Gopnik and Wellman, 1992). The “simulation theory" states that a person understands another or empathizes by imagining himself in the other’s situation and seeing it from his perspective. Mentally simulating one’s experience in turn enables us to understand and relate to what the other person is feeling. The “theory-theory" in contrast, states that the ability to understand what another person is feeling is based on existing knowledge defining how one should think and feel. This theory suggests that empathy is based on a more abstract and cognitive understanding of the other person’s mental state. Accordingly, people rely on their beliefs, desires, and intentions to infer the mental state of the other.

A recent work by Singer and Klimecki (2014) in the field of Neuroscience states that empathy refers to the general capacity of humans to resonate with others’ emotional states irrespective of their valence. However, when empathizing, they suggest that one should not confuse oneself with the other; i.e., one should still know that the emotion he resonates with is the emotion of another. The failure to separate that can lead to empathetic distress. According to them, the desirable way of empathizing with others is having compassion, which is a feeling of concern for another person’s suffering accompanied by the motivation to help. However, all this work does not describe in detail specific means through which humans show empathy, especially via natural language dialogues. Also, most of these studies on empathic states focus on reactions to negative rather than positive events (Buechel et al., 2018). Hence, empathy for positive events remains less understood. In our study, we explore the means through which humans empathize with others both in positive and negative scenarios.

### 2.2.2 Conditional Neural Response Generation

Chen et al. (2019a), propose a model that can generate comments to posts in social media so that they are not only relevant in topic but also in emotion. To fully understand how emotions are expressed in conversations, they first analyse NTCIR-12 STC-1 collection (Shang et al., 2016), a social-media conversation dataset. The results show that for posts with different emotions, the distributions of comment emotions are very different from each other, and only several emotions are appropriate for responding to a given post. Inspired by the findings, they extend the basic encoder-decoder neural network architecture (Vinyals and Le, 2015b) with an RNN-based response emotion estimator, which takes in a post and estimates how relevant an emotion is for responding to the post. This information is fed into the decoder when generating the response. In this, the classifier automatically determines the emotion of the response. Xu et al. (2018) incorporate dialogue acts as policies in their open-domain neural response generation model by performing learning with human-human conversations tagged with a dialogue act classifier. They jointly model dialogue act selection and response generation using a GRU based neural network consisting of a policy network and a generation network. The policy network first selects a dialogue act according to the conversation history, and then the generation network generates a response based on the conversation history and the selected dialogue act. They claim that with dialogue acts, they not only achieve significant improvement over response quality for a given context but also can explain why such achievements are possible. The above work motivated us to develop explicit empathetic response intents from the dataset. We believe they can inform the development of empathetic social chatbots by providing more control to and interpretation of the responses generated and render human-machine conversations more natural and engaging.

### 2.2.3 Dialogue-Act/Intent Taxonomies

Work has been conducted to establish dialogue act/intent taxonomies both by analysing human-human and human-machine conversation datasets. Stolcke et al. (2000) propose a taxonomy of 42 mutually exclusive dialogue acts with the intention of enabling computational dialogue act modeling for conversational speech. They follow the standard Dialog Act Markup in Several Layers (DAMSL) tag set (Core and Allen, 1997) and modify it in several ways so they can easily distinguish utterances in conversational speech. Using this taxonomy, they produce a large hand-labeled database of 1,155 conversations from the Switchboard corpus of spontaneous human-to-human telephone conversations (Godfrey et al., 1992), which is widely used to train and test dialogue act classifiers. Montenegro et al. (2019) propose a dialogue act taxonomy for a task-oriented virtual coach designed to improve the lives of the elderly. It is a multi-dimensional hierarchical taxonomy comprising of topic, intent, polarity, and entity labels at the top, in which the intent label classifies the utterance in classes related to the user's communicative intentions such as 'question', 'inform', and 'agree'. They use the taxonomy to manually annotate the user turns in 384 human-machine dialogues collected from a group of elderly. It aims to help the dialogue agent to detect goals, realities, obstacles, and ways forward of the particular topics the agent is designed to deal with.

Existing dialogue-act/intent taxonomies are either too general as they were constructed for open-domain conversations or too specific as they were constructed for specific task-oriented scenarios. These taxonomies do not necessarily model empathy in human social conversations. Also, the above approaches do not use automatic approaches to extend the manual annotations, which make their datasets comparatively smaller. In our study, we present how a smaller set of human labeled sentences can be extended using lexical methods and use it to train a classifier to automatically annotate a larger corpus.

## 2.3 Dataset

Many open-domain conversation datasets are publicly available mainly to assist tasks such as neural dialogue generation. Out of them, some datasets are multi-modal (e.g. IEMOCAP (Busso et al., 2008), SEMAINE (McKeown et al., 2011), MELD (Poria et al., 2019)) containing visual, acoustic and textual signals. Since they contain a lot of back-channel communication through facial expressions and speech tones, the text may not fully represent the contextual expression of intent. Datasets containing dialogues extracted from social media platforms such as Twitter (e.g., the Twitter Dialog Corpus (Serban et al., 2017a)) are often noisy, short, and different from real-world conversations and may contain a lot of toxic responses rather than compassionate ones. Also, datasets containing TV or movie transcripts (e.g., Emotionlines (Hsu et al., 2018), OpenSubtitles (Lison et al., 2018)) and telephone recordings (e.g. Switchboard corpus (Stolcke et al., 2000)) are a translation of voice into text, which does not fully model interactions that happen only through text. Even purely text-based daily conversation datasets such as DailyDialog (Li et al., 2017b) are not guaranteed to contain empathetic responses.

Rashkin et al. (2019) introduced the EmpatheticDialogues dataset consisting of 24,856 open-domain, human-human conversations as a benchmark dataset to train and evaluate dialogue systems that can converse in an empathetic manner. Each conversation in this dataset is based on a situation associated with one of 32 emotions, which are 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). The dialogues are collected using ParlAI (Miller et al., 2017a), integrated with Amazon Mechanical Turk (MTurk), recruiting 810 US workers. During construction, the workers were instructed to show empathy when responding to conversations initiated by their speaker counterparts. Since almost all the dialogues in this dataset are empathetic, purely text-based, and most of which do not contain any toxic responses, we chose it to derive our taxonomy. An example conversation from this dataset is given in Table 2.1. Table 2.2 shows the basic statistics of the dataset. The average number of turns per dialogue is close to 4. The maximum number of dialogue turns in the dataset is 8. However, not many dialogues exceed 4 turns. Close to 77% of the total number of dialogues contain only up to 4 turns, and only 1.4% of the dialogues contain up to 8 turns.

| | |
|---|---|
| Label: | Afraid |
| Situation: | Speaker felt this when... *"I've been hearing noises around the house at night"* |
| Conversation: | Speaker: *I've been hearing some strange noises around the house at night.* Listener: *oh no! That's scary! What do you think it is?* Speaker: *I don't know, that's what's making me anxious.* Listener: *I'm sorry to hear that. I wish I could help you figure it out* |

Table 2.1: Example conversation taken from the EmpatheticDialogues dataset (Rashkin et al., 2019).

## 2.4 Taxonomy of Empathetic Response Intents

To develop the taxonomy, we investigated which intents are frequently associated with listeners when responding to different emotional situations in EmpatheticDialogues. We took a subset of the dataset with situations associated with the Plutchik's 8 basic emotions (Plutchik, 1984) (joyful, anticipating, trusting, surprised, angry, afraid, sad, and disgusted), and manually analysed it to derive the listener intents associated with each type of emotion. In this process, 20 dialogues belonging to each emotion were randomly selected and each sentence in all listener utterances were manually annotated by an expert evaluator with a label that best

| Criteria | Statistics |
|---|---|
| Total no. of dialogues | 24,856 |
| Total no. of dialogue turns | 107,247 |
| Average no. of turns per dialogue | 4.31 |
| Maximum no. of turns per dialogue | 8 (345 dialogues) |
| Minimum no. of turns per dialogue | 1 (3 dialogues) |
| Total no. of speaker turns | 55,984 |
| Total no. of listener turns | 51,263 |
| Average no. of speaker tokens per dialogue turn | 17.88 |
| Average no. of listener tokens per dialogue turn | 13.69 |

Table 2.2: Statistics of the EmpatheticDialogues dataset used for analysis.

describes their intent. This resulted in 521 sentences manually annotated with intent labels. Because an utterance can have multiple sentences, we decided to annotate each sentence in a listener's utterance with a unique intent label. For example, the two sentences comprising the utterance *"Those symptoms are scary! Do you think it's Corona?"* would be annotated with separate intent labels "Acknowledging" and "Questioning", respectively. After analysing their occurrences and whether some of the intents can be grouped into a common intent, we were able to come up with a taxonomy of 15 empathetic response intents. Table 2.3 presents this taxonomy with corresponding examples and occurrence frequencies. Words and phrases that were most helpful in annotating these examples with their corresponding intents are underlined. Manual annotation of empathetic response intents was carried out with reference to the context preceding an utterance. This way, we were able to distinguish utterances using similar words in the same order depending on their context. For example, sentences such as *"I hope they find a vaccine soon."* can be categorised into two different intents, "Encouraging" and "Consoling" depending on whether the sentence follows a positive or negative emotional context, respectively.

| Category | Examples | Frequency |
|---|---|---|
| 1. Questioning (to know further details or clarify) | - *<u>What</u> are you looking forward to<u>?</u>* | 24.38% |
| 2. Acknowledging (Admitting as being fact) | - *That <u>sounds like</u> double good news. It was <u>probably fun</u> having your hard work rewarded.* | 22.46% |
| 3. Agreeing (Thinking/Saying the same) | - *That's a great feeling, <u>I agree!</u>* | 9.60% |
| 4. Consoling | - *<u>I hope</u> he gets the help he needs.* | 7.87% |
| 5. Encouraging | - *<u>Hopefully</u> you will catch those great deals!* | 5.37% |
| 6. Sympathizing (Express feeling pity or sorrow for the person in trouble) | - *So <u>sorry to hear</u> that.* | 5.37% |

| | | |
|---|---|---|
| 7. Wishing | - *Hey... congratulations to you!* | 4.41% |
| 8. Suggesting | - *Maybe you two should go to the pet store to try and find a new dog for him!* | 4.03% |
| 9. Sharing own thoughts/opinion | *I would love to have a boy too, but I'm not sure if I want another one or not.* | 4.03% |
| 10. Sharing or relating to own experience | *I had a friend who went through the same thing.* | 3.84% |
| 11. Advising | *Don't take too much money with you.* | 2.69% |
| 12. Expressing care or concern | *I hope the surgery went successfully and with no hassle.* | 2.30% |
| 13. Expressing relief | *Phew.. That's a relief., I am glad you were okay.* | 1.53% |
| 14. Disapproving | *But America is so great now! look at all the great things that are happening.* | 1.15% |
| 15. Appreciating | *You are very trusting. It's nice to have a friend like you.* | 0.95% |

Table 2.3: Taxonomy of empathetic response intents with corresponding examples and occurrence frequencies based on the manually annotated 521 listener utterances in the EmpatheticDialogues dataset.

## 2.5 Automatic Labelling of EmpatheticDialogues Using the Taxonomy

### 2.5.1 Annotation Procedure

To annotate all the speaker and listener utterances in the EmpatheticDialogues dataset with emotion labels and response intents, we trained a BERT transformer-based classifier, as suggested by Devlin et al. (2019). Prior to selecting BERT as the classifier, we trained and tested a FastText classifier on the annotation task, but its accuracy was lower compared to BERT. We proceeded with the 8 most frequent intents (questioning, acknowledging, agreeing, consoling, encouraging, sympathizing, wishing, and suggesting) in our taxonomy of empathetic listener intents and the 32 types of emotion categories given in the EmpatheticDialogues dataset. The rest of the listener intents were classified as 'neutral' since the emotion behind those intents were more on the neutral side. To expand the training data collected by manual annotation, we searched through the rest of the dataset using n-grams that are most indicative of the intent categories. For example, n-grams such as '100 %', 'absolutely', 'definitely', 'i agree', 'me neither', 'me too', and 'i completely understand' are indicative of the intent 'agreeing' and were used to collect more example utterances corresponding to that category. The most indicative n-grams used to collect more utterances for each of the intents are listed in Appendix A.1.

During training, we initialized the representation network with weights from the pre-trained language model, RoBERTA (Zhuang et al., 2021), and fine-tuned the model on situation descriptions given in the EmpatheticDialogues dataset tagged with 32 emotions and listener utterances tagged with 8 out of 15 intents from our taxonomy of empathetic response intents. The training, validation, and test sets comprised of 25023, 3544 and 3225 sentences respectively, which spanned equally across all emotion and intent categories. We trained the model with a peak learning rate of $2e^{-5}$ and a batch size of 32 for 10 epochs and obtained the classifier giving the lowest validation loss. The top-1 accuracy of our classifier with 41 labels over the test set was 65.88%, which is significantly higher than the accuracy of FastText (Joulin et al., 2017) and DeepMoji (Felbo et al., 2017) classifiers trained on 32 emotion labels in the EmpatheticDialogues dataset. The latter two were considered as the state-of-the-art at that time, and achieved 43% and 48% accuracy on the EmpatheticDialogues test set, respectively (Rashkin et al., 2019).

### 2.5.2 Analysis of Emotion-Intent Exchange Patterns

Based on the above annotations, we analysed the most frequent response intents corresponding to different emotions expressed by speakers. In Figure 2.2, we visualize the emotion-intent exchanges taking place between speakers and listeners in the EmpatheticDialogues dataset. In this, each chord connects emotion-intent pairs that co-occur together. The chord leaving a particular arc represents the speaker's emotion or intent and gets connected to the arc representing the listener's emotion or intent that immediately follows. It can be seen that a significant proportion of speaker utterances contain a particular emotion in the 32 different emotion categories defined in EmpatheticDialogues, while most of the listener utterances contain a particular intent defined in our taxonomy. Instead of conveying a particular emotion, the listeners show their empathy via specific means described in our taxonomy. And the proportions of the arcs for each intent resembles the frequencies in the manually annotated subset. It serves as a validation that our taxonomy is indeed true for listener responses as it was applied to the entire dataset.

It can be seen that 'questioning' and 'acknowledging' play a significant role in empathetic responses irrespective of the speaker's emotion—whether it is subtle or intense or has a positive or negative valence. Questioning enables the listener to sound more attentive and show interest in what the speaker describes. It prevents listeners from arriving at early conclusions, without knowing the situation in detail. It is also important to let speakers know that they have the right to feel the way they feel, even though listeners may not completely agree with their choices. Expressions of 'acknowledgment' serve this purpose. This type of emotional interaction allows the speaker to elaborate on his feelings and what he is going through and feel validated at the same time. It can also be seen that some listener intents such as 'encouraging' and 'wishing' are frequently associated with positive speaker emotions, and some intents such as 'sympathizing' and 'consoling' are frequently associated with negative speaker emotions. A list of example utterance-response pairs corresponding to some of the most frequent emotion-intent exchanges ($\geq 100$ times out of $\approx 50$k utterance-response pairs

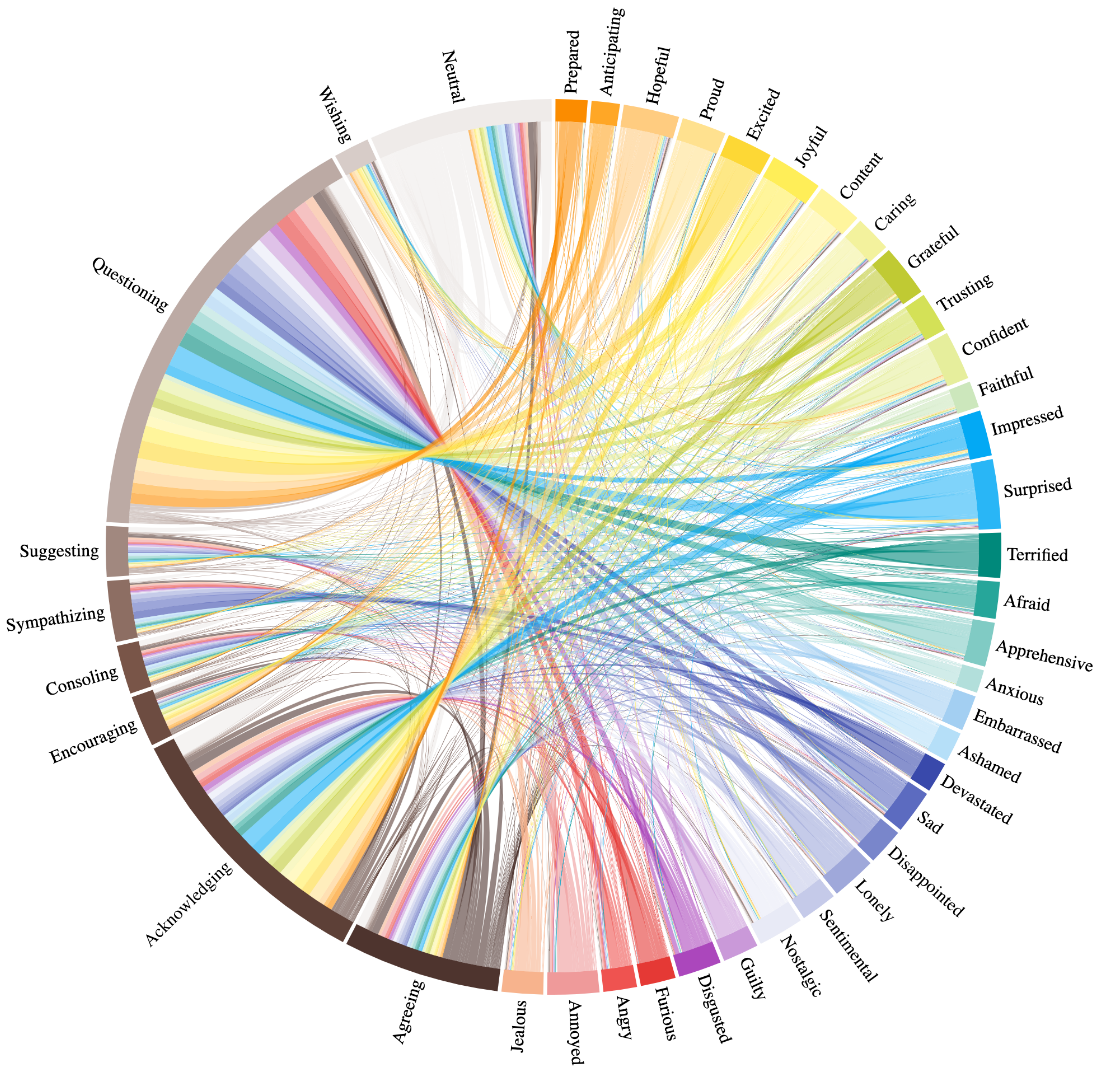



Figure 2.2: Visualization of emotion-intent exchanges between speakers and listeners in the EmpatheticDialogues dataset irrespective of the dialogue turn. Each chord connects co-occurring emotion-intent pairs. The chord leaving a particular arc represents the speaker's emotion or intent and gets connected to the arc representing the listener's emotion or intent that immediately follows in a conversation.

in EmpatheticDialogues) are given in Appendix A.2.

Next, we analysed how emotions and response intents shift over different turns in the dialogue as the dialogues progress in time. In this analysis, we discovered the most frequent emotion-intent flows that occur between speakers and listeners from the start to the end of conversations. To visualize the shift in emotions and intents over different dialogue turns, we computed the frequency of emotion-intent flow patterns up to 4 dialogue turns and plotted the ones having a frequency ≥ 5. The reason for selecting only the first 4 turns in the dialogues is the fact that close to 77% of the dialogues in the EmpatheticDialogues dataset contain up to

4 turns and from this only 1.8% of the dialogues go up to 8 turns. Since there is comparatively much fewer data over dialogue turns from 5 to 8, we omitted these turns in our analysis.

Figure 2.3 plots the most frequent emotion-intent flow patterns up to 4 turns in the dataset. Turns 1 and 3 correspond to speaker turns and turns 2 and 4 correspond to listener turns. According to the visualization, most emotions experienced by speakers are immediately followed by 'questions' as well as expressions of 'acknowledgment'. Expressions of 'sympathy' immediately follow more negatively intense emotions such as 'devastated' and 'sad'. Towards the end of dialogues, we can see more expressions of 'acknowledgment', 'agreement' and 'suggesting'. Expressions of 'encouragement' and 'wishing' can be seen in the case of positive emotional situations and 'sympathizing' and 'consoling' in the case of negative emotional situations. Another important observation is that towards the end of dialogues, listener utterances become more emotional compared to the beginning, as the speakers elaborate on their emotions. Such situations also reflect scenarios of personal distress—the phenomenon where one is unable to distinguish the emotion of their own from the emotion of another. Dialogue in Table 2.4 illustrates this phenomenon.

| |
|---|
| S: *Bleh, I just had the worst food ever.* (Disgusted) |
| L: *What did you eat?* (Questioning) |
| S: *I was at Mcdonalds and was given a rotten cheese burger. I almost puked after I ate it.* (Disgusted) |
| L: *Oh gross, makes me never want McDonalds again.* (Disgusted) |

Table 2.4: Example conversation that illustrates personal distress towards the end of the dialogue.

Still, they are not as frequent as how listeners choose to empathize healthily instead of making it a distress. But this sheds light on the fact that when the speaker goes on elaborating on his situation, sometimes the listener's ability to distinguish between the speaker's emotion and the emotion of his own may decrease, leading to consequences of personal distress. In the case of intense negative emotions, it can lead to avoidance or a deliberate change of conversation topic. This is a scenario commonly experienced by people engaged in therapeutic and health professions and is described in researches by Singer and Klimecki (2014) and Buechel et al. (2018). However, in order to verify this observation more solidly, we need a corpus with a larger number of turns per dialogue.

The taxonomy we have developed can be incorporated into the design of social chatbots to gain more controllability and interpretability of the responses generated. It can be achieved by feeding in the conversation history, in which each utterance is tagged with an emotion or an intent label into a neural network that jointly models dialogue intent selection and response generation. Dialogue intent selection module will select the most appropriate intent based on the conversation history we feed in, and the response generation module will generate an appropriate response conditioned on the selected intent label. To help ensure more

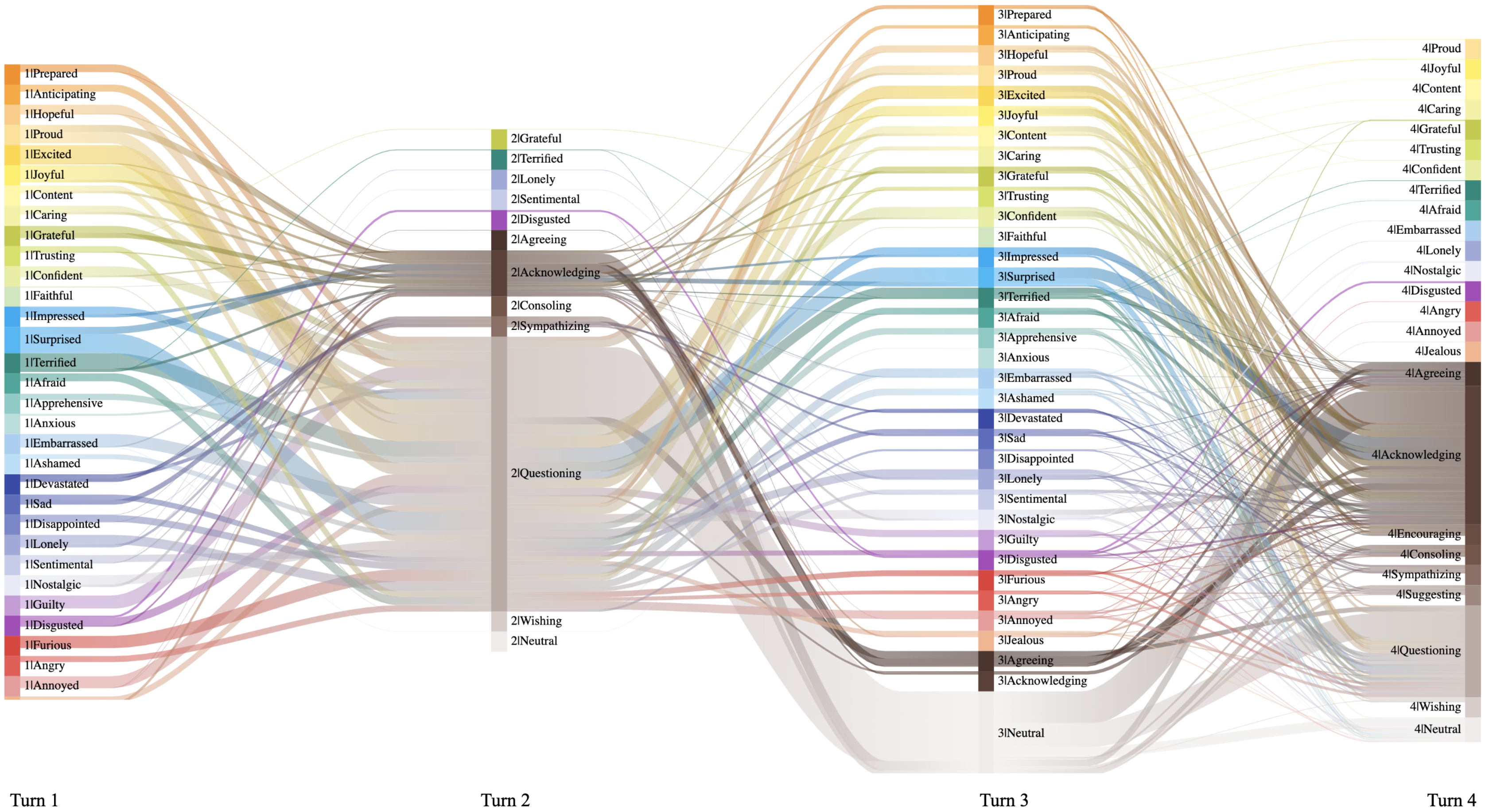



Figure 2.3: Visualization of the most common emotion-intent flow patterns (having a frequency ≥ 5) throughout the first four dialogue turns in the EmpatheticDialogues dataset.

robustness, it is also possible to repeatedly sample plausible intent labels during training and feed them into the response generation module. The overall goal of modeling chatbots in this manner is to lead the conversation in a healthy and desirable direction with the controllability and interpretability provided by the taxonomy. Moreover, the taxonomy can be used as an annotation scheme to label utterances in other datasets and analyse them in terms of their empathetic quality in the same way described here. It also has implications in distinguishing between multiple forms of empathy—compassion and personal distress, as recognized in psychology and neuroscience fields.

One limitation of this study is the analysis results are highly dependent on the EmpatheticDialogues dataset. In the next chapter, we curate a much larger empathetic dialogue dataset using a subset of the OpenSubtitles (≈8M dialogues) (Lison et al., 2018), which helps us develop a more accurate emotion classifier and validate the taxonomy on a different dataset. Another limitation is the emotion classifier trained to automatically label utterances in the EmpatheticDialogues dataset is a sentence-level classifier, which is unable to accurately distinguish similar utterances whose empathetic label can differ according to the context. We intend to improve it into a classifier based on dialogue history that will be able to more accurately distinguish such cases. Automated labeling of intents using lexical methods also has the possibility to injure the robustness of the model due to considering only the most indicative n-grams in individual sentences without accounting for the surrounding context.

## 2.6 Chapter Summary

In this chapter, we introduced a taxonomy of empathetic response intents capable of supporting automatic empathetic communication in social chitchat. The strategies relying on this taxonomy are essential for a chatbot to engage in prosocial conversations, expressing empathetic concern for its users, and keeping the users engaged. Another significant contribution from our work is to provide analysis on the EmpatheticDialogues corpus after automatically annotating it based on the most frequent intents from our taxonomy and 32 types of emotion categories defined in EmpatheticDialogues. We illustrated the most frequent emotion-intent exchange patterns in the dataset and how they vary temporally over the course of interaction. These results further validate the taxonomy of empathetic listener intents we derived and shed light on the frequent empathetic conversation patterns seen among humans when engaged in social chitchat. We explained how our taxonomy can be utilized in the development of an empathetic chatbot to achieve more controllability and interpretability in the responses generation process. The method described here can also be used as an annotation scheme to label utterances from other datasets and analyse them in terms of their empathetic quality. In the subsequent chapters, we use these findings to develop a social chatbot capable of effectively engaging in empathetic conversations.

# 3 A Large-Scale Dataset for Empathetic Response Generation

*This chapter is based on the work of Anuradha Welivita, Yubo Xie, and Pearl Pu (Welivita et al., 2021). The author of this thesis (Anuradha Welivita) was mainly responsible for the development of the data curation pipeline, turn segmentation of the OpenSubtitles dataset, the design and implementation of the human computation experiment, the development of the semi-supervised framework to expand the labels, and the comparison of the curated dataset with the state-of-the-art gold standard.*

## 3.1 Introduction

Researchers are increasingly inclined towards refining pre-trained language models with domain-specific datasets to achieve certain tasks (Devlin et al., 2019; Zhuang et al., 2021; Rashkin et al., 2019). One such area is the development of empathetic conversational agents that can understand human emotions and respond appropriately. The aim of the empathetic response generation task is to generate syntactically correct, contextually relevant, and more importantly emotionally appropriate responses following previous dialogue turns. Such tasks require the creation and availability of large dialogue datasets, in which each utterance is annotated with the correct intents and emotions. Though many such datasets have been developed in the past (Busso et al., 2008; Poria et al., 2019; Li et al., 2017a; Rashkin et al., 2019), due to the cost of manual labor, they are limited in size, thus insufficient to train robust conversational agents. Since collecting and manually annotating such gold standard data is expensive, replacing them with automatically annotated silver standard data has become a rising interest (Filannino and Di Bari, 2015). We show how such a large-scale silver standard dataset with sufficient quality can be curated and used to fine-tune pre-trained language models for the generation of empathetic responses.

Emotions revealed in social chitchat are rather complex. It has many categories of emotions to distinguish due to subtle variations present in human emotion. For example, *Sadness* and *Disappointment* are pursued and dealt with differently in human conversations even though both of them are negative emotions. Also, the listener's reaction to emotion is not always a straightforward mirroring effect of the speaker's emotion. Rather it can be more neutral and convey a specific intent, as is evident from the dialogue example in Table 3.1.

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

Table 3.1: An example showing the listener's reactions to emotions do not always mirror the speaker's emotions.

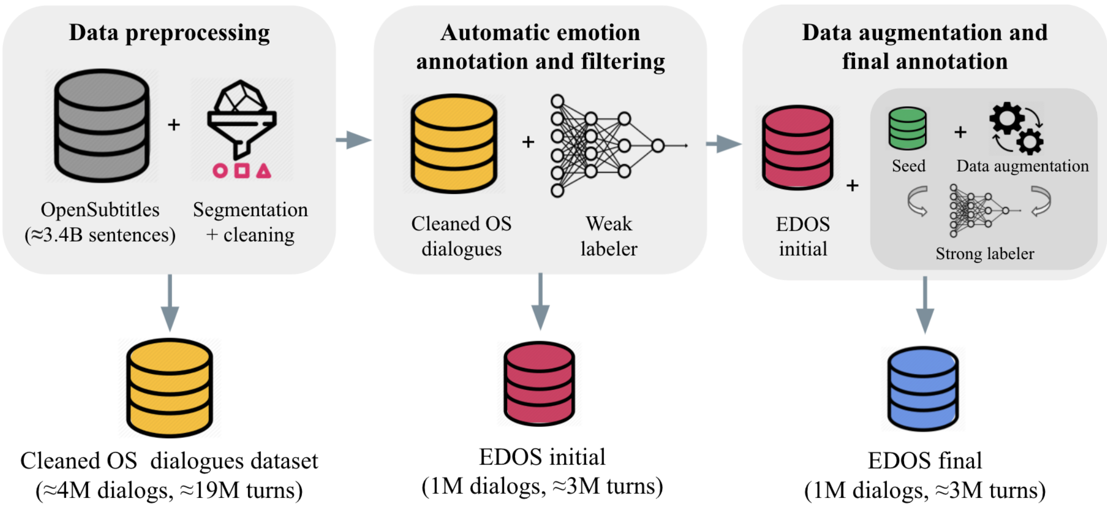



Figure 3.1: Steps for curating the EDOS dataset.

In Chapter 2, we analyzed listener responses in the EmpatheticDialogues dataset (Rashkin et al., 2019) and discovered eight listener-specific empathetic response intents contained in emotional dialogues: *Questioning*; *Agreeing*; *Acknowledging*; *Sympathizing*; *Encouraging*; *Consoling*: *Suggesting*; and *Wishing*. We annotated the EmpatheticDialogues dataset with 32 fine-grained emotions, eight empathetic response intents, and the *Neutral* category, and discovered frequent emotion-intent exchange patterns in empathetic conversations. We observed that this type of dataset tagged with fine-grained emotions and intents can be used to train neural chatbots to generate empathetically appropriate responses. But for this purpose, a large-scale emotion and intent-labeled dataset is even more desirable. Curating such a dataset is technically challenging since 1) annotating such a large-scale dataset require costly human labor, and 2) given the fine-granularity of the emotion and intent labels, the human labeling task is more difficult and error-prone compared to the more coarse grained *Angry-Happy-Sad* emotion categories. As a result, existing manually labeled emotional dialogue datasets such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and DailyDialogue (Li et al., 2017a) are smaller in scale and contain only a limited set of emotions (emotions derived from basic emotion models such as the Ekman's). Most importantly, existing datasets fail to distinguish between *Neutral* and *Questioning*, or any of the other eight empathetic response intents. They combine everything into a big label *Neutral* or *Other* when the utterance is not emotional. But *Questioning*, *Agreeing*, *Acknowledging*, *Sympathizing*, *Encouraging*, *Consoling*, *Suggesting*, and *Wishing* are important details in constructing empathetic dialogues. These eight response intents, which we call the plus categories, are novel in our work and contribute to the model's learning of important response patterns in the data.

To fill the above gap, we curate a novel large-scale silver dialogue dataset, **EDOS** (**E**motional **D**ialogues in **O**pen**S**ubtitles), containing 1M emotional dialogues from movie subtitles, in which each dialogue turn is automatically annotated with 32 fine-grained emotions, eight plus categories as well as the *Neutral* category. Movie subtitles are extensively used for emotion analysis in text in earlier and recent research (Kayhani et al., 2020; Merdivan et al., 2020; Giannakopoulos et al., 2009). The Nature article "How movies mirror our mimicry" (Ball, 2011) states *"screenwriters mine everyday discourse to make dialogues appear authentic"* and *"audiences use language devices in movies to shape their own discourse"*. Hence, it can be one of the major sources to train chatbots and learn emotional variations and corresponding response strategies in dialogues. To reduce the cost of human labeling and the complexity of labeling dialogues with fine-grained emotions and intents, we devised a semi-automated human computation task to collect fine-grained emotion and intent labels for a small set of movie dialogues (9K). We then followed automatic data augmentation techniques to expand the labeled data and trained a dialogue emotion classifier to automatically annotate 1M emotional dialogues. Table 3.2 compares the EDOS dataset with the state-of-the-art emotion-annotated dialogue datasets. Compared to existing datasets, EDOS is much larger in size and contains more fine-grained emotion as well as novel empathetic intent annotations.

The process of curating the dataset involved several stages. First, we applied automatic turn and dialogue segmentation methods, data cleaning and removal of duplicates on movie subtitles in the OpenSubtitles (OS) corpus (Lison et al., 2019) and obtained close to 4M dialogues. Then, we applied a weak labeler (a BERT-based sentence-level classifier) trained on the EmpatheticDialogues dataset (Rashkin et al., 2019), to label utterances in OS dialogues and filtered 1M emotional dialogues (EDOS initial). Thereafter, we applied data augmentation techniques on a small set of human-annotated data and used the manually annotated and extended labels to train a strong labeler that is used to annotate dialogues in EDOS initial and obtained the final 1M EDOS dataset. We evaluated the quality of the resultant dataset by comparing it against the EmpatheticDialogues dataset by means of offline experiments and visual validation methods. Figure 3.1 summarizes the process of creating EDOS. The data curation pipeline we followed substantially reduced the cost of human labor while ensuring quality annotations.

Our contributions in this chapter are three-fold. 1) We curate a large-scale dialogue dataset, EDOS, containing 1M emotional dialogues labeled with 32 fine-grained emotions, eight empathetic response intents (the plus categories), and *Neutral*. Compared to existing dialogue

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

Table 3.2: Comparison of emotion annotated dialogue datasets available in the literature against EDOS.

datasets tagged with emotions, EDOS is significantly larger (≈ 40 times larger than EmpatheticDialogues), and contains more fine-grained emotions and empathetic response strategies. 2) We outline the complex pipeline used to derive this dataset. 3) We evaluate the quality of the dataset compared to a state-of-the-art gold standard dataset using offline experiments and visual validation methods.

## 3.2 Methodology

This section describes the dialogue selection process, the design of the human annotation task, the data augmentation techniques used to expand human-labeled dialogues, and the development of a strong labeler to annotate the dataset.

### 3.2.1 Dialogue curation from movie subtitles

The OpenSubtitles 2018 corpus consists of 3.7M movie and TV subtitles. It comprises 3.4B sentences and 22.2B tokens. It is an excellent source to learn emotional variations in dialogue and corresponding response mechanisms. But due to the absence of speaker markers, movie subtitles do not contain an explicit dialogue turn structure (who speaks what) and specific indicators where one dialogue ends and the next dialogue begins. To overcome the first issue, we reproduced the work by Lison and Meena (2016) to build an SVM-based classifier that determines if two consecutive sentences are part of the same dialogue turn. Our classifier achieved a segmentation accuracy of 76.69%, which is close to the accuracy of 78% that the authors claim. The set of features that gave the best turn segmentation accuracy are:

- Unigram and bi-gram features of adjacent sentences after lemmatization
- First and final tokens of adjacent sentences
- First and final bi-grams of adjacent sentences
- Whether the two sentences belong to the same subtitle block or not (boolean)
- Genre of the movie (*Drama, Crime, Musical* etc.)
- Sentence density of the subtitles file (no. of sentences/subtitle duration)
- Quadratic combinations of the above features with itself and the rest

After performing turn segmentation on the OpenSubtitles corpus, we divided the turns into separate dialogues based on a simple heuristic. If the difference between the end time of the previous turn and the start time of the current turn is more than 5 seconds, we take these two turns as belonging to 2 different dialogues. An exception occurs if this timestamp information is missing in at least one of the turns. In this case, we assume that these two turns appear in the same subtitle block and consider them as belonging to the same dialogue. This way, we formed 9M dialogues from the OpenSubtitles corpus altogether. The choice of 5 seconds to separate dialogues is based on a histogram of time intervals between adjacent subtitle blocks in the OpenSubtitles corpus, which is denoted in Figure 3.2. As it can be observed in the histogram, most of the time gaps fall below 3 seconds. A clear drop in count was observed between 3-5 seconds.

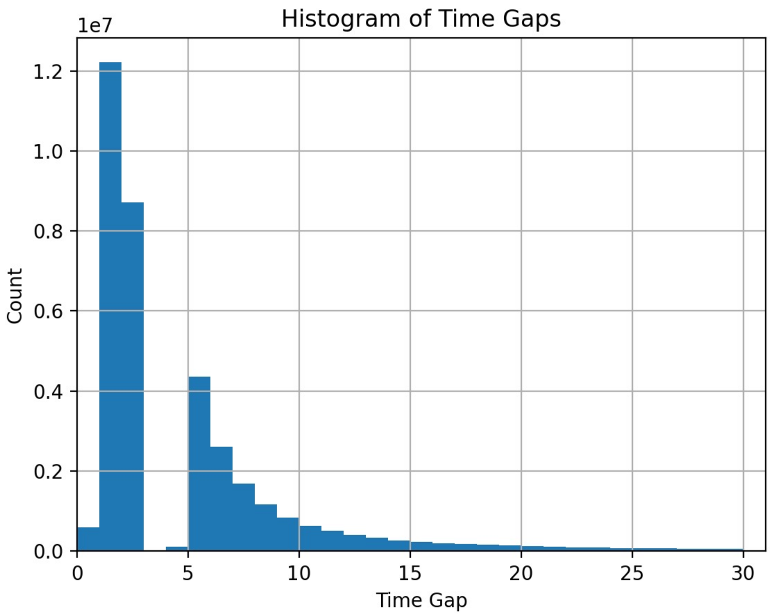



Figure 3.2: Histogram of time intervals between adjacent subtitle blocks in the OpenSubtitles corpus.

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

To filter out dialogues containing emotional statements and empathetic responses from the cleaned OS dialogues dataset, we employed the weak labeler developed in Chapter 2 (BERT transformer-based sentence level classifier), which was trained on 25K situation descriptions from EmpatheticDialogues (Rashkin et al., 2019) tagged with 32 emotion classes, and 7K listener utterances tagged with eight empathetic response intents and the *Neutral* category. This classifier had a high top-1 classification accuracy of 65.88%. We call it a weak labeler since it predicts emotion or intent only at the sentence level and is trained on a different dataset other than OS. We filtered the top 1M dialogues having the highest label confidence as predicted by this classifier to form the 1M EDOS (initial) dataset.

### 3.2.2 Human computation

To train a dialogue emotion classifier that can identify both fine-grained emotions and empathetic response intents, we devised an Amazon Mechanical Turk (AMT) experiment to collect an initial set of ground truth labels for OS dialogues. But annotating dialogue turns with one of 41 labels is a daunting task. To make the task less exhaustive, we devised a semi-automated approach using our weak labeler. By applying the weak labeler on each turn of the cleaned OS dialogue dataset, we filtered out the turns having prediction confidence $\geq 0.9$, along with their dialogue history. Next, we ranked these dialogues according to their readability and selected the highest readable dialogues from each class to be labeled. This is to reduce the time spent by the workers in having to read long and complicated dialogues. The steps followed in computing dialogues' readability are included in Appendix A.3. Workers had to select a label from the top-3 predictions made by the weak labeler. If none of the top-3 predictions matched, they could manually specify the correct class. Appendix A.4 shows the user interface of the AMT annotation task. The main purpose of incorporating a weak labeler here was to make the task less daunting for the crowd worker. Otherwise, having to choose a label out of 41 labels may lead to even worse results due to the complicated nature of the task. The risk of reduced data reliability is avoided by taking only the labels with the majority vote.

After ranking the dialogues according to readability, we selected the top 250 dialogues in each category for the AMT task. We bundled 15 dialogues in a HIT with 5 quiz questions that served as checkpoints to evaluate the crowd workers' quality. Situation descriptions from the EmpatheticDialogues dataset for which we already knew the emotion labels were used to formulate the quiz questions. Finally, we obtained dialogues where we had 2 out of 3 worker agreements, which resulted in $8,913$ dialogues altogether. Table 3.3 shows the results of the AMT task.

| **Description** | **Statistics** |
|---|---|
| Total number of dialogues | $10,250$ |
| Number of dialogues labeled with majority vote | $8,913$(86.96%) |
| Inter-annotator agreement (Fleiss' Kappa) | 0.46 (moderate) |
| Number of times workers got $3/5$ quiz questions correct | 77.75% |
| Number of dialogues in which the workers manually specified the label | 425 |

Table 3.3: The results of the AMT task for annotating a subset of the EDOS (initial) dataset.

### 3.2.3 Data augmentation and annotation

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

1. Using the SBERT model, first, we computed dialogue turn embeddings (each with a vector representation of 768 dimensionalities) for all the turns ($\approx$19M) in the cleaned OS dataset.

2. Then, we calculated dialogue embeddings for human-annotated and unlabeled dialogues from the cleaned OS dialogues dataset. For this, we applied a decaying weight starting from the last turn and took the weighted average of the turn embeddings of each dialogue. We used half decaying, i.e, if we have a dialogue with turn embeddings $v_1$, $v_2$, and $v_3$, the final dialogue embedding would be $(4/7)\,v_3 + (2/7)\,v_2 + (1/7)\,v_1$.

3. Next, we calculated the cosine similarity between annotated and unlabeled dialogue embeddings and ranked the results.

4. Finally, we applied a similarity threshold and obtained all the unlabeled dialogues with a cosine similarity that exceeds this threshold and tagged them with the same crowd annotated class label. Here, we used a threshold of 0.92 after manually inspecting a random subset of the results obtained for a range of thresholds. Examples from this stage are denoted in Table 3.4.

| **Manually annotated dialogues** | **Dialogues discovered using similarity matching (with similarity ≥ 0.92)** |
|---|---|
| - ***That 's beautiful !.*** **(Acknowledging)** | *- Now , let 's take a look at this beautiful piece of work*<br>*- Oh , my God . It 's beautiful .*<br>- ***Oh . That 's beautiful .*** |

| | |
|---|---|
| *- I thought the coils were closer to me .*<br>*- Oh well, It was a good one nonetheless .*<br>- ***I 'm so happy !*** **(Joyful)** | *- Actually , I just wanted to say I love you . And I 'm sorry if I 'm a bit edgy about my book , but all that counts for me is you . You becoming my wife .*<br>*- That 's what really matters .*<br>- ***I 'm very happy .*** |
| *- Hey ! Don 't eat at my house anymore .*<br>- ***You 're disgusting .*** **(Disgusted)** | *- I thought I told you to stay the fuck away from me if you were back on that shit .*<br>- ***You 're disgusting .*** |
| *- Was the team mad , then ?*<br>*- I wasn 't happy !*<br>- ***That 's pretty bad .*** **(Acknowledging)** | *- It 's starting to hurt so bad .*<br>*- Really ? That bad ?*<br>- ***Really bad .*** |

Table 3.4: Examples of similar dialogues discovered above a cosine similarity threshold of 0.92. The last turn in each dialogue discovered through similarity matching was labeled with the emotion or intent of that of the last turn of the manually labeled dialogue.

We extended the original crowd annotated dialogue dataset by $3,196$ more dialogues with distantly annotated class labels using the above method. Thereafter, using the crowd-annotated and extended labels, we trained an initial classifier that we used to annotate the rest of the dialogues and add more labels to our dataset that had annotation confidence over 0.9. This method is termed self-labeling (Triguero et al., 2015), a semi-supervised learning technique that can be used to grow labeled data. With this, we were able to extend the labeled data by $4,100$ more dialogues. Next, we again applied SBERT over the self-labeled data and extended them by $2,118$ more dialogues. Finally, we were able to have $\approx 14K$ labeled dialogues altogether. We used this data to train a final dialogue emotion classifier to annotate the rest of the unlabeled data. This resulted in a classifier with precision 64.11%, recall 64.59%, macro F1-score 63.86%, and accuracy 65.00%, which is comparable with the state-of-the-art dialogue emotion classifiers (as denoted in Table 3.5). The next section elaborates on the design of the dialogue emotion classifier.

### Dialogue Emotion Classifier

Our dialogue emotion classifier consists of a representation network that uses the BERT architecture, an attention layer that aggregates all hidden states at each time step, a hidden layer, and a softmax layer. We used the BERT-base architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters as the representation network. It was initialized with weights from RoBERTa (Zhuang et al., 2021). We fed in a dialogue turn along with the preceding context in the reverse order as input to the representation network. To give more importance to the dialogue turn for which prediction has to be made and the turns that immediately precede it, we multiplied the token embeddings belonging to each turn by a decreasing weight factor. Its input representation is constructed by summing the corresponding token embedding multiplied by the weighting factor and its position embedding.

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

Table 3.5: Comparison of the performance of the dialogue emotion classifier used for annotation with performance of the state-of-the-art dialogue emotion classifiers. F1-score reported here is the macro-F1 score.

Using decreasing weights for context utterances is based on the intuition that in human dialogues, more attention is paid to the most recent utterances in dialogue history. This idea is backed up by time-decay functions used in neural dialogue understanding approaches (See et al., 2019). We conducted an ablation study with and without using decreasing weights in the model. The performance of the unweighted models was lower than the performance of weighted models yielding final F1 scores of 63.44 and 64.86 for unweighted and weighted models, respectively. Details including the hyper-parameters used are included in the Appendix A.5.

## 3.3 Quality Analysis

The statistics of the EDOS dataset are given in Table 3.6. More detailed statistics including the number of dialogues per emotion are included in Appendix A.6.

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

Table 3.6: Statistics of the EDOS dataset.

Table 3.7 shows some example dialogues taken from the EDOS dataset along with annotations and confidence scores. By observing the examples, it could be noticed that even for less confident predictions, the label quite accurately describes the emotion or intent of the corresponding dialogue turn.

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

| | |
|---|---|
| Turn 4 | **(Faithful, 0.86)** We're not leaving. |

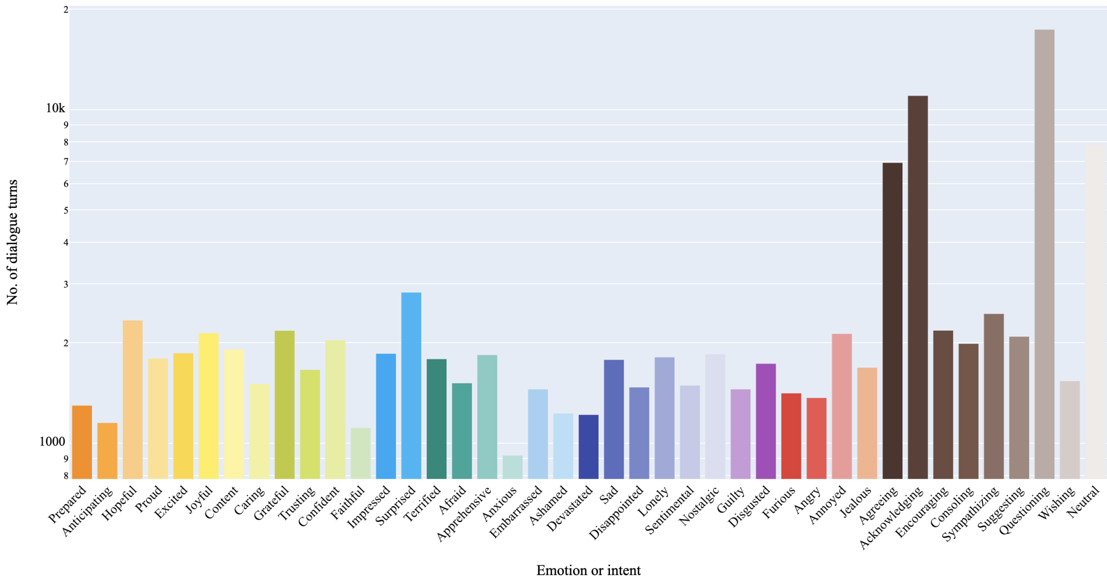
Table 3.7: Example dialogues from the EDOS dataset along with annotations and confidence scores.

We also conducted a qualitative comparison of the annotations in the EDOS dataset with EmpatheticDialogues (Rashkin et al., 2019), the state-of-the-art gold standard dataset for empathetic conversations. Figure 3.3 compares the distributions of emotions and intents in the two datasets. It is observed that in both datasets, intent categories take prominence over individual emotion classes. This is in par with the observations made in Chapter 2, where we noticed that one or more intents from the taxonomy of empathetic intents are mostly utilized when responding to emotions in dialogue, rather than similar or opposite emotions. Especially, the intent *Questioning* takes the highest percentage among the annotations in EmpatheticDialogues and EDOS. We also computed the KL-divergence ($\geq 0$) of the emotion and intent distribution of EDOS with respect to that of EmpatheticDialogues, which measures how one probability distribution is different from a second, reference probability distribution (Kullback and Leibler, 1951). It resulted in a KL-divergence value of 0.2447, which indicates that there is considerable similarity between the two distributions (the lower the KL divergence, the more similar the distributions).

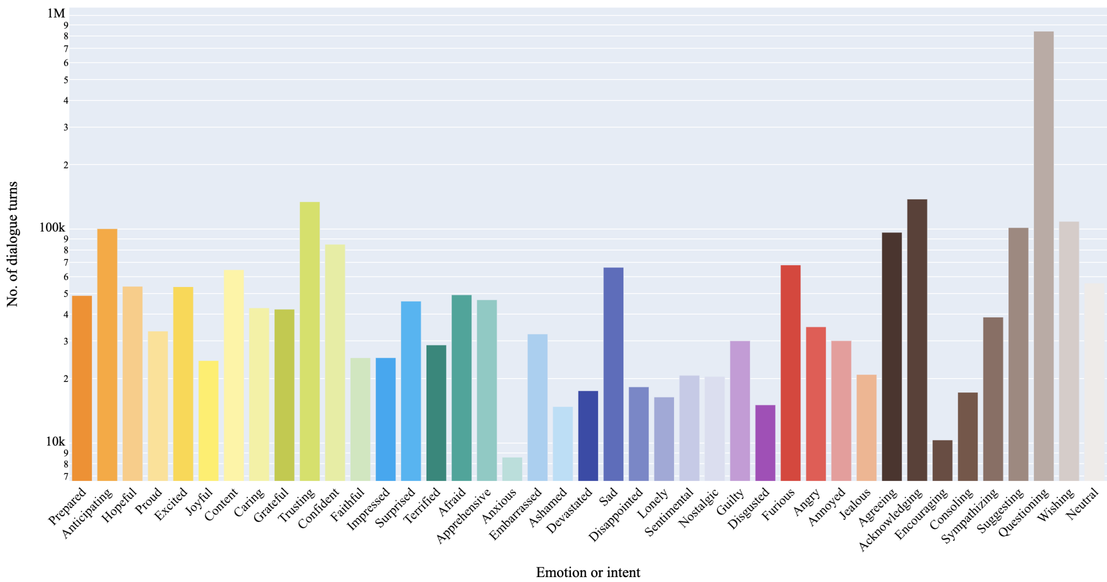
Figures 3.4 and 3.5 show the emotion-intent flow patterns in EmpatheticDialogues and EDOS. In the visualization corresponding to EmpatheticDialogues, the $1^{st}$ and $3^{rd}$ dialogue turns correspond to the speaker and the $2^{nd}$ and $4^{th}$ dialogue turns correspond to the listener. However, in EDOS, we cannot distinguish the dialogue turns as speaker and listener turns due to the absence of speaker annotations. Though this is the case, we could still observe some conversational dynamics present in EmpatheticDialogues are preserved in EDOS. For example, in both datasets, the speaker mostly starts the conversation with some emotional statement and in the subsequent turn, the response tends to be of the intent *Questioning*. In both datasets, intents *Agreeing* and *Acknowledging* follow emotions seen in the first turn irrespective of whether they are positive or negative. As the dialogues proceed, it could be seen in both datasets the emotions deescalate as more empathetic response intents emerge.

## 3.4 Experimental baselines

We propose some experimental baselines using the curated dataset for empathetic response generation and compare the performance against a dialogue model trained on the EmpatheticDialogues dataset. For this purpose, we trained a transformer (Vaswani et al., 2017b) model with various training settings. Specifically, the following datasets were involved: **1) OS dialogues** (As described in Section 3.2.1, these dialogues were obtained by segmenting the movie subtitles. Note that for the purpose of pre-training, we excluded the EDOS dialogues, resulting in around 3M dialogues.); **2) EDOS** (1M dialogues); and **3) EmpatheticDialogues** (25K dialogues). All three datasets were split into a training (80%), validation (10%), and test




(a) EmpatheticDialogues dataset




(b) EDOS dataset

Figure 3.3: Comparison of the distribution of emotions and intents in the EmpatheticDialogues and EDOS datasets.

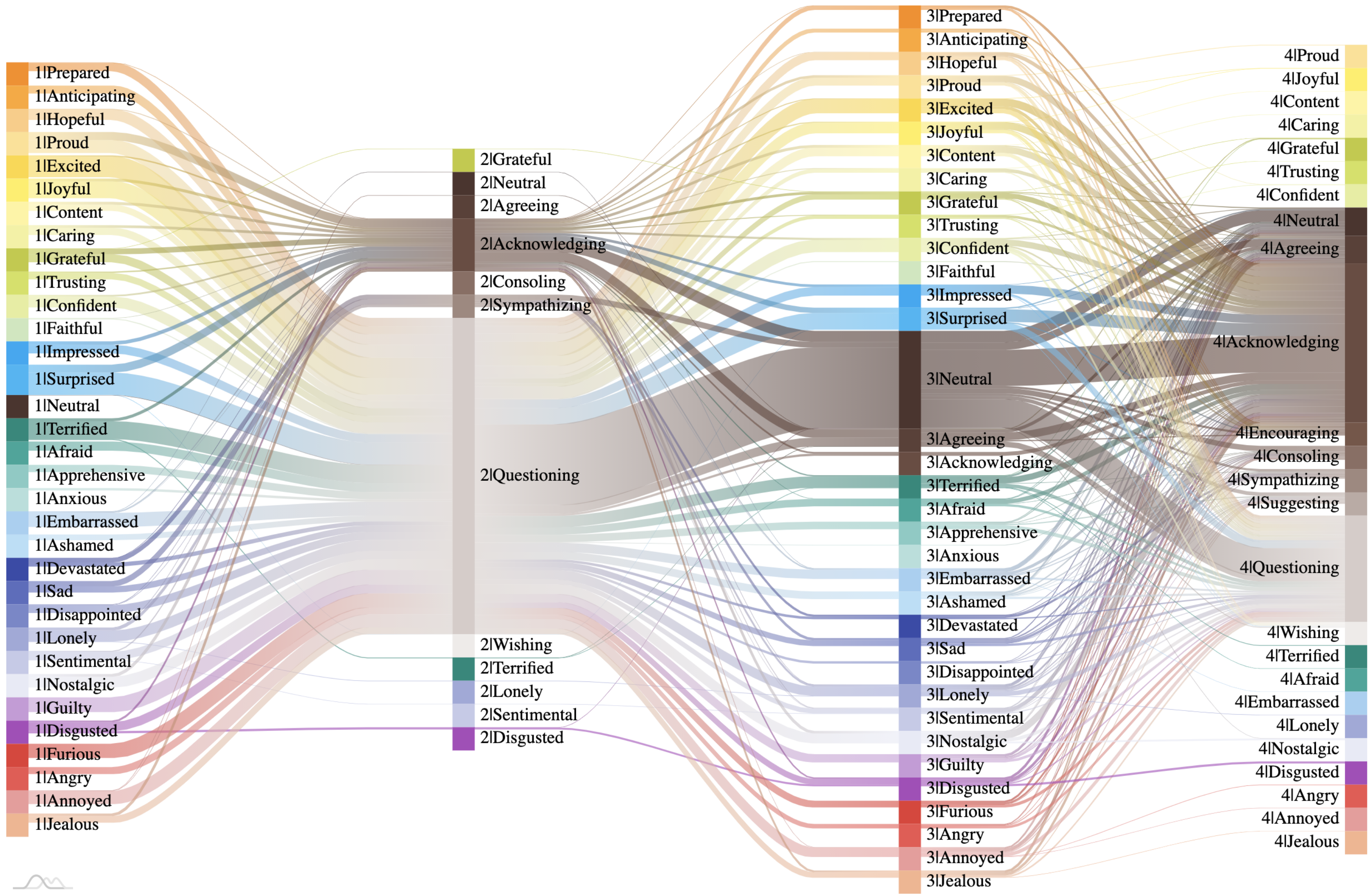



Figure 3.4: The emotion-intent flow patterns of the EmpatheticDialogues dataset. For simplicity, only the first four dialogue turns are visualized.

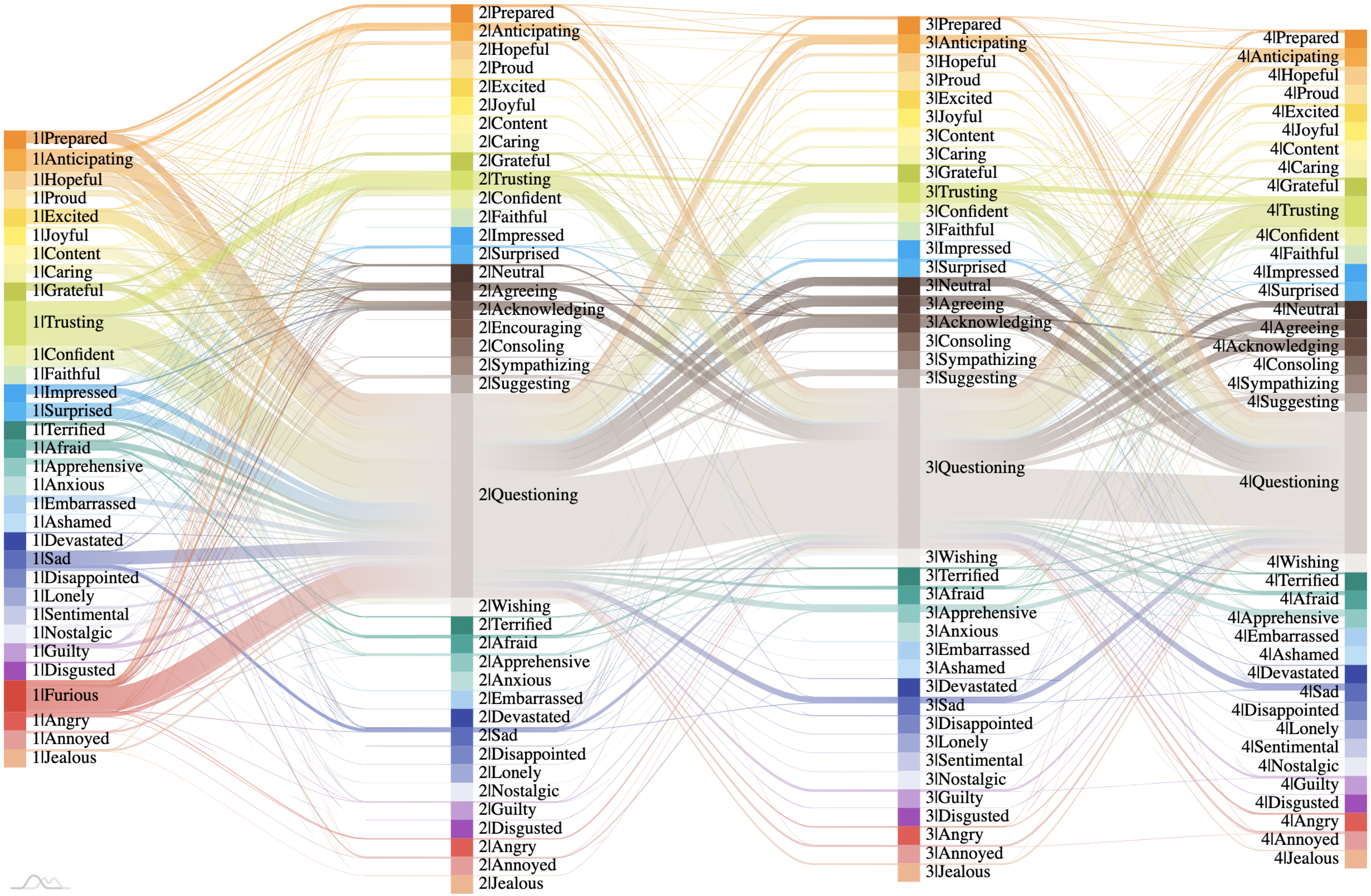



Figure 3.5: The emotion-intent flow patterns of the EDOS dataset. For simplicity, only the first four dialogue turns are visualized.

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

Table 3.8: Dialogue model evaluation results. Here PPL denotes perplexity, D1 and D2 denote Distinct-1 and -2, and SES denotes the sentence embedding similarity. : held-out, : zero-shot.

(10%) sets. Based on the training strategies, we have the following models:

1. **Pre-trained**—to take advantage of transfer learning, we pre-trained the transformer model on the 3M OS dialogues. The large scale of this training set is expected to provide a good starting point for fine-tuning.

2. **Fine-tuned**—we took the pre-trained transformer and then fine-tuned it on EDOS and EmpatheticDialogues datasets respectively. All the models have 4 layers, 6 multi-heads, and a hidden size of 300, and were trained until the minimum validation loss was reached. For inference, we used beam search with beam size 32 and 4-gram repeats blocking.

Additional training details about the experimental baselines are included in Appendix A.7.

To evaluate the performance of the dialogue models, we adopted the following metrics:

- **Perplexity**. Perplexity measures how well a probability model predicts a given sample. In our case, a lower perplexity score indicates better capability of generating the ground-truth response.

- **Distinct-1 and -2**. The Distinct-1 and -2 metrics (Li et al., 2016b) measure the diversity of the generated responses by calculating the ratio of unique unigrams or bigrams over the total number of unigrams or bigrams in the generated responses.

- **Sentence Embedding Similarity**. We use Sentence-BERT (Reimers and Gurevych, 2019) to obtain an embedding for the generated response as well as the ground-truth, and then calculate the cosine similarity between the two embeddings.

The performance of the dialogue models was tested in held-out and zero-shot settings. The evaluation results are shown in Table 3.8.

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

## 3.5 Chapter Summary

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

Obtaining a large dataset is important only if the quality can be assured. The qualitative comparison conducted between EDOS and the state-of-the-art EmpatheticDialogues dataset by means of visual validation was one way to confirm that. The results of the comparison confirmed that most of the conversational dynamics present in EmpatheticDialogues were observed in EDOS. We also proposed some experimental baselines by training a transformer model for empathetic response generation on OS, EDOS, and EmpatheticDialogues datasets and tested them in held-out and zero-shot settings. The results showed that the model fine-tuned on EDOS scored the best in terms of diversity metrics. This dataset can be readily utilized to develop empathetic conversational agents and for fine-grained emotion analysis in dialogues. The pipeline we present can be used when creating similar large-scale datasets in similar or even different domains.

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

# 4 Use of a Taxonomy of Empathetic Response Intents to Control and Interpret Empathy in Neural Chatbots

*This chapter is based on the work of Anuradha Welivita, and Pearl Pu (Welivita and Pu, 2023b). The author of this thesis (Anuradha Welivita) was mainly responsible for the development of the rule-based and neural intent prediction approaches and training and evaluation of the neural conversational models based on EmpatheticDialogues and EDOS datasets.*

## 4.1 Introduction

End-to-end neural dialogue response generation has revolutionized the design of open-domain conversational agents or chatbots due to requiring little or no manual intervention and its ability to largely generalize (Sordoni et al., 2015; Shang et al., 2015; Vinyals and Le, 2015a). It overcomes many limitations of traditional rule-based response generation techniques such as the cost of domain expertise and predictability of responses. But due to the black-box nature of these end-to-end models, they offer very little controllability to the developer and generate responses that are difficult to interpret (Xu et al., 2018; Wu et al., 2021; Gupta et al., 2021), making these approaches less reliable and fail-safe (Garcez and Lamb, 2023). A recent example is Microsoft's Taybot that started producing unintended, and offensive tweets denying the Holocaust as a result of learning from racist and offensive information on Twitter (Lee, 2016). Having control over the generated responses would have enabled the chatbot to avoid malicious intentions and carefully choose how to converse. Thus, it is important to look at ways how developers can gain control over the responses generated by end-to-end neural response generation models and how they can be made interpretable.

Recent research has taken efforts to induce controllability and interpretability into end-to-end models. For example, Xu et al. (2018) explore how the flow of human-machine interactions can be managed by introducing dialogue acts as policies to the dialogue generation model. Sankar and Ravi (2019) show that conditioning the response generation process on interpretable dialogue attributes such as dialogue acts and sentiment helps to eliminate repetitive responses and makes the model more interesting and engaging.

| | |
|---|---|
| **Dialogue context:** | |
| Speaker: | *I think that the girl of my dreams likes somebody else. I feel very sad about it.* |
| Listener: | *Ooh, am so sorry about that. Have you tried to talk to her?* |
| Speaker: | *It's tough as she has been out of the country for a month, so I will likely discuss it when she returns.* |
| **Possible responses:** | |
| (No control) | *Have you talked to her about it yet?* (Repetitive) |
| (No control) | *I don't think that's a good idea.* (Not encouraging to the speaker) |
| (Conditioned on: **Encouraging**) | *I hope everything works out for you.* (Empathetically appropriate) |

Table 4.1: An example dialogue showing how controllability affects response generation.

In contrast to task-oriented dialogue systems designed to help people complete specific tasks, open-domain chatbots are designed to engage users in human-machine conversation for entertainment and emotional companionship (Wu and Yan, 2018). Hence, in open-domain conversations, controllability should also be studied with respect to aspects such as humor, personality, emotions, and empathy, which cannot be achieved using generic dialogue acts. In this study, our focus is on controlling empathy in open-domain chatbot responses, which requires understanding conversational strategies used in human-human empathetic conversations.

Earlier studies gain control in this aspect by conditioning the response on either manually specified (Zhou et al., 2018a; Zhou and Wang, 2018; Hu et al., 2018; Song et al., 2019) or automatically predicted (Chen et al., 2019b) sentiment or emotion labels. However, our analysis of human-human conversations of the EmpatheticDialogues dataset (Rashkin et al., 2019) revealed that listeners are much more likely to respond to positive or negative emotions with specific empathetic intents such as *acknowledgment*, *consolation* and *encouragement*, rather than expressing similar or opposite emotions. These observations resulted in the development of a taxonomy of empathetic response intents. In this chapter, we explore how end-to-end response generation can be combined with more advanced control of empathy by utilizing this taxonomy of empathetic response intents in addition to existing emotion categories. To provide a glimpse of what we aim to achieve, in Table 4.1 we show how conditioning the response on an empathetic response intent chosen based on the dialogue history can serve in producing a more empathetically appropriate response. It avoids repetitive or sub-optimal responses generated by end-to-end approaches without any control.

Our empathetic response generation model consists of two modules: 1) a response emotion or intent prediction module; and 2) a response generation module. We experiment with both rule-based and neural approaches for predicting the next response's emotion or intent. For the rule-based approaches for predicting the response emotion/intent, we develop two decision tree-based response emotion and intent prediction methods. For the neural approach for predicting the response emotion/intent, we develop a classifier based on the BERT transformer architecture (Vaswani et al., 2017a; Devlin et al., 2019). The reason why we evaluate the performance of rule-based approaches is that they are much simpler than neural models and save a lot of training time and resources. Thus, if considerable performance can still be achieved through rule-based approaches compared to the baselines, it is worth considering the use of such simpler approaches over sophisticated neural approaches, especially in resource-limited environments. The emotions and intents predicted by these methods are then used to condition the responses generated by the response generation module. For training and evaluating these models, we use two state-of-the-art dialogue datasets containing empathetic conversations: 1) the EmpatheticDialogues dataset (Rashkin et al., 2019); and 2) the EDOS (Emotional Dialogues in OpenSubtitles) dataset developed in Chapter 3. The automatic and human evaluation results confirm the importance of the use of the taxonomy in generating more diverse and empathetically more appropriate responses than end-to-end models.

Our contributions in this chapter are three folds. 1) We explore the ability of the taxonomy of empathetic response intents in controlling and interpreting the responses generated by open-domain conversational agents for emotional prompts. 2) We propose an empathetic response generation model consisting of a response emotion/intent prediction module and a response generation module to generate empathetic responses in a controllable and interpretable manner. 3) We experiment with both rule-based and neural approaches in predicting the next response's emotion or intent and evaluate their performance in conditional generation of empathetic responses using automatic and human evaluation metrics against standard baselines.

## 4.2 Related Work

### 4.2.1 Controllable Response Generation

Recent research has focused on methods to control and interpret the responses generated by open-domain neural conversational agents. Mainly we find three methods they use to control the generated response: 1) by a manually specified value (Zhou et al., 2018a; Zhou and Wang, 2018; Hu et al., 2018; Song et al., 2019); 3) by rules that are predefined or derived from the training data (Hedayatnia et al., 2020); 3) by an automatically predicted value from a neural network model (Xu et al., 2018; Sankar and Ravi, 2019; Santhanam et al., 2020; Ke et al., 2018; Lee et al., 2020). In studies addressing emotional response generation, a manually specified sentiment, emotion (Zhou et al., 2018a) or an emoji (Zhou and Wang, 2018) was used to control the sentiment or emotionality of the responses generated. Later, more and more research focused on automatically predicting values or deriving rules such that they could be used to control the generated response without manual intervention. For example, Sankar and Ravi (2019) used an RNN based policy network to predict the next dialogue act given previous dialogue turns and dialogue attributes. Hedayatnia et al. (2020) used rules designed as a set of dialogue act transitions from common examples in the Topical-Chat corpus (Gopalakrishnan et al., 2019) to plan the content and style of target responses.

But all the above work focused on achieving controllability using generic dialogue acts or generating controlled emotional responses conditioned on similar or opposite emotions, emojis, or sentiment tags. As discovered in Chapter 2, these labels do not suffice the controlled generation of meaningful empathetic responses because humans demonstrate a wide range of emotions and intents when regulating empathy. Previous work also lacks comparisons between rule-based and automatic conditioning methods used to control response generation. In this work, we address the above gaps by investigating how empathy in neural responses can be controlled using a taxonomy of eight empathetic response intents, in addition to 32 emotion categories, while evaluating the applicability of both rule-based and automatic control mechanisms for this task.

### 4.2.2 Evaluation Methodologies

Various automatic as well as human evaluation metrics are used to evaluate open-domain dialogue response generation. The same metrics can be applied to evaluate the quality of the responses generated by controllable and interpretable dialogue response generation models. Perplexity is a popular automatic evaluation metric, which is model dependent and measures how well a probability model predicts a given sample. This is usually computed by exponentiation the cross-entropy loss. A lower perplexity means, better capability of generating the ground-truth response. Other automatic metrics that exist to evaluate how appropriate the responses are to a given context, falls mainly into three categories: **1) Word-overlap based metrics**; **2) Embedding-based metrics**; and **3) Diversity-based metrics**. Word overlap-based metrics measure the amount of word overlap between the proposed response and the ground-truth response. The BLEU (Papineni et al., 2002) Score, METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) are three popular metrics used to evaluate this. But they have been mainly used in evaluating machine translation responses and have been shown to poorly correlate with human judgments (Liu et al., 2016). Embedding-based metrics take into account the meaning of each word as defined by a word embedding and approximate sentence-level embedding using them. The sentence-level embeddings between the candidate and the target response is then compared using a measure such as the cosine distance. The most common approaches for embedding-based comparison are Greedy Matching (Rus and Lintean, 2012), Embedding Average (Foltz et al., 1998; Landauer and Dumais, 1997; Mitchell and Lapata, 2008), and Vector Extrema (Forgues et al., 2014). Diversity-based metrics such as Distinct-1 and Distinct-2 scores (Li et al., 2015) measure the diversity of the generated responses by calculating the ratio of unique unigrams or bigrams over the total number of unigrams or bigrams in the generated responses.

In addition, human evaluation is widely adopted to measure the appropriateness of a generated response, since most of the above approaches measure how close a generated response is to the ground-truth, but in reality there are many possible responses that could be given to a certain context, which can be equally appropriate. Existing work mostly uses A/B testing or ranking model-generated responses according to how appropriate the response is to the given context. In our experiments, we mainly measure the empathetic appropriateness of the generated dialogue responses by recruiting crowd workers. They are shown the dialogue context preceding the generated response and asked to evaluate how empathetically appropriate the generated response is to the given context on a scale of *Good*, *Okay*, and *Bad*. Since there is no standard way described in the literature to evaluate the empathetic appropriateness of open-domain dialogue responses, the workers are guided by a tutorial containing examples of responses that are *Good*, *Okay*, and *Bad* in terms of empathetic appropriateness. However, this type of evaluation can still have some bias since workers come from diverse cultures and backgrounds whose understanding and ways of expressing empathy can be different.

## 4.3 Methodology

Our controllable and interpretable empathetic response generation architecture consists of two modules: 1) the response emotion/intent prediction module; and 2) the response generation module. The emotion or intent predicted by the first module is input into the second to condition the response generated by the second module. In the following sections we discuss the datasets used for our experiments, the different rule-based and automatic emotion/intent prediction methods we propose, how the emotions and intents predicted by these modules are used to generate responses that are both controllable and interpretable, and the different evaluation methods we utilize to compare the performance of these approaches on two state-of-the-art dialogue datasets containing emotional dialogue prompts.

### 4.3.1 Datasets

We utilized the EmpatheticDialogues (ED) dataset proposed by Rashkin et al. (2019), and the cleaned OS (OpenSubtitles) and EDOS dialogue datasets developed in Chapter 3 to train and evaluate our models. Even though the speakers' and the listeners' turns in the OS and EDOS datasets are not clearly defined, we assumed the odd-numbered turns (1, 3, 5, ...) as speaker turns and even-numbered turns (2, 4, 6, ...) as listener turns for our experiments. We used the OS dialogues dataset containing ≈3M dialogues for pre-training and the ED and EDOS datasets to separately fine-tune the models. A summary of the statistics of these datasets are denoted in Table 4.2. From each dataset, 80% of the data was used for training, 10% for validation, and the remaining 10% for testing.

We used the BERT (Devlin et al., 2019) transformer-based dialogue emotion classifier developed in 3 to automatically annotate all dialogue turns in the above datasets. The classifier was trained on 25K situation descriptions from EmpatheticDialogues labeled with 32 emotion classes, 7K EmpatheticDialogues listener turns labeled with eight empathetic response intents

| Dataset | Dialogues | Turns | Turns/dialogue |
|---|---|---|---|
| OS | 2,989,774 | 11,511,060 | 3.85 |
| ED | 24,847 | 107,217 | 4.32 |
| EDOS | 1,000,000 | 2,940,629 | 2.94 |

Table 4.2: Statistics of the datasets used for training and evaluating the models.

and *Neutral*, and 14K emotion and intent annotated dialogue turns from the EDOS dataset. It demonstrated a final annotation accuracy of 65.88% over 41 labels, which is significant compared to the other state-of-the-art dialogue emotion classifiers. We use the emotion and intent labels suggested by this classifier as ground-truth labels for our experiments.

### 4.3.2 Response Emotion/Intent Prediction

To generate controlled and interpretable empathetic responses, we utilized 32 fine-grained emotions on which the dialogue situations in the EmpatheticDialogues dataset are based, the most frequent eight intents from the taxonomy of listener-specific empathetic response intents developed in Chapter 2, and the *Neutral* category. A summary of the intents that we utilized for this work is denoted in Table 4.3 along with corresponding examples. To predict the emotion or intent of the next response, we propose several rule-based and neural response emotion/intent prediction methods, which are described in the following subsections.

| Empathetic intent | Example response |
|---|---|
| 1. Questioning | *What's the matter? What's wrong?* |
| 2. Agreeing | *Exactly, I get that entirely!* |
| 3. Acknowledging | *Sounds awesome!* |
| 4. Encouraging | *Just give it a trial.* |
| 5. Consoling | *I hope everything works out for you.* |
| 6. Sympathizing | *I am sorry to hear that.* |
| 7. Wishing | *Congrats, that's a step forward.* |
| 8. Suggesting | *Maybe you should talk to her.* |

Table 4.3: The taxonomy of listener specific empathetic response intents used to achieve controllability and interpretability in the responses generated.

#### Baselines

The first and the most basic baseline that we used when generating responses is the plain end-to-end transformer model proposed by Vinyals et al. (2017a), in which no conditioning is used when generating the response. As a second baseline, we sample a response emotion or intent from the set of eight empathetic response intents plus the most recent emotion encountered in the last $k(k = 3)$ dialogue turns to condition next dialogue response on. This is based on our observations in Chapter 2 on the EmpatheticDialogues dataset. We observed that in human empathetic conversations, the listener's response to emotional prompts mostly contains an empathetic response intent identified by our taxonomy of empathetic intents or a statement with similar emotion. This baseline is inspired by the work of Hedayatnia et al. (2020), in which the response dialogue act is chosen among the most frequently seen dialogue acts based on an equal probability distribution.

### Rule-based Decision Tree Approaches

We propose two non-neural, decision tree-based response intent prediction methods that leverage the knowledge of the emotion-intent flow of the dialogues in the training dataset. The basic idea of a decision tree for this context is denoted along with an example in Figure 4.1. The probabilities of emotions and intents in the branches in the decision tree are learned from the training data itself by traversing through dialogues using a window of size $k$, where $k$ is the maximum depth of the decision tree. The window is moved forward two dialogue turns at a time capturing the probability of speaker-listener emotion-intent exchanges in the training dataset.

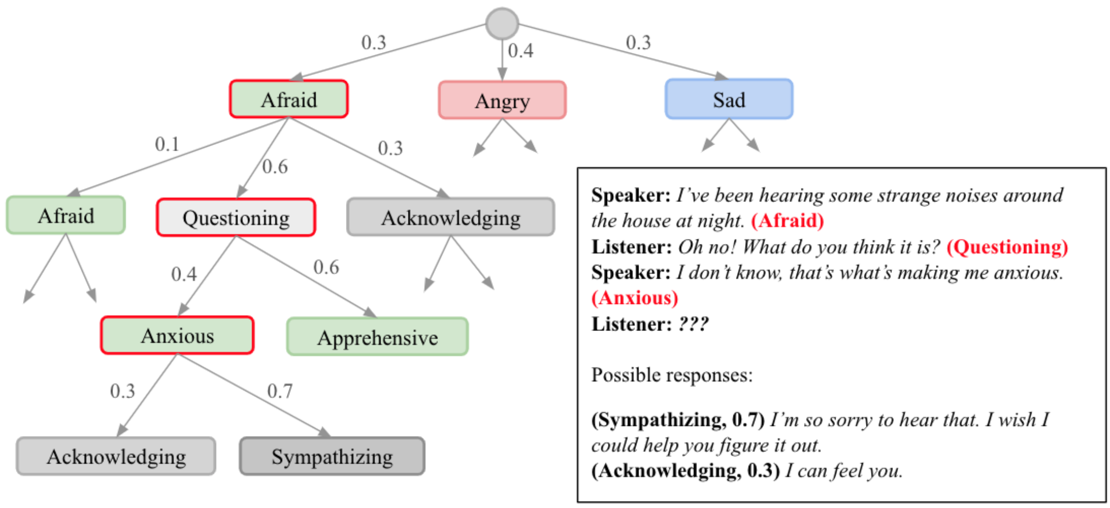



Figure 4.1: Visualization of a simpler version of our decision tree approach to predict the response emotion or intent.

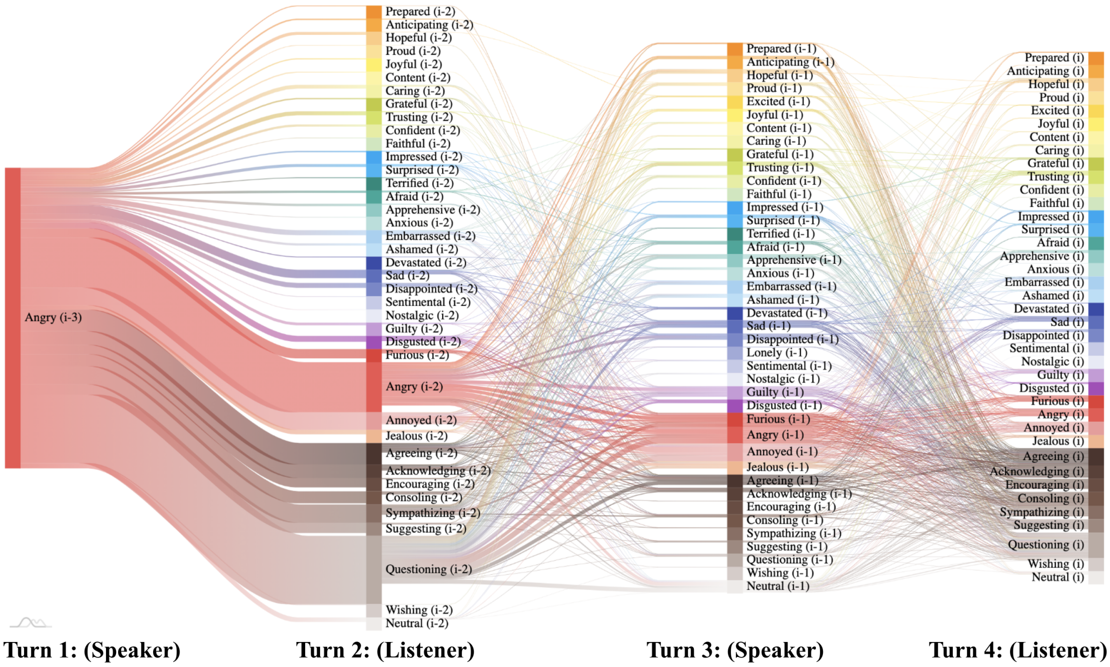
Here, we used a window of size 4 mainly because most dialogues contained in the ED, OS, and EDOS datasets were limited to four dialogue turns. During inference, an emotion or an intent is sampled based on the sequence of emotions and intents in the previous $(k-1)$ dialogue turns. We used two different methods: **1) argmax**; and **2) probabilistic sampling**, to sample the response emotion or intent from the decision tree. In the argmax method, we chose the emotion or intent with the highest probability in the decision tree based on the sequence of emotions and intents in the previous $(k-1)$ dialogue turns. In the probabilistic sampling method, we sampled an emotion or an intent based on the distribution of probabilities in the decision tree given the sequence of emotions and intents in the previous $(k-1)$ dialogue turns. We refer to these two decision tree-based methods as **DT (argmax)**, and **DT (prob. sampled)**.

We have more control over the above methods than neural response intent prediction methods since we can foresee where the dialogue will be directed by visualizing the decision trees beforehand. For example, the decision trees generated using the EmpatheticDialogues and EDOS training datasets when the emotion of the beginning dialogue prompt is *Angry* are denoted in Figure 4.2. As it could be observed, in the ED dataset, the listeners mostly respond to speakers' emotions with one of the intents from the taxonomy of empathetic response intents. The EDOS dataset by nature is more dramatic, in which both the speaker and the listener become emotional. This phenomenon is called "emotional contagion" in the psychological literature (Hatfield et al., 1993). For example in EDOS, if the speaker is angry, the listener also tends to reply back with anger. These communication patterns could clearly be visualized with the decision trees created and the developer can predict beforehand how the chatbots whose responses are conditioned on these emotion-intent patterns would behave for a given emotional prompt.

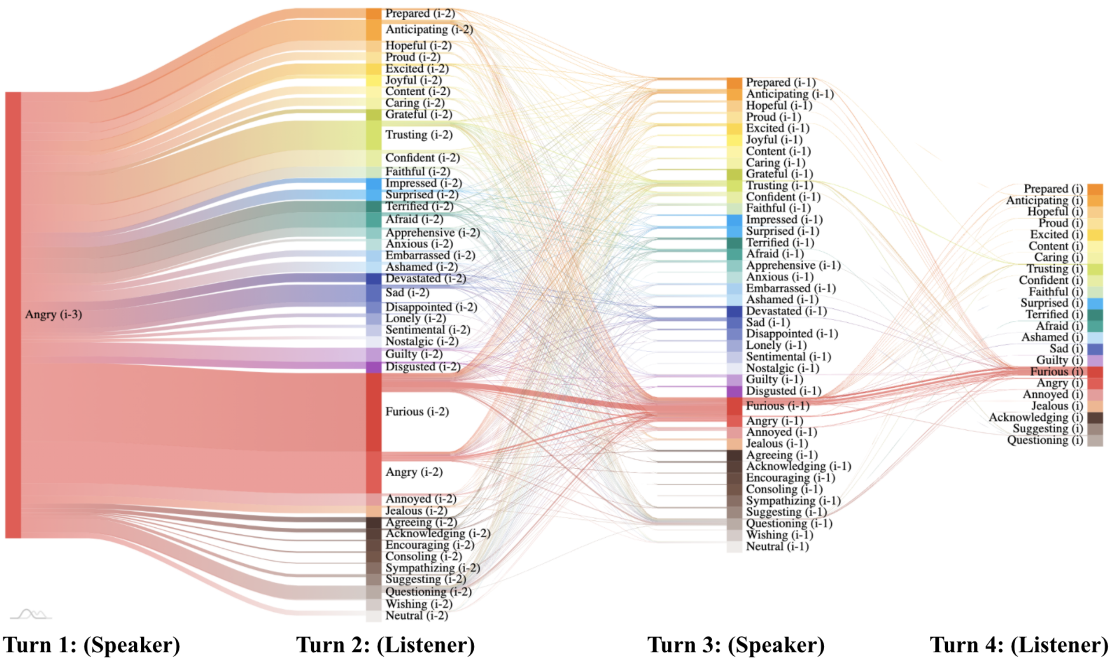
#### Neural Response Emotion and Intent Predictor

An automatic way of predicting the next response's emotion or intent is using a neural network-based response emotion/intent predictor. An advantage of using neural approaches to determine the emotion or intent of the next response is that they can leverage clues from the semantic content of the previous dialogue turns in addition to the flow of emotions and intents when predicting the response emotion or intent. Our neural response emotion/intent predictor consists of a BERT transformer-based encoder architecture (representation network) followed by an attention layer for aggregating individual token representations, a hidden layer, and a softmax as depicted in Figure 4.3. The BERT-base architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters is used as the representation network. It is initialized with weights from the pre-trained language model RoBERTa (Zhuang et al., 2021).

We concatenate the previous $k$ dialogue turns as depicted in Figure 4.3 and they are input to the encoder of the model. The emotions and intents corresponding to these $k$ dialogue turns are added to the word embeddings and positional embeddings in the original transformer architecture. This additional knowledge helps the model to get a better understanding of the flow of emotions and intents in the previous dialogue turns. The emotions and intents are embedded into a vector space having the same dimensionality as the word and position embeddings so they can add up. In addition, we also incorporate segment embeddings that differentiate between speaker and listener turns. We pre-trained the model on the OS dialogues dataset and fine-tuned it separately on ED and EDOS datasets. The hyper-parameters used during training and other training details are described in Appendix A.8.




(a) Dataset: EmpatheticDialogues




(b) Dataset: EDOS

Figure 4.2: Decision trees generated using the EmpatheticDialogues and EDOS training datasets when the emotion of the beginning dialogue prompt is *Angry*.

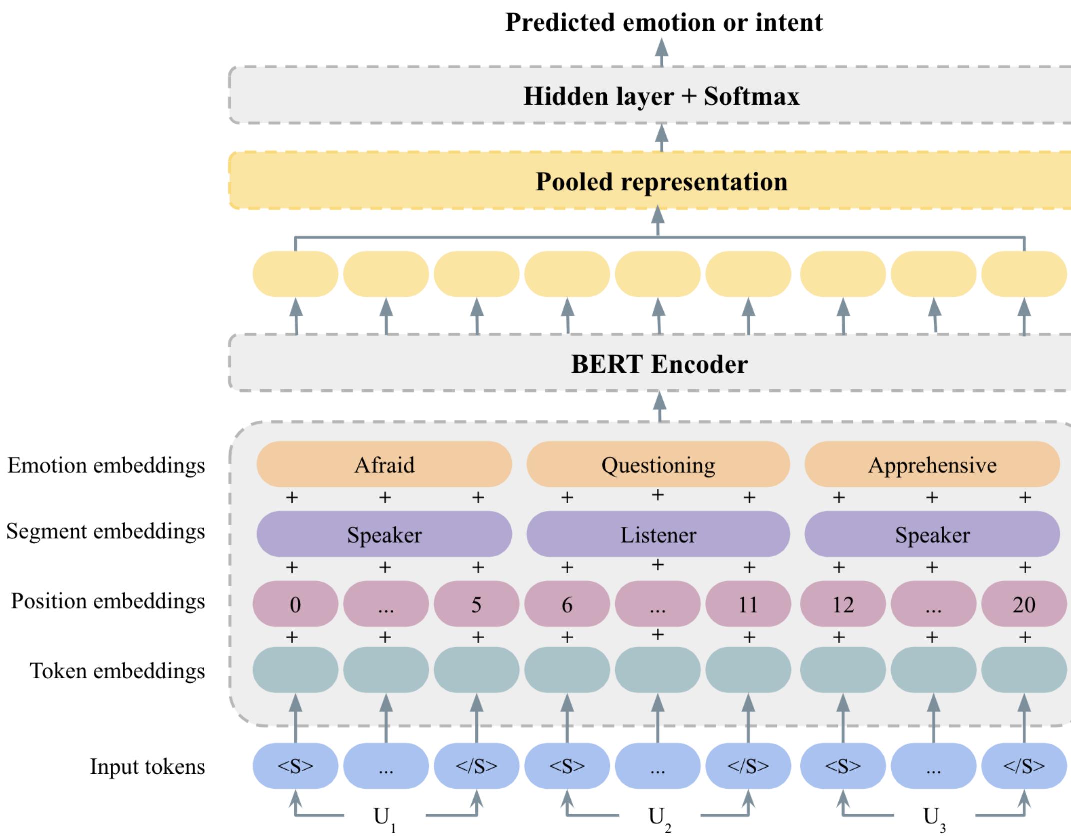



Figure 4.3: Architecture of the neural response emotion/intent predictor.

### 4.3.3 Response Generation

To generate controlled empathetic responses, we incorporated the different response emotion/intent prediction methods described above as input to the decoder. Figure 4.4 shows the high-level architecture of our models. The input representation for the encoder of the generation model is the same as the input representation used for the neural response emotion/intent predictor described in section 4.3.2. The vector representation generated by the encoder is input into the decoder along with the embedding of the emotion or intent predicted by the response emotion/intent predictor. During training, instead of the predicted emotion or intent, we used the ground-truth emotion or intent. The generation model is first pre-trained on OS dialogues and then fine-tuned on ED and EDOS datasets separately.

## 4.4 Evaluation and Results

### 4.4.1 Automatic Evaluation Results

Evaluation by means of automatic metrics was carried out separately for response emotion/intent prediction and conditional response generation. The following subsections describe the

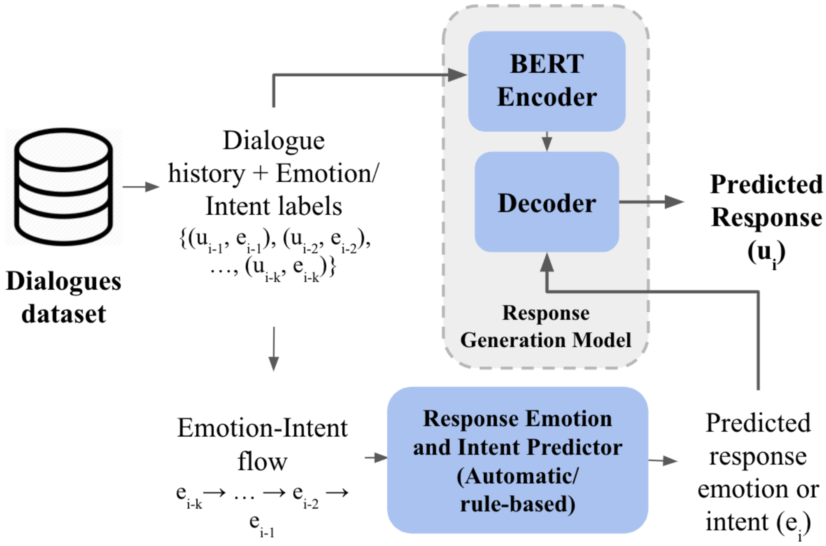



Figure 4.4: Overall architecture of the controllable and interpretable empathetic response generation model.

results obtained in these evaluations.

### Prediction Performance

The weighted precision, recall, F1, and balanced accuracy scores computed for different response emotion/intent prediction methods across ED and EDOS testing datasets are indicated in Table 4.4. According to the weighted precision, recall, F1, and accuracy scores, the neural emotion/intent predictor performed the best compared to other prediction methods. Among rule-based approaches for response emotion/intent prediction, the DT (argmax) method performed the best. The DT (argmax) method had considerable improvement in recall, F1, and accuracy scores over the equally sampled baseline.

| Model | Trained on: OS + ED<br>Tested on: ED | | | | Trained on: OS + EDOS<br>Tested on: EDOS | | | |
|---|---|---|---|---|---|---|---|---|
| | Prec. | Recall | F1 | Acc. | Prec. | Recall | F1 | Acc. |
| Equally sampled | 0.1138 | 0.0667 | 0.0638 | 0.0410 | 0.0981 | 0.0221 | 0.0232 | 0.0285 |
| DT (argmax) | 0.0959 | 0.0883 | 0.0883 | 0.0692 | 0.0755 | 0.1016 | 0.0799 | 0.0419 |
| DT (prob. sampled) | 0.0715 | 0.0663 | 0.0680 | 0.0480 | 0.0627 | 0.0616 | 0.0619 | 0.0345 |
| Neural predictor | 0.1634 | 0.1636 | 0.1472 | 0.1163 | 0.1306 | 0.1712 | 0.1181 | 0.0679 |

Table 4.4: Weighted precision, recall, F1 and accuracy scores computed for ED and EDOS test datasets. The cells in dark green indicate the best scores and the cells in light green indicate the second best scores.

### Generation Performance

To evaluate the performance of response generation, we computed the perplexity, diversity metrics (distinct unigram and distinct bigram scores) (Li et al., 2016a), and vector extrema cosine similarity (Forgues et al., 2014) on ED and EDOS testing datasets. They are denoted in Table 4.5. We also evaluated the responses generated by a model conditioned on the ground-truth emotion or intent of the next response to see how well the taxonomy of empathetic response intents alone contributes to better empathetic response generation performance.

| Model | Trained on: OS + ED<br>Tested on: ED | | | | Trained on: OS + EDOS<br>Tested on: EDOS | | | |
|---|---|---|---|---|---|---|---|---|
| | PPL | D-1 | D-2 | Vector extrema | PPL | D-1 | D-2 | Vector extrema |
| GT emotion/intent | 11.74 | 0.0823 | 0.2812 | 0.5181 | 12.57 | 0.0846 | 0.2552 | 0.4539 |
| End-to-end model | 12.26 | 0.0544 | 0.1612 | 0.5015 | 13.13 | 0.0784 | 0.228 | 0.4365 |
| Equally sampled | 13.48 | 0.0761 | 0.2469 | 0.4824 | 14.20 | 0.0754 | 0.2229 | 0.433 |
| DT (argmax) | 13.23 | 0.0865 | 0.2977 | 0.4892 | 14.14 | 0.0727 | 0.2419 | 0.4458 |
| DT (prob. sampled) | 13.37 | 0.0795 | 0.2761 | 0.4828 | 14.23 | 0.0763 | 0.2418 | 0.436 |
| Neural predictor | 13.15 | 0.0835 | 0.2811 | 0.4851 | 13.97 | 0.0805 | 0.2415 | 0.4403 |

Table 4.5: Perplexity (PPL), diversity metrics (distinct unigrams: D-1; and distinct bigrams: D-2), and vector extrema cosine similarity calculated on ED and EDOS testing datasets.

According to the results, the models whose response was conditioned on the ground-truth response emotion or intent performed the best in terms of perplexity and embedding extrema in both ED and EDOS datasets and in terms of diversity metrics in the EDOS dataset. These results emphasize the usefulness of the taxonomy of empathetic response intents and the 32 fine-grained emotion categories in generating controlled empathetic responses. The models incorporating the DT (argmax) approach scored the best in terms of diversity metrics in the ED test dataset.

### 4.4.2 Human Evaluation

In addition to the automatic metrics, we carefully designed a human evaluation experiment in Amazon Mechanical Turk (AMT) to evaluate responses' empathetic appropriateness. We selected a total of 1,000 dialogue cases: 500 ED and EDOS dialogues for testing. The AMT workers had to drag and drop responses generated by five models (end-to-end; models whose response was conditioned on the equally sampled baseline, DT argmax, DT prob. sampled and the neural predictor) into areas *Good*, *Okay*, and *Bad*, depending on their empathetic appropriateness. We neglected responses conditioned on the ground-truth emotion or intent since we are more interested in automatically predicted labels. We bundled 10 dialogues into a HIT (Human Intelligence Task) so that one worker works on at least 10 cases to avoid too much bias between answers. To evaluate the quality of the work generated, we included three quiz questions equally spaced in a HIT. In these, we included the ground-truth response among the other responses generated by the models. If a worker rated the ground-truth response either

| Model | Trained on: OS + ED Tested on: ED | | | | Trained on: OS + EDOS Tested on: EDOS | | | |
|---|---|---|---|---|---|---|---|---|
| | Good | Okay | (Good + Okay) | Bad | Good | Okay | (Good + Okay) | Bad |
| End-to-end model | 46.94 | 15.72 | 62.66 | 37.34 | 24.48 | 27.02 | 54.50 | 45.50 |
| Equally sampled | 25.22 | 25.22 | 50.44 | 49.57 | 19.35 | 29.84 | 49.18 | 50.82 |
| DT (argmax) | 37.61 | 33.70 | 71.31 | 28.69 | 21.70 | 36.72 | 58.42 | 41.58 |
| DT (prob. sampled) | 24.17 | 21.68 | 45.85 | 42.13 | 21.96 | 31.78 | 53.74 | 46.26 |
| Neural predictor | 50.00 | 17.39 | 67.39 | 32.61 | 24.82 | 33.01 | 57.83 | 42.17 |

Table 4.6: Human evaluation results (as a percentage) corresponding to ED and EDOS testing datasets.

as *Good* or *Okay*, then a bonus point was added. To encourage attentiveness to the task, for those who obtained at least two out of three quiz questions correct, we gave a bonus of 0.1$. Three workers were allowed to work on a HIT and only the ratings that were agreed by at least two workers, both who have obtained bonuses, were taken to compute the final scores. As a result, 8.33% of the answers were disqualified. The results of the experiment are denoted in Table 4.6. The experiment yielded an inter-rater agreement (Fleiss' kappa) score of 0.2294 indicating fair agreement. Additional details pertaining to the human evaluation experiment are denoted in Appendix A.9

According to the results, the neural predictor scored the highest percentage of *Good* ratings in both ED and EDOS testing datasets. The models that use the equally sampled approach performed the worst producing the highest percentage of responses ranked *Bad*. An interesting observation is that the DT (argmax) method scored the most number of combined *Good* and *Okay* responses in ED and EDOS testing datasets confirming that rule-baled approaches such as the decision tree approach we propose could be used to control and interpret the responses without losing significant accuracy.

### 4.4.3 Case Study

In Table 4.7 we show some example responses generated by different models for a given dialogue context. It could be noticed that having no response control mechanism and having a response conditioned on an inappropriate intent both can result in responses that are empathetically inappropriate with the dialogue context. The neural predictor, as well as the decision tree-based mechanisms, generate some emotion or intent that is appropriate to the dialogue context, enabling the generation model to generate responses that are more empathetically appropriate, guiding the conversation in a meaningful direction.

| **Dialogue context:** | | | |
|---|---|---|---|
| Speaker: *Last week when Toys R Us closed it really made me start thinking of the 90s and my childhood* **(Nostalgic)** | | | |
| Listener: *I've heard a lot of people say something similar. Did you go there a lot?* **(Questioning)** | | | |
| Speaker: *I did! I was talking to my husband about the different toys I had. Even started talking about the music and shows from that era.* **(Nostalgic)** | | | |
| Model | Predicted emotion/intent | Generated response | Human eval. |
| Ground-truth emotion/intent | Jealous | *That sounds like a lot of fun. I wish I had a husband like that.* | Good |
| End-to-end model | - | *I'm sorry to hear that. What kind of toys were you talking about?* | Bad |
| Equally sampled | Wishing | *I wish you the best of luck!* | Bad |
| DT (argmax) | Questioning | *What kind of toys were you talking about?* | Good |
| DT (prob. sampled) | Encouraging | *I bet that was fun* | Good |
| Neural predictor | Encouraging | *I bet that was fun* | Good |

Table 4.7: An example dialogue showing that both lack of controllability and conditioning the response on an inappropriate emotion or intent can lead to responses that are empathetically inappropriate with the dialogue context.

## 4.5 Chapter Summary

This chapter investigated the use of a taxonomy of empathetic response intents along with 32 fine-grained emotions in controlling and interpreting the responses generated by open-domain conversational agents for emotional prompts. In this regard, several rule-based and automatic response control methods were proposed and were compared in terms of their prediction and generation performance on two state-of-the-art dialogue datasets containing emotional dialogues. It was observed that the neural response emotion/intent predictor we proposed outperformed the rest including the end-to-end model in terms of evaluation metrics related to both prediction and generation. This implies the importance of leveraging semantic clues in addition to the flow of emotions and intents in the previous turns when predicting the next response's emotion or intent. However, there are some disadvantages to using this approach: 1) developers cannot foresee the label that the model would predict next; and 2) cost of time and resources spent for training the model. As a remedy, we proposed two decision tree-based response emotion/intent prediction approaches. Across evaluation metrics for prediction and generation, the performance of the decision-tree methods was considerably better than the end-to-end approach and the equally sampled baseline. The decision tree (argmax) method performed the best in terms of diversity metrics related to response generation. In the human evaluation stage, we saw that the DT (argmax) method produced the most number of combined *Good* and *Okay* responses in ED and EDOS test datasets, pointing to the fact that the rule-based approaches we proposed can still be used without a significant degrade in performance in resource-limited environments. On the whole, the results of this study inform developers about the utility of the taxonomy of empathetic response intents in controlling the responses generated by open-domain chatbots and which optimal methodology to use (rule-based or automatic conditioning) based on the operational environment.

# 5 HEAL: A Knowledge Graph for Distress Management Conversations

*This chapter is based on the work of Anuradha Welivita, and Pearl Pu (Welivita and Pu, 2022b). The author of this thesis (Anuradha Welivita) was mainly responsible for the summarization of the distress narratives, clustering, development of the knowledge graph, HEAL, implementation of the retrieval-based chatbot, and evaluating it against the state-of-the-art empathetic chatbots.*

## 5.1 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 (Xie, 2017; Fitzpatrick et al., 2017; Inkster et al., 2018; Ghandeharioun et al., 2019; Mousavi et al., 2021).

With the development of sophisticated neural network architectures such as the transformer (Vaswani et al., 2017a) and pre-trained language models such as BERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021) and GPT-3 (Brown et al., 2020a), fine-tuning neural response generation models on unstructured text has become one of the common approaches to build chatbots. But as it was seen in Chapter 4, the lack of controllability and the black-box nature make these models less reliable and fail-safe. Also, generic and often repetitive responses produced by these models make the interaction quite less engaging and uninformative. This is especially problematic when the user is undergoing a distressful situation where he is sensitive to inappropriate comments. 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., 2018b; Young et al., 2018) 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., 2019).

Though large-scale knowledge graphs such as ConceptNet (Speer et al., 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 in the Reddit (www.reddit.com) online peer support platform.

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 5.1. The rationale behind selecting different types of nodes includes the identification of topics related to distress and the corresponding response types leading to positive feedback from the speaker. 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. The identification of speaker's expectations and the connections between speaker's expectations and responses can facilitate models in generating more targeted responses addressing specific questions raised in the dialogue prompt. The affective states that we identify include the 32 fine-grained emotions, the 8 main empathetic response types identified in Chapter 2, and the *Neutral* category, which are derived by analysing empathetic human-human conversations in the EmpatheticDialogues dataset (Rashkin et al., 2019). These states can further narrow down the selection of particular types of responses given particular types of affective states in the dialogue prompt following the controllable and interpretable response generation architecture introduced in Chapter 4.

As depicted in Figure 5.2, 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.

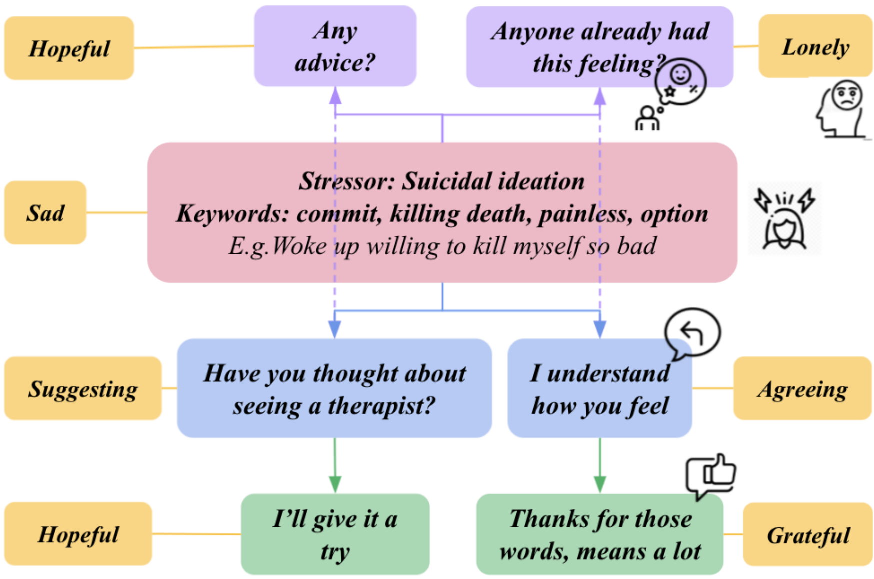



Figure 5.1: An illustration of part of *HEAL*. The red, purple, blue, green, and yellow nodes represent the stressors, speaker expectations, response and feedback types, and associated affective states respectively.

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 (2020); 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.

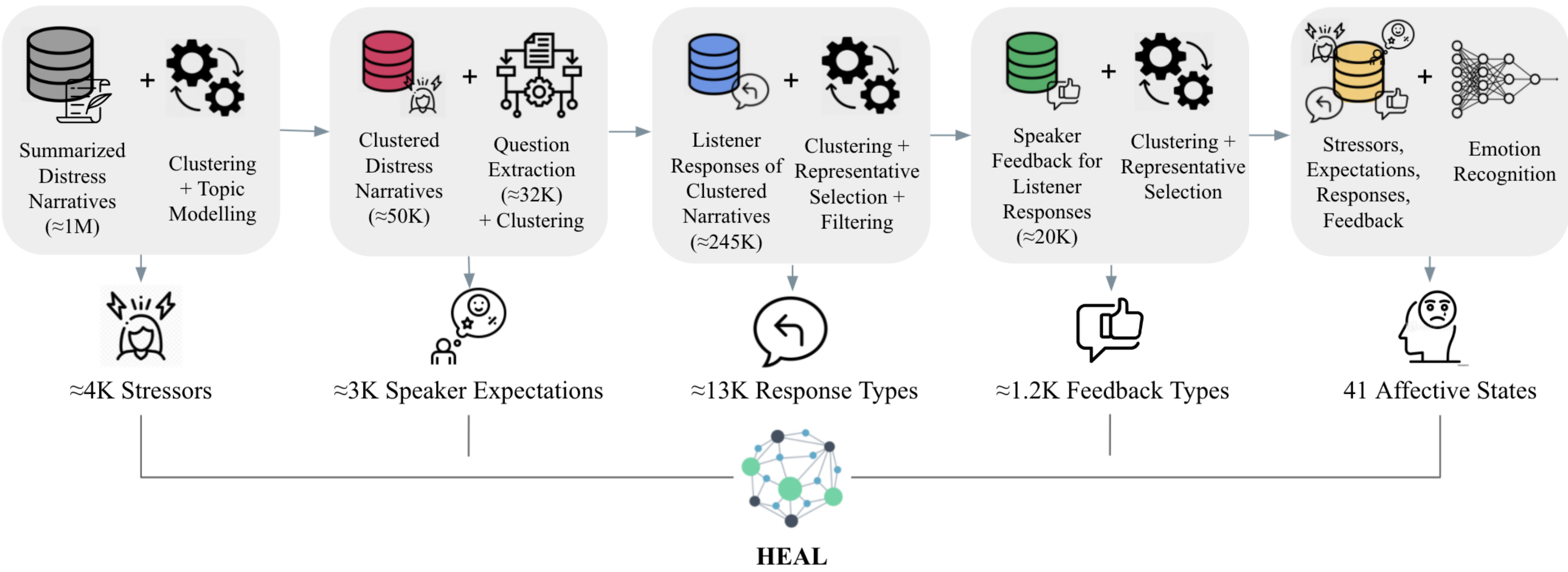



Figure 5.2: Step-by-step process for developing the knowledge graph, HEAL.

## 5.2 Related Work

Knowledge graphs have attracted the attention of the natural language processing community due to their usefulness 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 (Sankar and Ravi, 2019). 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 et al., 2017), ATOMIC (Sap et al., 2019), and ASER (Zhang et al., 2020a) 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., 2018b), 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.

### 5.2.1 Reddit Emotional Distress (RED) Dialogue Dataset

Publicly available emotional dialogue datasets such as EmpatheticDialogues (Rashkin et al., 2019), EmotionLines (Hsu et al., 2018) and EmoContext (Chatterjee et al., 2019b), 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 et al., 2016; Zhang and Danescu-Niculescu-Mizil, 2020) are not directly accessible due to ethical reasons. Thus, to develop the knowledge graph, we curated a new dataset from the online peer support platform, Reddit, containing dialogues that discuss real-world distressful situations. We call this dataset the RED (**R**eddit **E**motional **D**istress) dialogue dataset.

Online peer support platforms such as Reddit and TalkLife (www.talklife.com) encourage open discussion of often stigmatized psychological concerns and personal distress (De Choudhury and De, 2014; Sharma et al., 2017). They provide alternative means for connection and support when other means of care are less accessible. The anonymity in such platforms facilitates self-disclosure and such discussions help people to feel more supported and less stressed in times of crisis (De Choudhury and De, 2014; Smith-Merry et al., 2019). Reddit is one such platform, which ranks among the most visited websites in the world (Sharma et al., 2017). Reddit users can create community forums called “subreddits" to discuss and support each other on a breadth of topics. Reddit policies also allow researchers to scrape its data and use them for research. Since many people interact in Reddit in a day-to-day basis, the distress-related topics it covers are abundant and have a wide variety. Because of these reasons we chose Reddit to curate conversations that provide support for people in distress.

For this purpose, we choose 8 subreddits: *depression*; *depressed*; *Off My Chest*; *SuicideWatch*; *Depression Help*; *sad*; *Anxiety Help*; and *Mental Health Support*, where such conversations were abundantly present. We used the Pushshift API (Baumgartner et al., 2020) to scrape English textual conversations from the above subreddits. We extracted one dyadic dialogue per conversation thread, thereby diversifying the conversation topics in the dataset. To preserve anonymity, we replaced the usernames with *speaker*, *listener_1*, ... *listener_n*. Next, we removed HTML tags and URLs from the data, and replaced numerals with a special tag <NUM>. But punctuation marks, emoticons, and emojis were preserved as they can be useful indicators to identify users' emotions. We applied *profanity-check* (Zhou et al., 2020b), a fast and robust library to detect profane language in listeners' turns and removed such sentences containing profanity. But we retained profane speaker turns as they can inform the speakers' state of mind. The resultant RED dataset contains ≈1.3 million dyadic conversations. It consists of ≈3 turns per dialogue on average. Table 5.2 displays the summary of descriptive statistics of conversations present in the dataset as well as in individual subreddits.

## 5.3 Methodology

### 5.3.1 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 distress 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*. The results are shown in Table 5.1. Accordingly, the highest percentage of summaries rated as *Good* were generated by our custom implementation of the SMMRY algorithm. Hence it was selected to summarize lengthy dialogue turns (turns with ≥ 100 tokens).

Our implementation of the SMMRY algorithm involved the following steps. We customized

| | SMMRY | BART | GPT-2 | XLNET | T5 |
|---|---|---|---|---|---|
| Good | **80%** | 55% | 30% | 15% | 0% |
| Okay | 20% | 35% | 40% | 50% | 30% |
| Bad | 0% | 10% | 30% | 35% | 70% |

Table 5.1: Percentage of summaries rated as *Good*, *Okay*, and *Bad* among randomly selected 100 distress narratives.

| Subreddit | # Dialogues | # Turns | # Tokens | Avg. # turns per dialog | Avg. # tokens per dialogue | Avg. # tokens per turn |
|---|---|---|---|---|---|---|
| r/depression | 510,035 | 1,396,044 | 106,967,833 | 2.74 | 209.73 | 76.62 |
| r/depressed | 10,892 | 23,804 | 1,940,000 | 2.19 | 178.11 | 81.50 |
| r/offmychest | 437,737 | 1,064,467 | 109,459,738 | 2.43 | 250.06 | 102.83 |
| r/sad | 18,827 | 42,293 | 3,088,562 | 2.25 | 164.05 | 73.03 |
| r/SuicideWatch | 262,469 | 791,737 | 59,267,000 | 3.02 | 225.81 | 74.86 |
| r/depression_help | 23,678 | 51,849 | 5,412,390 | 2.19 | 228.58 | 104.39 |
| r/Anxietyhelp | 8,297 | 18,351 | 1,428,287 | 2.21 | 172.14 | 77.83 |
| r/MentalHealth Support | 3,551 | 7,931 | 772,952 | 2.23 | 217.67 | 97.46 |
| **All** | **1,275,486** | **3,396,476** | **88,336,762** | **2.66** | **226.06** | **84.89** |

Table 5.2: Descriptive statistics of the conversations in the RED dataset.

the algorithm by adding the last step in order to avoid summaries beginning with sentences containing transition phrases and expressions coreferencing previously mentioned entities.

1. Tokenize and lemmatize the text.
2. Create a frequency dictionary by calculating the occurrence of each token in the text.
3. Assign each token with points depending on their popularity.
4. Split up the text into individual sentences and rank them by the sum of their tokens' points.
5. Return the K most highly ranked sentences in chronological order (K is dynamically chosen such that the returned text does not exceed 100 tokens).
6. If the first sentence returned contains coreferences that are not resolved as detected by the *neuralcoref* python library[1] or if it starts with a transition phrase, select the next top ranked sentence from the preceding sentences.

Approximately 43% of the dialogue turns were summarized using this. After the application of the SMMRY algorithm, we further filtered dialogues in which at least one turn exceeds 100 tokens (as tokenized by the Roberta tokenizer). These were likely to be resulting due to absence of clear sentence separation markers that disallow correct execution of the summarization algorithm on them.

### 5.3.2 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 agglomerative clustering algorithm is particularly useful when the number of clusters are unknown, which is true in our scenario. It differs from other clustering methods such as K-means, LDA (Blei et al., 2003) and NMF (Dhillon and Sra, 2005) with the ability to specify a threshold for the clustering—clusters above this threshold are merged. It also does not force data points to clusters as they are considered outliers and gives more flexibility to generate meaningful clusters through hyperparameter tuning.

The linkage distance between the clusters 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

[1] github.com/huggingface/neuralcoref

document-level embeddings. Among several models the authors of Sentence-BERT have proposed, we used the *distilbert-base-nli-stsb-quora-ranking* model trained for similar questions detection.

### 5.3.3 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 et al., 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 (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 Appendix A.10). 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.

To derive topics from the clusters of distress narratives based on their content, we make use of a class-based TF-IDF (c-TF-IDF) score, which generates easily interpretable topics while preserving important words in the topic descriptions. In this, instead of applying TF-IDF as usual on a set of documents, it treats all documents in a single cluster as a single document and applies TF-IDF on that, which enables ranking the most important words belonging to a topic/cluster at the top. After applying c-TF-IDF-based topic modeling on these clusters, we uncovered some clearly distinguishable stressors, which further validated the goodness of clustering. Table 5.3 shows some stressors identified in this process.

| **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 5.3: Some stessors identified in the clusters of distress narratives using TF-IDF.

### 5.3.4 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 5.4. We randomly selected a member of each cluster as the cluster representative.

| Type | Threshold | No. of clusters | Largest cluster size | Tot. no. of 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 5.4: 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 $[0, \infty)$, respectively. The more positive these values are the better.

### 5.3.5 Affective State Modelling

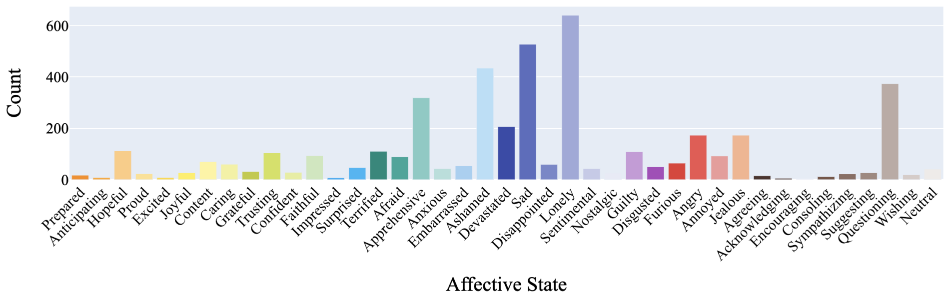
To associate each of the stressors, expectation, response and feedback clusters with an affective state, we used a BERT transformer-based classifier proposed in Chapter 2 that was trained on the EmpatheticDialogues dataset with a significant classification accuracy of 65.88%, which is comparable with 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, 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.

Table 5.5 shows some examples of frequent expectation, response, and feedback types discovered after clustering, along with corresponding percentages of occurrence and the most prominent affective states associated with the cluster.

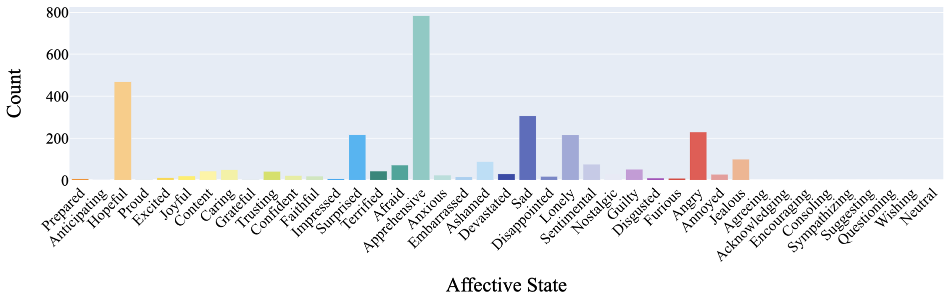
## 5.4 HEAL: Statistical Analysis

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.

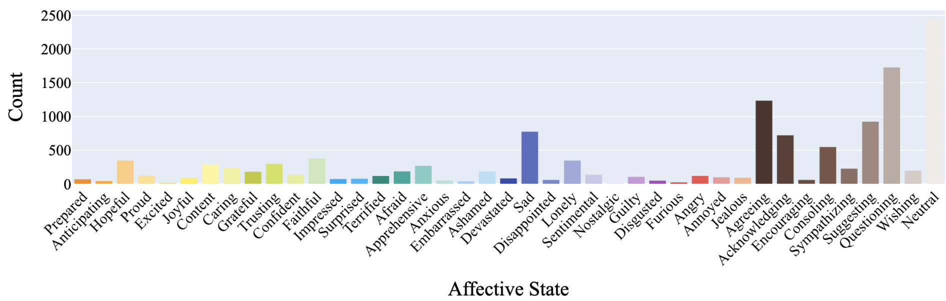
Figure 5.3 shows the distribution of affective states associated with the stressors, expectations, responses, and feedback types. According to the statistics, 73.60% of the stressors are associated with negative affective states. Out of them, emotions *Lonely*, *Sad*, *Ashamed* and *Apprehensive* are associated with 44.01% of the stressors. Most of the 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 of useful response types in HEAL that can deescalate the negative affective states of people suffering from distress.



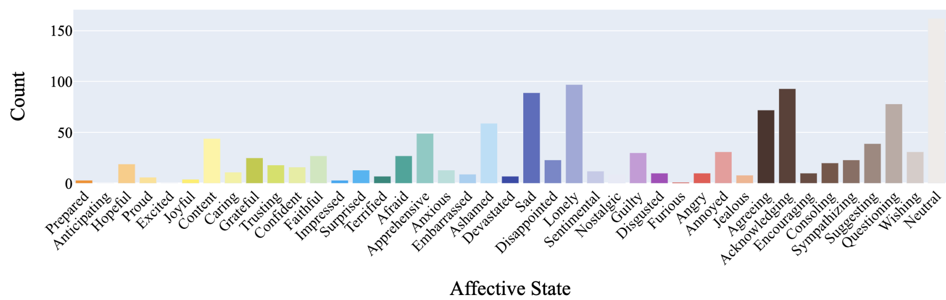

(a) Stressors




(b) Expectations




(c) Responses




(d) Feedback

Figure 5.3: Distribution affective states pertaining to stressors, expectations, responses and feedback in HEAL.

| Representative phrase | Other examples | Affective state | Occurrences % |
|---|---|---|---|
| **Expectations:** | | | |
| - *What should I do?* | - *What do I do?*<br>- *What should I do about this?* | Sad | 2.99% |
| - *Can anyone relate?* | - *Anyone relates or wants to share their thoughts?*<br>- *Does anyone else relate to this?* | Lonely | 0.48% |
| - *Any tips on how to stop?* | - *Does anyone know how to stop this?*<br>- *What do you all do to stop it?* | Hopeful | 0.21% |
| **Responses:** | | | |
| - *I'm sorry you feel like this.* | - *I'm sorry you are feeling down.*<br>- *I'm so sorry to hear that you are feeling this way.* | Neutral: Sympathizing | 1.31% |
| - *That's terrible.* | - *Yeah it's horrible.*<br>- *That's really horrible.* | Neutral: Acknowledging | 0.05% |
| - *Of course you deserve love, everyone deserves love.* | - *You are worthy of affection, love and care.*<br>- *You deserve to be happy and feel loved.* | Caring | 0.02% |
| **Feedback:** | | | |
| - *Thank you.* | - *Thank you for your reply*<br>- *Thanks I appreciate it* | Grateful | 36.16% |
| - *That makes a lot of sense.* | - *This actually makes a lot of sense!*<br>- *that definitely does make sense.* | Neutral: Acknowledging | 0.24% |
| - *In my situation I don't think therapy is the solution.* | - *I am seeking therapy, but it's not helping right now*<br>- *Because therapy can't help what lives inside of me.* | Sad | 0.12% |

Table 5.5: Examples of common expectation, response, and feedback types discovered through clustering, their associated affective states, and percentages of occurrence in the dataset.

## 5.5 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 5.4. The size of the nodes corresponds to the size of the respective clusters and the width of the 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 are 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.

## 5.6 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 10% of the dialogues from the RED dataset that were 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 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.

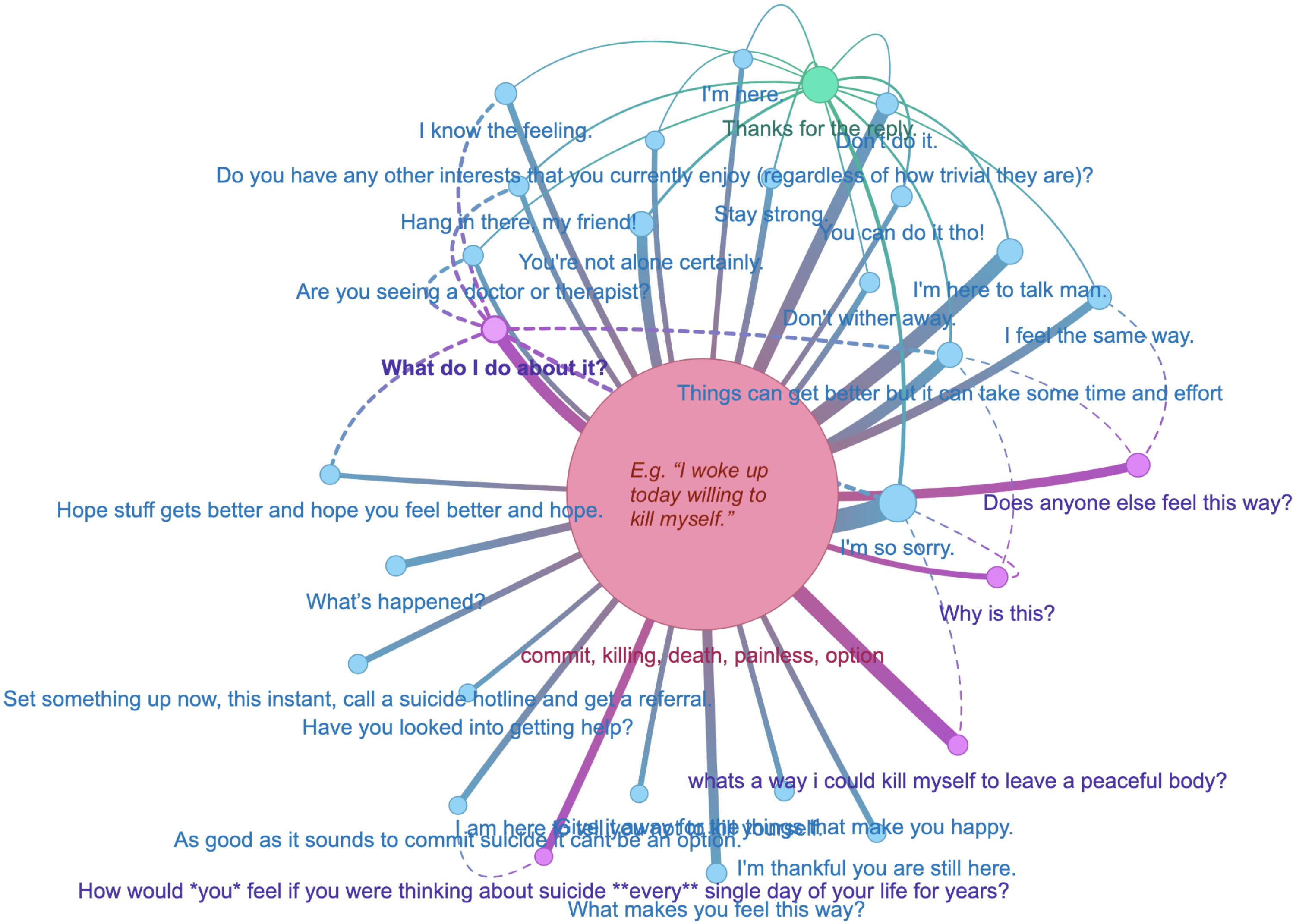



Figure 5.4: Visualization of part of HEAL by vis.js. The stressors, expectations, response and feedback types are indicated in colors red, purple, blue, and green, respectively. Only connections with significant edge weights are visualized to avoid clutter.

We compare responses retrieved by HEAL-ranked with two state-of-the-art empathetic response generation models, one developed by Xie and Pu (2020) and Blender (generative) (Roller et al., 2021). The former is a multi-turn emotionally engaging dialogue generation model based on RoBERTa (Zhuang et al., 2021). 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.

### 5.6.1 Automatic Evaluation

Table 5.6 includes the automatic metrics computed on the responses produced by the above models for RED dialogue prompts. We can observe HEAL-ranked outperforms the rest in terms of Distinct-N metrics used to measure response diversity (Li et al., 2016a). This shows the utility of HEAL in producing more diverse responses than existing neural response generation models. We justify this further in Table 5.7 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 (Examples are included in Appendix A.11). 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.

| Dataset | Model | D1 | D2 | D3 | D4 | BLEU1 | BLEU2 | METEOR | ROUGE | GM |
|---|---|---|---|---|---|---|---|---|---|---|
| Reddit | (Xie et al., 2020) | 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 5.6: 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., 2016a) and GM stands for the Greedy Matching score (Rus and Lintean, 2012).

| | |
|---|---|
| Prompt | *My brother passed away 2 years ago and i'm still so heartbroken. It still hurts so much* |
| (Xie et al., 2020) | *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 et al., 2020) | *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 5.7: Responses produced by the generative models that are repetitive in responding to different prompts (repetitive responses are marked with the same superscript symbol).

### 5.6.2 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 RED test dataset to be evaluated by the crowd-workers. The workers were instructed to drag and 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 are included in the Appendix A.12. The results of the experiment are denoted in Table 5.8. 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.

| Model | Good | Okay | Bad |
|---|---|---|---|
| (Xie et al., 2020) | 10.48% | 27.42% | 62.10% |
| Blender | 51.75% | 21.05% | 27.19% |
| Heal-ranked | **61.19%** | **31.34%** | **7.46%** |

Table 5.8: Human evaluation results for empathetic appropriateness of the responses produced by the three models.

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.

### 5.6.3 Case Study

In Table A.15, 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 are included in Appendix A.11).

| | |
|---|---|
| Prompt | *If I just knew how, I would have totally done it. Maybe tomorrow I should just jump in front of a train :(* |
| (Xie et al., 2020) | *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 et al., 2020) | *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 5.9: Examples of responses showing toxicity in responding to distress prompts (by superscript †).

## 5.7 Chapter Summary

In this chapter, 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 the 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. In the future, 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 as introducing more controllability and interpretability for those models, thereby increasing reliability. There is also 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.

# 6 Curating a Large-Scale Motivational Interviewing Dataset using Peer Support Forums

*This chapter is based on the work of Anuradha Welivita, and Pearl Pu (Welivita and Pu, 2022a). The author of this thesis (Anuradha Welivita) was mainly responsible for the design and implementation of the human annotation experiment, the analysis of the curated dataset, and the derivation of recommendations to boost distress support dialogue responses based on the analysis.*

## 6.1 Introduction

World Health Organization estimates psychological distress affects 29% of people in their lifetime (Steel et al., 2014). However, the shortage of mental health workers and the stigma associated with mental health demotivates people from actively seeking out help. As emphasized in Chapter 5, providing mental health support through AI-driven conversational agents to complement traditional therapy has become an interesting area of research. But one challenge associated with developing such agents is the lack of large-scale psycho-therapeutic conversations. They are either limited or are not available publicly due to ethical reasons.

Nowadays, with the expansion of social media, it could be observed that people use social media platforms such as Reddit to vent their distress and peers are seen to actively respond to such posts. These conversations are available in abundance and are publicly accessible through web scraping APIs. Thus, conversations scraped from such platforms are seen as an alternative to overcome the above challenge. Prior work has found that responses from peers contain higher empathic concern for posts for seeking help as many peers share similar distressful experiences (Hodges et al., 2010). But the extent to which responses from peers align with responses from trained counselors remain a major limitation. Knowing these differences can shed light on to what extent such conversational data could be used in training therapeutic chatbots and what pre-processing or rephrasing steps that one should take if they are being used for such purposes.

Though studies have been conducted independently to assess the competency of counselors and peers offering support (Pérez-Rosas et al., 2016; Klonek et al., 2015; Gaume et al., 2009; Sharma et al., 2020a; De Choudhury and De, 2014), studies that comparatively analyse the differences between them are limited. One such study was conducted by Lahnala et al. (2021), where they show that a classifier can distinguish between responses provided to help-seeking posts regarding mental health by professionals and peers. Mousavi et al. (2021) conducted a similar analysis between responses collected from psychotherapists and non-expert dialogue writers and noted linguistic variability in the two types of responses. However, all these analyses are limited to the lexical level.

To address the above gaps, we comparatively analyze responses from professional counselors and peers by annotating these responses using labels adapted from a well-established behavioral coding scheme named Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2014). The MITI code is used in psychological literature to evaluate how well a mental health practitioner responds to those seeking help with their mental health related issues. Specific response types from the MITI code have shown to increase the likelihood of positive health outcomes (Pérez-Rosas et al., 2018; Gaume et al., 2009). For this purpose, we make use of post-response pairs scraped from the CounselChat website (counselchat.com) that contains high-quality therapist responses to emotional distress related questions from individuals and dialogues curated from several mental-health related subreddits in Reddit, in which peers engage in actively to respond to those seeking help.

Annotating dialogue responses with labels from the MITI coding system is known to be very time consuming and expensive as it requires expert annotators trained in the practice of psychology and careful attention to the labelling task (Pérez-Rosas et al., 2016). This human labour is difficult to find. But the availability of crowdsourcing platforms such as UpWork [1] and Fiverr [2] have made it more accessible to find human labour that satisfy our requirements. Thus, we were able to recruit professionally trained mental health practitioners through UpWork to annotate dialogue responses with labels adapted from the MITI code. Our annotation pipeline consisted of three stages as depicted in Figure 6.1. Two workers were involved in the first stage and high-quality workers who scored high observed agreements with a peer in the first stage acted as judges to resolve conflicting labels in the second and third stages contributing to increased observed agreement and inter-rater reliability.

Using these annotations, we conducted a comparative analysis between responses from peers and counselors to identify to what extent they align with each other. Based on these findings, we recommend ways of boosting peers' responses to match as close as possible to counselors' responses. The recommendations made in this paper can contribute to improve the perceived therapeutic effectiveness of a chatbot trained on data from peer support forums.

Our contributions are threefold. 1) We develop an MI dataset having client-counselor and peer-peer dialogues, in which ≈17K listeners' utterances are annotated with labels adapted

[1] http://upwork.com
[2] https://www.fiverr.com

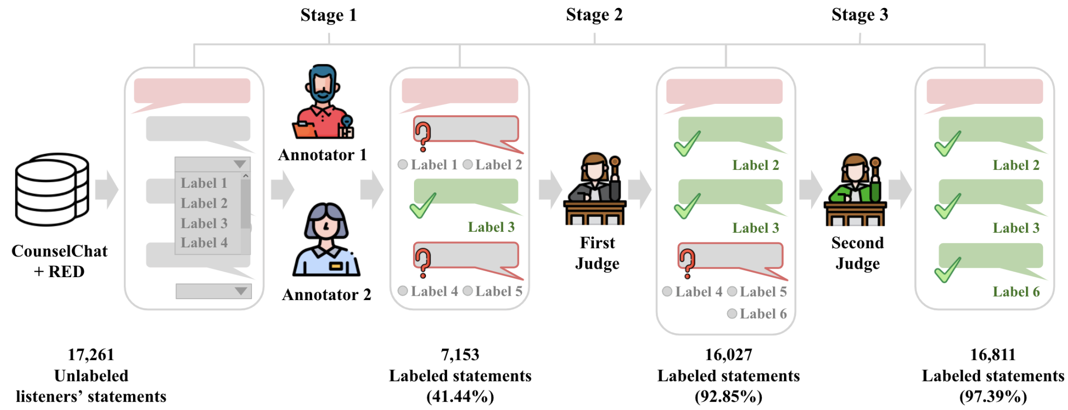



Figure 6.1: The annotation process to label the listeners' statements in the CounselChat and RED datasets with labels adapted from MITI. The process was conducted in three main stages.

from the MITI code. 2) We discuss the details of the annotation process followed in increasing the agreement between the workers when annotating with MITI codes. 3) Based on these annotations, we conduct a comparative analysis between counselors' and peers' responses and draw useful conclusions on to what extent responses from peers align with responses from trained counselors and recommend ways of boosting peers' responses such that it can increase the perceived effectiveness of therapeutic chatbots trained on such data.

## 6.2 Related Work

Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2003) exclusively focuses on the verbal behaviour of a counselor and is used to increase clinical skills in the practice of motivational interviewing (MI). This coding system has been used extensively to improve the understanding of the counseling practice alone (Can et al., 2012; Pérez-Rosas et al., 2018, 2019). Pérez-Rosas et al. (2016) developed an MI dataset consisting of ≈22K counselors' responses during Motivational Interviewing encounter annotated with 10 behavioral codes from the MITI. Althoff et al. (2016) conducted a quantitative study on counseling conversations to measure how various linguistic aspects of conversations are correlated with conversation outcomes. However, the datasets used in such analyses are not publicly available due to ethical reasons. Thus, it is difficult to use such resources in training therapeutic chatbots even though real counselor responses are the ideal candidates for this purpose.

There are a number of research efforts taken to develop therapeutic chatbots (Fitzpatrick et al., 2017; Inkster et al., 2018; Welch et al., 2020; Mousavi et al., 2021). Among them, recent work focuses on using dialogue data from peer-support forums (Sharma et al., 2020b; Welivita and Pu, 2022b). Some studies specifically focus on attributes such as perceived empathy and information richness in mental health-related discourse in peer support forums that suggests they are good candidates for training such chatbots (Nambisan, 2011; De Choudhury and De, 2014; Sharma et al., 2020a,b). But these studies lack comparisons with responses generated by professional counselors. In our work, we mainly attempt to address this limitation.

## 6.3 Methodology

In this section, we describe the methodology including the labels chosen to annotate the listeners' statements, the datasets used, and different stages of the annotation process.

### 6.3.1 Labels Adapted from MITI

The labels we used for annotation were adapted from MITI code 2.0 (Moyers et al., 2003) and 4.2.1 (Moyers et al., 2014). Table 6.1 shows the MITI labels that were used for annotation with descriptions and examples. Altogether, there are 15 labels. They also include labels *Self-Disclose* and *Other*, which are not included in the MITI code. We included the label *Self-Disclose* because in conversations involving peer support, *Self-Disclosure* is an important aspect that enables the sharing of lived experience and is seen to occur quite frequently in the majority of such conversations (Truong et al., 2019). The above labels were used to annotate each sentence in the listeners' utterances. The MITI does not contain an exhaustive list of all possible codes; thus not all sentences can be mapped to a label from the MITI code. In this case, the annotators were asked to select *Other*. Also, the labels from the MITI code are mutually exclusive. Thus, the same sentence could not receive more than one label.

### 6.3.2 Datasets

Along with the RED (Reddit Emotional Distress) dialogue dataset that was curated in Chapter 5 by scrapping dialogues from the peer support platform, Reddit, for comparison purposes we used the CounselChat dataset that contains responses from professional counselors. The CounselChat dataset consists of high-quality therapist responses to emotional distress related questions from individuals. This data is scraped from the CounselChat website (counselchat.com), which is an online platform that helps counselors to make meaningful contact with potential clients. On the website, professional counselors respond to questions posed by users suffering from mental health issues and emotional distress. The dataset consists of 2,129 post-response pairs that span across 31 distress related topics, the most frequent topics being *depression*, *relationships*, and *intimacy*. This dataset is publicly available. [3] Out of this data, we randomly selected 1K post-response pairs to be annotated with labels derived from the MITI code. The RED dataset contains utterances from peer-supporters as response for posts containing emotional distress. Out of the ≈1.3M dyadic conversations present in the dataset, 1K dialogues were randomly selected for annotation.

[3] https://github.com/nbertagnolli/counsel-chat

| MITI label | Description | Examples |
|---|---|---|
| 1. Closed Question | Questions that can be answered with an yes/no response or a very restricted range of answers. | *Do you think this is an advantage?* |
| 2. Open Question | Questions that allow a wide range of possible answers. | *What is your take on that?* |
| 3. Simple Reflection | Repetition, rephrasing, or paraphrasing of speaker's previous statement. | *It sounds like you're feeling worried.* |
| 4. Complex Reflection | Repeating or rephrasing the previous statement of the speaker but adding substantial meaning/emphasis to it. | **Speaker:** *Mostly, I would change for future generations.* **Listener:** *It sounds like you have a strong feeling of responsibility.* |
| 5. Give Information | Educating, providing feedback, or giving an opinion without advising. | *Logging your cravings is important as cravings often lead to relapses.* |
| **MI Adherent Behaviour Codes:** | | |
| 6. Advise with Permission | Advising when the speaker asks directly for advice. Indirect forms of permission can also occur, such as when the listener says to disregard the advice as appropriate. | *If you agree with it, we could try to brainstorm some ideas that might help.* |
| 7. Affirm | Encouraging the speaker by saying something positive or complimentary. | *You should be proud of yourself for your past's efforts.* |
| 8. Emphasize Autonomy | Emphasizing the speaker's control, freedom of choice, autonomy, and ability to decide. | *It is really up to you to decide.* |
| 9. Support | Statements of compassion or sympathy. | *I know it's really hard to stop drinking.* |
| **MI Non-Adherent Behaviour Codes:** | | |
| 10. Advise without Permission | Making suggestions, offering solutions or possible actions without first obtaining permission from the speaker. | *You should simply scribble a note that reminds you to take a break.* |
| 11. Confront | Directly and unambiguously disagreeing, arguing, blaming, criticizing, or questioning the speaker's honesty. | *Yes, you are an alcoholic. You might not think so, but you are.* |
| 12. Direct | Giving orders, commands, or imperatives. | *Don't do that!* |

| | | |
|---|---|---|
| 13. Warn | A statement or event that warns of something or that serves as a cautionary example. | *Be careful, DO NOT stop taking meds without discussing with your doctor.* |
| **Other:** | | |
| 14. Self-Disclose | The listener discloses his/her personal information or experiences. | *I used to be similar where I get obsessed about how people look.* |
| 15. Other | Statements that are not classified under the above codes | *Good morning, Hi there.* |

Table 6.1: The set of labels adapted from the MITI code, which were used to annotate listeners' responses.

### 6.3.3 Annotation Experiment

We used UpWork, a leading crowdsourcing platform to recruit trained counselors to annotate dialogue responses from CounselChat and RED datasets. Altogether 12 workers were recruited to annotate 2K dialogues, 1K from CounselChat, and 1K from RED. They contained 17,261 individual sentences in the listener utterances in total.

The task was carried out in three stages. First, the workers were asked to annotate each sentence contained in the listener utterances of the dialogues from CounselChat and RED datasets with one of fifteen MITI labels. We bundled ten dialogues into one batch (a batch contained five CounselChat and five RED dialogues interchangeably) and assigned two workers per batch. In the beginning, a tutorial about the labels accompanied by a practice task was offered to self-validate the workers' answers. As the task was ongoing, we computed the observed agreement of each worker with peers and offered more batches for the workers whose observed agreement was better than the others.

At the end of stage 1, we noticed that the two workers assigned to each batch did not agree on a common label for more than half of the listeners' sentences in the two datasets. Manual inspection of their answers revealed that the majority of the disagreements occurred because there are highly confusing pairs of labels that need more careful attention to differentiate (e.g. *Complex Reflection* and *Give Information* can be easily confused). Hence, we launched a second stage of the experiment by asking four workers who scored the highest observed agreement with a peer in the first stage to act as judges to resolve these conflicting labels. A judge was presented with the two labels the workers specified in the first stage along with the listener's sentence and the dialogue context and was asked to select either one of the two labels if one of them agreed with the listener's sentence. Only if none of the labels agreed with the listener's sentence, he was instructed to select a label from the rest.

At the end of stage 2, most of the conflicting labels were resolved by the judge's annotations. But there was a small percentage of listeners' sentences for which a label was still not agreed upon. We noticed for 68.15% of such unresolved sentences, at least one annotation was given by a relatively poor performing worker whose observed agreement score with a peer was below average as measured in the first stage. We decided such labels are not worth considering since they cloud the decision process and chose to launch a third stage of the experiment by removing one annotation given by the poorest performing worker for each such unresolved sentence. Similar to stage 2, we recruited the same judges and presented them with the two remaining labels to be resolved. This entire annotation pipeline is illustrated in Figure 6.1. The user interfaces of the annotation experiment are included in Appendix A.13.

### 6.3.4 Worker Quality

In stage 1 of the annotation process, to motivate the workers to pay attention to the task, we offered to pay them a bonus of $5 for each batch of dialogues that scored an above average observed agreement with a peer worker. Out of 400 worker assignments (200 batches × 2 workers per batch), 140 of them (35%) were able to receive this bonus. As the task progressed, those who scored higher observed agreements with the peer workers were allocated more batches to annotate.

In the second and third stages, to validate the quality of the judges and their attentiveness to the task, hidden checkpoints were included to measure the workers' attentiveness to the task. These checkpoints were based on the labels agreed upon by the two workers in the first stage of the task. In each batch of 10 dialogues, we randomly selected 10 sentences for which a label was agreed upon in the first stage. For each such sentence, we randomly sampled another label out of the remaining labels and showed it along with the correct label for the judge to select from. The four judges we recruited were able to get in overall 84.3% questions correct in stage 2 of the annotation task. The scores for each of the four judges were 80.00%, 86.47%, 86.47%, and 87.50%. In the third stage, they were able to get in overall 82.93% questions correct. Their individual scores were 83.00%, 83.64%, 82.00%, and 83.00%. All the scores being above 80% in both stages indicates that they all were paying significant attention to the task.

## 6.4 Results

Table 6.2 shows the statistics of the results from each stage of the experiment. At the end of stage 1, out of 17,261 listeners' sentences, 7,152 received a label as agreed by the two annotators. Altogether, by end of stage 1, we could yield an observed agreement percentage of 41.43% and an inter-rater agreement (Fleiss' kappa) score of 0.3391 indicating fair agreement. At the end of stage 2, another 8,875 labels were resolved, yielding an observed agreement of 87.79%. The updated inter-rater agreement (Fleiss' kappa) after this stage was 0.5292, which is a significant increase compared to the previous stage. After the end of completion of stage 1 and stage 2 of the annotation process, from among the total of 17,261 listeners' sentences in CounselChat and RED datasets, 16,027 of them were able to receive a label as agreed by at least two workers. This is 92.85% of the entire data.

From the remaining 1,234 sentences for which a label was not agreed upon, 841 (68.15%) sentences were annotated by at least one poor worker whose observed agreement with a peer was below average. At the end of stage 3 of the experiment, which was conducted by removing such annotations given by the poor workers, a second judge was able to agree with one of the two remaining labels for 784 sentences, yielding an observed agreement of 93.22%. The updated inter-rater agreement (Fleiss' kappa) after the third stage was 0.5453 (moderate agreement), showing a slight increase compared to the score in the previous stage. The lower kappa scores are potentially due to the inherent difficulty of distinguishing some of the MI labels, which we elaborate below. A similar annotation experiment conducted by Perez-Rosas et al. (2016) reports similar kappa scores ranging from 0.31 to 0.64 on different MI labels.

| **Description** | **CC** | **RED** | **CC + RED** |
|---|---|---|---|

| | | | |
|---|---|---|---|
| **Stage 1:** | | | |
| Total number of listeners' sentences annotated | 9,893 | 7,368 | 17,261 |
| Sentences for which a label was agreed upon by both annotators | 4,067 | 3,085 | 7,152 |
| Observed agreement between the two annotators | 41.11% | 41.87% | 41.43% |
| Inter-rater agreement (Fleiss' kappa) | 0.3059 | 0.3577 | 0.3391 |
| **Stage 2:** | | | |
| The number sentences, for which, the label had to be resolved | 5,826 | 4,283 | 10,109 |
| The number of times the judge agreed with one of the given labels | 5,111 | 3,764 | 8,875 |
| Observed agreement between the judge and one of the two annotators | 87.73% | 87.88% | 87.79% |
| Updated inter-rater agreement (Fleiss' kappa) | 0.5029 | 0.5440 | 0.5292 |
| **Stage 3:** | | | |
| The number sentences, for which, the label had to be resolved | 479 | 362 | 841 |
| The number of times the judge agreed with one of the given labels | 450 | 334 | 784 |
| Observed agreement b/w the judge and one of the remaining annotators | 93.95% | 92.27% | 93.22% |
| Updated inter-rater agreement (Fleiss' kappa) | 0.5193 | 0.5601 | 0.5453 |

Table 6.2: Statistics of the three stages of the annotation experiment. The CounselChat dataset is abbreviated as CC.

The confusion matrices computed at each stage of the experiment are denoted in Appendix A.14. We could observe that the label pair *Complex Reflection - Give Information* had the highest percentage of disagreement between the two workers in both CounselChat and RED datasets. In addition, the label pairs *Advise with Permission - Advise without Permission* and *Give Information - Advise without Permission* were highly confusing to differentiate in the CounselChat dataset. Whereas, in RED, the label pairs *Affirm - Support* and *Advise without Permission - Direct* were difficult to be differentiated. These observations were quite intuitive since these pairs of labels either contained similar semantic content (e.g. *Complex Reflection - Give Information, Advise with Permission - Advise without Permission, Give Information - Advise without Permission, Advise without Permission - Direct*) or used similar language constructs (e.f. *Affirm - Support, Advise without Permission - Direct*).

Final aggregated statistics of the three stages of the annotation process is shown in Table 6.3. It could be observed how the labels grew to cover a larger portion of the listeners' sentences as the annotation process advanced through the stages. Finally, close to 97% of the listeners' sentences (16,812 in total) were annotated with the MITI labels.

| **Description** | **CC** | **RED** | **CC + RED** |
|---|---|---|---|
| # listener sentences | 9,893 | 7,368 | 17,261 |
| # labels agreed in stage 1 | 3,085 (41.11%) | 4,067 (41.87%) | 7,152 (41.43%) |
| # labels agreed in stage 2 | 9,178 | 6,849 | 16,027 |

| | | | |
|---|---|---|---|
| | (92.96%) | (92.77%) | (92.85%) |
| # labels agreed in stage 3 | 9,628 | 7,183 | 16,811 |
| | (97.49%) | (97.32%) | (97.39%) |

Table 6.3: Final aggregated statistics of the three stages of the annotation process.

## 6.5 Analysis of the MI Dataset

In Figure 6.2, we show the distribution of labels adapted from the MITI code in CounselChat and RED datasets, separately. Table 6.4 shows the specific number of each label in CounselChat and RED datasets and the increase/decrease in each label in the two datasets compared to each other. Based on these statistics, we discuss seven major differences observed between responses from counselors and peers and state our recommendations when using this data to train therapeutic conversational agents.

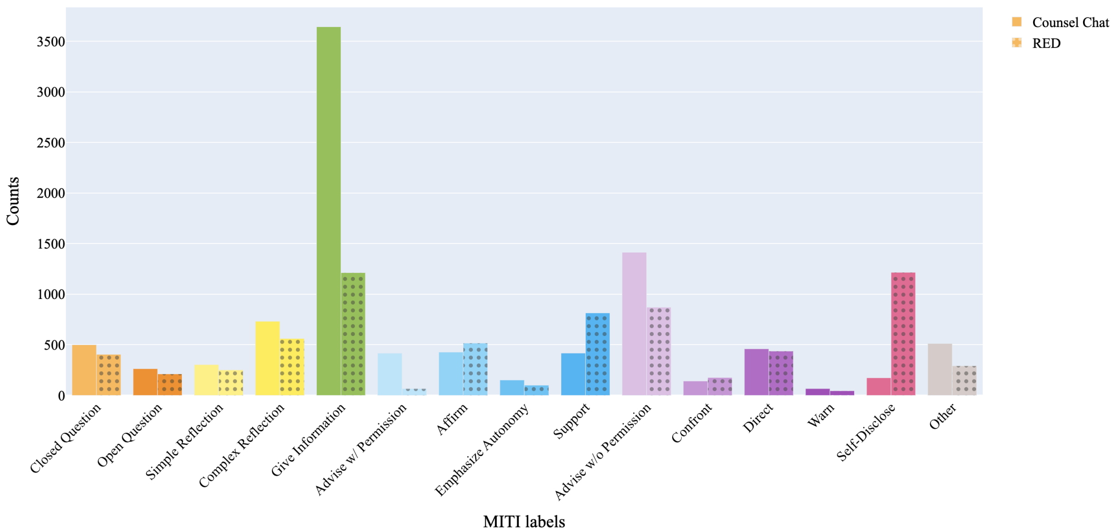



Figure 6.2: Distribution of MITI labels in CounselChat and RED datasets.

| **Label** | **No. of labels in CounselChat** | **No. of labels in RED** | **Increase in CC compared to RED** | **Increase in RED compared to CC** |
|---|---|---|---|---|
| Closed Question | 500 | 405 | 23.46% ↑ | -19.00% ↓ |
| Open Question | 264 | 212 | 24.53% ↑ | -19.70% ↓ |
| Simple Reflection | 304 | 252 | 20.63% ↑ | -17.11% ↓ |
| Complex Reflection | 732 | 562 | 30.25% ↑ | -23.22% ↓ |
| Give Information | 3,643 | 1213 | 200.33% ↑ | -66.70% ↓ |
| **MI Adherent Behavior Codes:** | | | | |
| Advise with Permission | 417 | 67 | 522.39% ↑ | -83.93% ↓ |
| Affirm | 428 | 517 | -17.21% ↓ | 20.79% ↑ |
| Emphasize Autonomy | 152 | 101 | 50.50% ↑ | -33.55% ↓ |
| Support | 418 | 815 | -48.71% ↓ | 94.98% ↑ |

| | | | | |
|---|---|---|---|---|
| **MI Non-Adherent Behavior Codes:** | | | | |
| Advise without Permission | 1,414 | 871 | 62.34% ↑ | -38.40% ↓ |
| Confront | 142 | 176 | -19.32% ↓ | 23.94% ↑ |
| Direct | 460 | 438 | 5.02% ↑ | -4.78% ↓ |
| Warn | 67 | 46 | 45.65% ↑ | -31.34% ↓ |
| **Other:** | | | | |
| Self-Disclose | 174 | 1216 | -85.69% ↓ | 598.85% ↑ |
| Other | 513 | 292 | 75.68% ↑ | -43.08% ↓ |

Table 6.4: Statistics of MITI labels in CounselChat and RED datasets and the increase/decrease in each label in the two datasets compared to each other. The increases/decreases that are favourable for the interaction are indicated in green while those that are unfavourable are indicated in red. The increases/decreases in *Self-Disclose* and *Other* are not assigned a color as their role in therapeutic interventions are quite blurry and subjected to debate.

**Questions:** There is ≈23% and ≈26% increase in closed and open questions, respectively, in counselor responses from CC compared to peer-support responses from RED. Questioning plays a central role in therapeutic interactions as it builds up mutual dialogue between client and therapist (Poskiparta et al., 2000). But Hill et al. (1983) observed with time-limited counseling, fact finding through closed questions is rated lower in helpfulness. It can result in the speaker saying less and less and the listener feeling pressured to ask more questions to keep the interaction going. However, in both CounselChat and RED datasets, the number of open questions is nearly half of the number of closed questions. Hence, mechanisms should be devised to increase the percentage of open questions to balance the number of closed questions. This combination would be more effective than a disproportionate reliance on closed questions.

**Reflections:** The number of reflections is positively associated with the perceived empathy (Klonek et al., 2015). It is also a competence indicator in assessing MI competency (Moyers et al., 2003). Non-surprisingly, both simple and complex reflections are observed to be higher (≈20% and ≈30% increase in simple and complex reflections, respectively) in counselors' responses compared to peers'. Thus, it would be beneficial to boost the percentage of reflections among peer support dialogues when using them to train therapeutic agents.

Scholars emphasize that listeners should formulate more reflections than closed questions (Klonek et al., 2015). As we observed, some closed questions such as *"Are you eating because you are bored?"* are identical to reflections, differing only in the voice intonation at the end. They could be easily reformulated into reflections such as *"It seems that you are eating because you are bored"*.

**Giving information:** In counselor responses, there is a 200.33% increase of *Give Information* type of sentences compared to responses from peers. It is quite unsurprising since counselors are relatively knowledgeable about the subject being discussed and hence are in a position to provide information that can help the speaker. Informed by this observation, steps should be taken to boost the amount of information in peer-support responses.

**MI Adherent Behavior Codes:** Supporting the client with statements of compassion and sympathy are surprisingly higher among peers (≈95% increase) compared to counselors. Affirming the speaker by saying something positive or complimentary is also seen to be comparatively higher in RED (≈21% increase). These are very good indicators that show peer-support responses if utilized in training therapeutic agents will reflect more compassion, sympathy, positivity, and compliments towards the user in distress. On the other hand, emphasizing the speaker's control and autonomy is observed to be higher in counselors' responses (≈50% increase) compared to responses from peers.

**Advising with and without permission:** Giving advices is generally seen to be higher in counselor responses. There is ≈522% increase in advising after asking for permission and ≈62% increase in advising without asking for permission among counselors' responses compared to those from the peers. Advising without asking for permission takes a portion of 77.22% of the total number of advices given in counselor responses. Thus, counselors, though professionally trained, tend to make the mistake of advising without prior asking for the speaker's permission. This percentage is higher in peer-support responses in which advising without permission takes a portion of 92.86% of the total number of advices given by the peers. Thus, in both datasets, steps should be taken to reformulate advices in a way that the agent asks for the speakers' permission before giving advice.

**MI Non-Adherent Behavior Codes:** Confronting the client by directly disagreeing, arguing, or criticizing is higher in peers' responses (≈24% increase) compared to those of the counselors'. Such interactions reflect uneven power sharing, accompanied by disapproval and negativity (Moyers et al., 2003). Directing the speaker by giving orders and also warnings are quite surprisingly seen to be slightly higher in the responses given by the counselors compared to the responses of the peers (≈5% and ≈46% increase for *Direct* and *Warn*, respectively). These are non-favourable response types that negatively affect the therapeutic interaction between the speaker and the listener and thus should be detected and eliminated as a preliminary step before using such responses to train chatbots.

**Self-Disclosure:** The role of self-disclosure in therapeutic interventions is quite blurry. For example, psychoanalysts believe that self-disclosure is counterproductive as it distorts client's transference. Conversely, Cognitive Behavioural therapists believe that self-disclosure can be a useful tool in therapy as it models and reinforces new perspectives for the client. Digging deep, there are two broad types of self-disclosure used by counsellors: 1) *intra-session disclosure*, where the counselor discloses a feeling about the client that is relevant to the therapeutic process; and 2) *extra-session disclosure*, where the counselor reveals information about themselves that occurs outside the session. In most cases, *intra-session disclosure* is the most useful type of self-disclosure (Levitt et al., 2016).

As we manually inspected the statements labeled as *Self-disclosure* in CounselChat and RED datasets, it was found out that *Intra-session disclosure* is seen higher in CC compared to RED, whereas *Extra-session disclosure* is seen higher in RED compared to CC. Table 6.5 provides some examples of such statements. This suggests that counselors are more careful when disclosing information about themselves and when they do they make sure that the information they disclose is relevant to the therapeutic process. Extra-session disclosure too has its place in therapeutic interactions specially contributing to building rapport between the client and the therapist. However, as suggested by R. Schwartz (2021), this type of disclosure must be used wisely with caution since it can as well be counterproductive distorting client's transference.

| |
|---|
| **Examples of *intra-session disclosure* in CounselChat:** |
| - *Personally, I can tell you that I would want my clients to tell me about anxiety they feel 100% of the time.* |
| - *I have had clients asking the same question and there is often an underlying fear that they "can't be helped" or they will "be too much for their therapist."* |
| **Examples of *extra-session disclosure* in RED:** |
| - *You remind me a lot of my best friend that I had when I was young. Being her friend was exhausting.* |
| - *I too suffer from psychosis from my schizo-affective disorder, yelled at my former best friend for gangstalking me, called her all kinds of horrible names.* |

Table 6.5: Examples of different types of self-disclosure observed in CounselChat and RED datasets.

## 6.6 Chapter Summary

In this chapter, we discussed the curation process of a large-scale distress consoling dialogue dataset containing utterances from trained counselors and peers. A carefully designed annotation process was followed to annotate each response statement with labels adapted from the MITI code. We saw the effectiveness of our annotation process as it contributed to increasing the observed agreement and inter-rater reliability as the process advanced through different stages. Based on the comparative analysis between responses from counselors and peers, we reported seven major differences between them, highlighting the strengths and limitations of using abundantly available peer-support dialogues for purposes such as training therapeutic chatbots. In summary, peers' responses tend to be more supportive, compassionate and encouraging than counselors' as observed by the increased percentage of *Support* and *Affirm* labels. But important therapeutic techniques such as asking more open questions than closed ones, reflections, giving information and advices with permission, and emphasizing the speaker's autonomy require further boosting. MI non-adherent behaviors such as confronting are also seen higher among peers and thus should be eliminated. Careful attention should also be paid to self-disclosure among peers as the majority of such statements are of the type *extra-session disclosure*, which is less useful for the therapeutic process. Curating this dataset is the first step in our goal of boosting the therapeutic competency in dialogue agents trained on this data, which will be described in the following chapter.

# 7 Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy

*This chapter is based on the work of Anuradha Welivita, and Pearl Pu (Welivita and Pu, 2023a). The author of this thesis (Anuradha Welivita) was mainly responsible for the design and training of the MI classifier and the MI rephrasing models, the development of pseudo-parallel corpora for training, and automatic and human evaluation of the MI rephrasing models.*

## 7.1 Introduction

As it was emphasized in Chapter 6, due to the lack of availability of large-scale psychotherapeutic conversations, researchers are using data scraped from online peer support forums such as Reddit to train chatbots that can offer distress consolation (Alambo et al., 2019; Roller et al., 2021). But since peers are not professionals, the responses contained in such forums can sometimes be unfavourable to address distress (e.g. confrontations, judgments, orders etc.). So, using this data can have severe risks. One solution for this is identifying favourable and unfavourable response types that appear in distress support dialogues and developing automatic means that can propose omission or rephrasing of such unfavourable response types. Figure 7.1 shows an example.

In Chapter 6, we analyzed the types of responses in distress support dialogues, using labels adapted from the Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2014). We developed the MI dataset, to have a comparative understanding of the differences between online support provided by peers and trained counselors by hiring professional counselors to annotate responses given by peers and counselors with labels derived from the MITI code. During analysis, we observed that peers' responses tend to be more supportive, and encouraging than counselors' (as observed by the increased percentage of *Support* and *Affirm* labels). But it was also observed that important therapeutic techniques, such as asking more *open questions* than *closed* ones, *reflections*, *giving information*, *advices with permission*, and *emphasizing speaker's autonomy* were lacking in peers' responses and hence require further boosting. One of the major observations was that among the advises given by the peers, 92.86% of them belonged to the category *Advise without permission*, which is MI non-adherent.

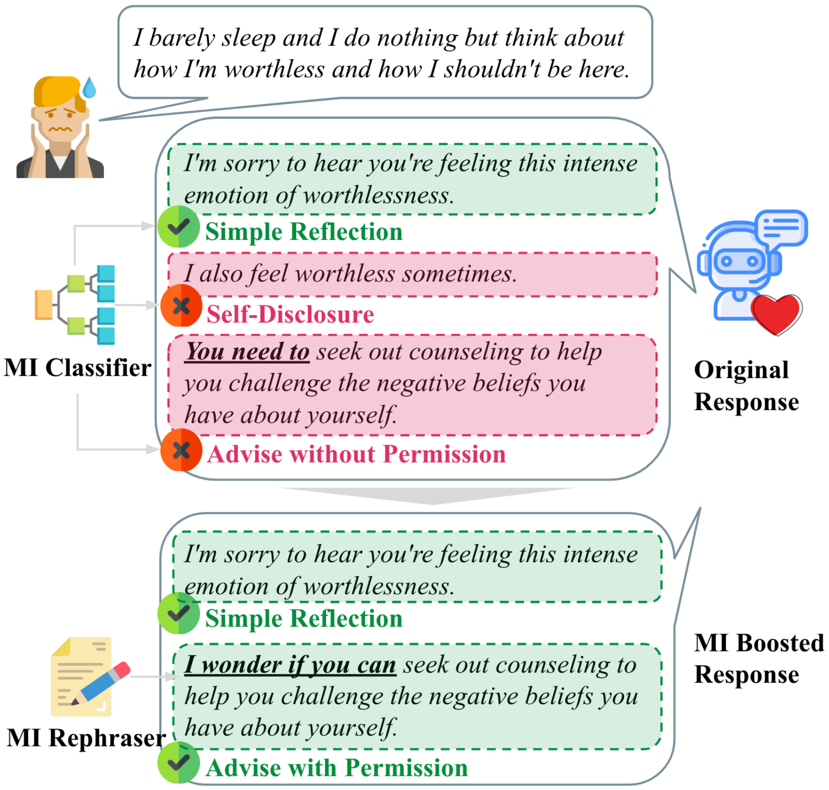



Figure 7.1: Example of detecting unfavourable and favourable response types in distress support dialogues and boosting the responses by omitting unfavourable responses or rephrasing them into more favourable ones.

This percentage was lower in counselor responses, but still accounted for 77.22% of the advises given by counselors.

In this work, we aim to detect such *Advise without permission* responses among distress support dialogues and build a rephraser that can rephrase such responses into *Advise with permission*, which is more MI-adherent. First, we detect such responses through a classifier trained on an augmented version of the MI dataset. Next, as we do not have human written responses rephrasing *Advise without permission* responses into *Advise with permission*, we use automatic methods such as template-based replacement and retrieval to construct a pseudo-parallel training corpus containing pairs of *Advise without permission* and *Advise with permission* sentences. Since rephrasing is a labor-intensive task compared to labeling and we require professionally trained counselors to do this in the distress consolation setting, using our already labeled dataset to construct a pseudo-parallel corpus saved us both time and cost. We apply the same methods on the augmented version of the MI dataset to form a much larger pseudo-parallel training corpus and use these corpora to fine-tune BlenderBot (Roller et al., 2021) and GPT3 (Brown et al., 2020b). Some of the models we fine-tune incorporate different forms of prompting with the aim of obtaining a better outcome with less training examples. We evaluate the rephrasers using automatic and human evaluation. The results mainly show when the training dataset is small, prompting improves the performance of the rephrasers across style transfer and semantic similarity dimensions. They also suggest that when the training dataset is large (in our case through data augmentation), pseudo-parallel data generated through simpler methods such as template replacement produce better results.

Our contributions are four-fold. 1) We develop an MI classifier that can predict 15 different favourable and unfavourable response types derived from the MITI code. 2) We propose a methodology to rephrase responses detected as *Advise without Permission* into more MI-adherent *Advise with Permission*. We show how this can be done in the absence of human written rephrasings by developing pseudo-parallel corpora using different automatic methods. 3) We evaluate these rephrasers using automatic and human evaluation and show how prompting and data augmentation can improve the performance of the rephrasers when there is less training data. 4) Finally, we discuss how this method can be applied to boost chatbot responses, making them more compliant with the MI strategy.

## 7.2 Related Work

Rephrasing responses recognized as *Advise without Permission* into *Advise with Perrmission* can be identified as a sub-task falling under the task of Text Style Transfer (TST), in which the goal is to automatically control the style attributes (e.g. sentiment, politeness, humor, etc.) of text while preserving the content (Jin et al., 2022). The field of TST involves traditional linguistic approaches as well as deep learning approaches. Traditional approaches to TST rely on term replacement and templates (Mairesse and Walker, 2011; Sheikha and Inkpen, 2011). With the success of deep learning, various neural methods have been recently proposed for TST. Given datasets in which there are direct mappings between the text of the source style and the text of the target style, which are referred to as parallel corpora, standard sequence-to-sequence models are often directly applied for TST (Rao and Tetreault, 2018; Shang et al., 2019; Xu et al., 2019). But parallel corpora are challenging to find because the development of such data often requires costly human labor. Thus, TST on non-parallel corpora has become an emerging area of research (Li et al., 2018; Jin et al., 2019; Liu et al., 2022).

Parallel and nonparallel datasets have been proposed for common sub-tasks of TST such as sentiment (Shen et al., 2017), topic (Huang et al., 2020), formality (Rao and Tetreault, 2018), politeness (Madaan et al., 2020), and humor (Gan et al., 2017) transfer. But to the best of our knowledge, this is the first attempt at introducing a new sub-task and releasing a nonparallel corpus for style transfer between MI non-adherent *Advise without Permission* and MI adherent *Advise with Permission* responses. This task is more challenging than the other sub-tasks because it requires the expertise of professional counselors to generate training data. In this work, we release a nonparallel corpus that can be utilized for this task, which is annotated by professional counselors. We also show how automatic methods could be applied to create pseudo-parallel corpora using this dataset, which can be used to train neural models for this task.

## 7.3 Datasets

For this work, we used the CounselChat dataset containing 2,129 post-response pairs, in which we get responses from verified counselors and the RED dataset containing 1,275,486 dyadic conversations, in which we get responses from peer supporters. We also used the MI dataset developed in Chapter 6, in which a subset of 1,000 dialogues each from CounselChat and RED datasets were annotated with labels adapted from the MITI code 2.0 (Moyers et al., 2003) and 4.2.1 (Moyers et al., 2014). Out of the 15 labels the dataset is annotated with, we are interested in the labels *Advise with Permission* and *Advise without Permission*, which are respectively considered MI-adherent and MI non-adherent response types. The MI dataset contains 16,811 annotated responses, out of which 2.87% (484) and 13.5% (2,285) responses are labeled as *Advise with Permission* and *Advise without Permission*, respectively.

To further augment the MI dataset, we used automatic labeling to expand the 15 labels into unlabeled dialogue responses from CounselChat and RED datasets. We used two automatic methods for this purpose: 1) n-gram-based matching; and 2) similarity based retrieval.

**N-gram Based Matching:** By tokenizing the responses in the MI dataset and computing the frequencies, we discovered the most frequent n-grams (four-grams and five-grams) occurring among the 15 labels. Examples of them are shown in Appendix A.15. Next, we searched for the presence of these indicative n-grams (first five-gram and then four-grams) among individual sentences that appear in dialogue responses of the unlabeled CounselChat and RED datasets. If an indicative N-gram was found in a sentence, we labeled that sentence with the label that n-gram is indicative of. The sentences with overlapping labels were discarded due to ambiguity. In this way, we were able to automatically label 1,918 and 340,361 sentences in CounselChat and RED datasets, respectively.

**Similarity Based Retrieval:** For each unlabeled sentence among the responses in CounselChat and RED datasets, we computed the cosine similarity with each of the labeled sentences in the MI dataset. Next, for each unlabeled sentence, we retrieved the labeled sentences whose cosine similarity is higher than a certain threshold (the thresholds were different for each of the 15 labels, which were selected after manually inspecting randomly selected pairs of unlabeled and labeled sentences corresponding to different labels). Next, we used a majority voting scheme to select the label we can associate the unlabeled sentence with. When we encountered ties, we computed the average similarities across the clusters of retrieved sentences with different labels that held a tie and selected the label based on maximum average similarity. In Figure 7.2, we show an example elaborating this procedure. Using this method, we were able to automatically annotate 2,881 and 1,196,012 sentences in CounselChat and RED datasets, respectively.

Appendix A.16 shows the statistics of the labels extended through both n-gram matching and similarity-based retrieval in CC and RED datasets. Using the union and the intersection of the labels retrieved from n-gram-based matching and similarity-based retrieval and combining them with the gold labels from the MI dataset, we created two augmented-labeled MI datasets

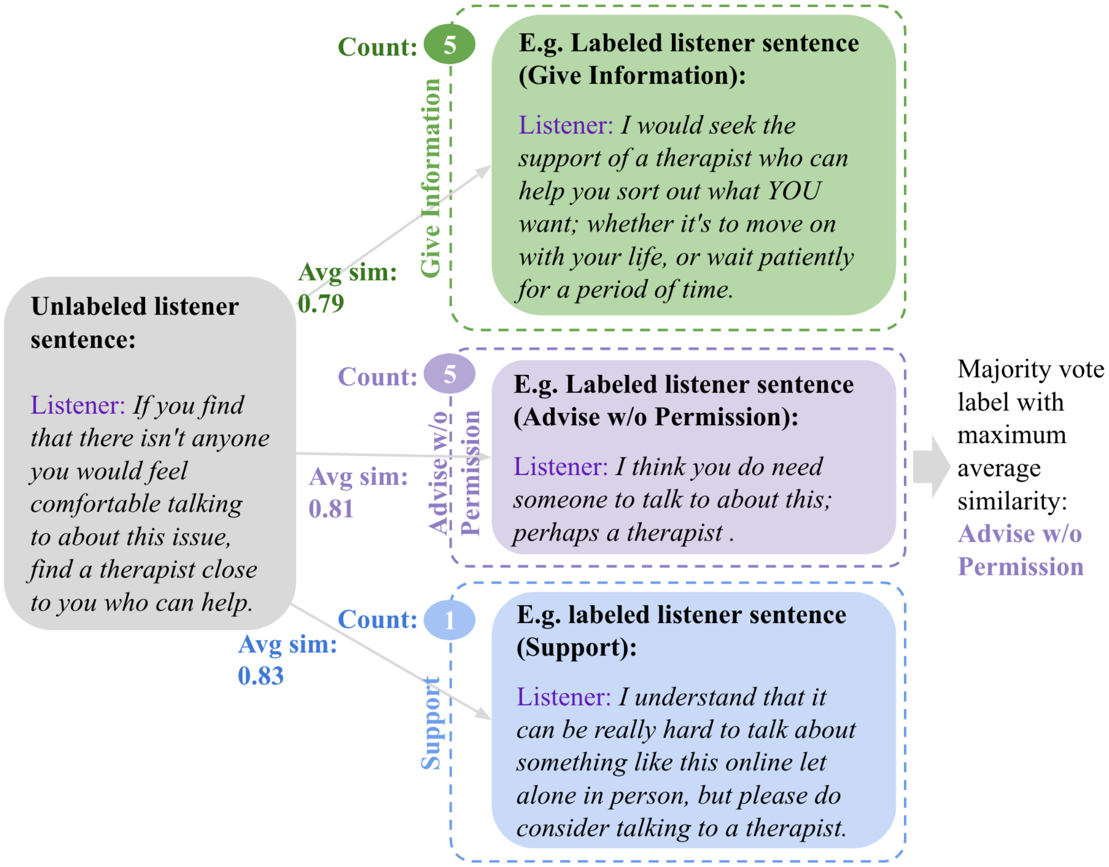



Figure 7.2: An example of automatically labeling an unlabeled sentence by computing the cosine-similarity with labeled sentences. The label is chosen based on majority voting. But this example shows a tie. Thus, we compute the average similarity of the sentence clusters that hold a tie and select the label of the sentence cluster with the maximum average similarity.

having 1,378,469 and 84,052 labeled sentences, respectively. For simplicity, we will refer to them as *MI Augmented (Union)* and *MI Augmented (Intersection)* datasets. Appendix A.17 shows the statistics corresponding to each label in the MI Augmented (Union) and MI Augmented (Intersection) datasets.

## 7.4 MI Classifier

We developed a classifier to automatically classify responses in distress-support dialogues into one of the 15 labels mentioned above. This is an important step that should be followed before rephrasing, since first it should identify the unfavourable responses types. For this purpose, we developed a classifier that consists of a representation network that uses the BERT architecture (Devlin et al., 2019), an attention layer that aggregates all hidden states at each time step, a hidden layer, and a softmax layer. We used the BERT-base architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters as the representation network. It was initialized with weights from RoBERTa (Zhuang et al., 2021). We trained three classifiers. The first one was trained on the smaller human-annotated MI dataset (MI Gold) taking 80% of the data for training and leaving 10% each for validation and testing. The other two were trained on the MI Augmented (Union) and MI Augmented (Intersection) datasets, leaving out the data used for validation and testing in the first case. In all cases, the optimal model was chosen based on average cross entropy loss calculated between the ground truth and predicted labels in the human-annotated validation set.

The classifiers trained on MI Gold, MI Augmented (Intersection), and MI Augmented (Union) datasets reported accuracies of 68.31%, 67.13%, and 73.44% in the MI Gold test set, respectively. The reported accuracies on the MI Gold validation set were 67.08%, 64.07%, and 72.67%, respectively for the three classifiers. Accordingly, the labels collected through the union of n-gram matching and cosine similarity-based methods improved the accuracy of the classifier by 8.33% and 7.5%, respectively on the validation and test sets compared to the accuracies reported when trained on the gold-labeled MI dataset. Additional technical details related to the MI classifier is indicated in Appendix A.18.

## 7.5 MI Rephraser

After identifying the favourable and unfavourable response types, we can choose to omit the unfavourable responses or if possible, rephrase them into a more MI adherent form. A label pair that this rephrasing strategy can be applied directly are *Advise without Permission* and *Advise with Permission*. Through N-gram analysis, we could discover some N-gram patterns that are indicative of the label pair *Advise without Permission* (e.g. *You should*, *You need to*, *You musn't*) and *Advise with Permission* (e.g. *It maybe helpful to*, *I wonder if you can*, *You may want to consider*). These could be identified as style attributes that vary across the responses identified as *Advise without Permission* and *Advise with Permission*. Thus, given a response identified as *Advise without Permission*, the goal of the rephraser would be to rephrase the response to be indicative of *Advise with Permission*, without changing the semantic content of the response.

As mentioned in Section 7.2, this can be identified as a sub-task under the task of Text Style Transfer (TST). TST is formally defined as, given a target utterance $x'$ and the target discourse style attribute $a'$, model $p(x'|a,x)$, where $x$ is a given text carrying a source attribute value $a$. In our case, $x$ corresponds to the response identified as *Advise without Permission*, $a$ corresponds to *Advise without Permission*, and $a'$ corresponds to *Advise with Permission*.

### 7.5.1 Pseudo-Parallel Corpora

As discussed in Section 7.2, the most recent methods for TST involve data-driven deep learning models. The prerequisite for using such models is that there exist style-specific corpora for each style of interest, either parallel or nonparallel. With the human-annotated MI dataset, we are in possession of a non-parallel corpus containing 2,285 *Advise without Permission* and 484 *Advise with Permission* type of responses. With the MI Augmented (Union) dataset, we have 199,885 *Advise without Permission* and 3,541 *Advise with Permission* type of responses. Since creating parallel corpora consumes human labor and cost, using the above data, we decided to create pseudo-parallel corpora that contain pairs of *Advise without Permission* and *Advise with Permission* responses to train our rephrasers. We used two automatic methods to create these pseudo-parallel corpora: 1) template-based replacement method; and 2) retrieval method.

### Template-Based Replacement Method

We used frequency-based N-gram analysis accompanied by human inspection to determine the linguistic templates that represent *Advise with Permission* and *Advise without Permission* responses. Table 7.1 shows the linguistic templates discovered for *Advise without Permission* (on left) and *Advise with Permission* (on right). In template-based replacement, if the algorithm detects any linguistic template on the left among the responses labeled as *Advise without Permission*, it will randomly select a template from the right to replace it with, giving a pair of *Advise without Permission* and *Advise with Permission* responses that contain the same semantic content but differ in style.

| Advise without Permission | Advise with Permission |
|---|---|
| - *You can* (verb) ___ | - *It maybe helpful to* (verb) ___ |
| - *You could* (verb) ___ | - *You may want to* (verb) ___ |
| - *You need to* (verb) ___ | - *I encourage you to* (verb) ___ |
| - *You should* (verb) ___ | - *Perhaps you can* (verb) ___ |
| - (Verb) ___ | - ___, *if you would like.* |
| - *You can try to* (verb) ___ | - *It would be good idea to* (verb) ___ |
| - *I think you should* (verb) ___ | - *It may be important to* (verb) ___ |
| - *I suggest that you* (verb) ___ | - *I would encourage you to* (verb) ___ |
| - *I suggest you* (verb) ___ | - *I wonder if you can* (verb) ___ |
| - *Maybe you can* (verb) ___ | - *Maybe it is important to* (verb) ___ |
| - *Maybe you could* (verb) ___ | - *An option would be to* (verb) ___ |
| | - *You may want to consider* (present continuous form of the verb) ___ |
| | - *You may consider* (present continuous form of the verb) ___ |
| | - *I would recommend* (present continuous form of the verb) ___ |
| | - *I wonder if you can consider* (present continuous form the verb) ___ |

Table 7.1: Linguistic templates corresponding to *Advise without Permission* and *Advise with Permission* responses.

We constructed two pseudo-parallel corpora by applying this method to the MI Gold and MI Augmented (Union) datasets, which contained 2,285 and 199,885 responses labeled as *Advise without Permission*, respectively. They respectively gave us 240 and 38,559 response pairs.

#### Retrieval Method

Given the non-parallel corpus containing *Advise without Permission* and *Advise with Permission* responses, we computed the semantic similarity between the *Advise without Permission* and *Advise with Permission* responses and retrieved the response pairs whose similarity is above a certain threshold. We used Sentence-BERT (Reimers and Gurevych, 2019) to generate embeddings of the two types of responses and compared them using cosine similarity. After manually inspecting a random subset of response pairs over a range of similarity thresholds, we chose 0.7 as the final threshold to determine the semantically similar response pairs. Similar to template-based replacement, we used this method to construct two pseudo-parallel corpora by applying the method to the gold-labeled and augmented-labeled MI datasets and obtained 104 and 54,956 response pairs, respectively. For simplicity, we will refer to the corpus constructed using the gold-labeled MI dataset as pseudo-parallel (PP) corpus and the corpus constructed using the augmented-labeled MI dataset as pseudo-parallel augmented (PPA) corpus. We used 80% of the data from each of the corpora for training our rephrasers, and 10% each for validation and testing. In section 7.7, we gauge the quality of the above corpora using human ratings.

In Figure 7.3, we visualize the process of creating Pseudo-Parallel (PP) and Pseudo-Parallel Augmented (PPA) corpora along with statistics corresponding to each dataset.

### 7.5.2 Rephrasing Models

Using the above corpora, we fine-tuned two pre-trained language generation architectures Blender (Roller et al., 2021) and GPT-3 (Brown et al., 2020b). Blender is a standard Seq2Seq transformer-based dialogue model. We used the 90M parameter version of Blender. Though it is a dialogue generation model, we used it mainly because it is pre-trained on Reddit discussions containing ≈1.5B comments and is already aware of the language constructs used in peer support. GPT-3 is a language model that utilizes standard transformer network having 175 billion parameters. We used the smallest but fastest version of GPT-3, Ada, to build our rephrasers. The main reason to use GPT-3 is that it has demonstrated strong few-shot learning capability on many text-based tasks. Both Blender and GPT-3 were fine-tuned on template-based, retrieval-based, and combined PP and PPA corpora.

Prior work has shown large language models can perform various tasks given a clever prompt prepended to the input (Brown et al., 2020b). So, we developed two variations of Blender and GPT3 models by appending a generic prompt and an N-gram-based prompt to the end of the training data. In generic prompting, we simply appended the label ***Advise with permission:*** to the end of the input text. In N-gram prompting, we detected if there is any N-gram that is indicative of *Advise with permission* in the output text. If there is, we appended it to the end of the input text. Table 7.2 shows training examples with generic and N-gram-based prompts.

---

**Training example with generic prompting:**

| | |
|---|---|
| Input: | *try to learn from your mistakes and meet some new people .* ***Advise with permission:*** |
| Output: | *It may be important to try to learn from your mistakes and meet some new people.* |
| **Training example with N-gram based prompting:** | |
| Input: | *try to learn from your mistakes and meet some new people .* ***It may be important to:*** |
| Output: | ***It may be important to*** *try to learn from your mistakes and meet some new people.* |

Table 7.2: Examples with generic and N-gram prompts.

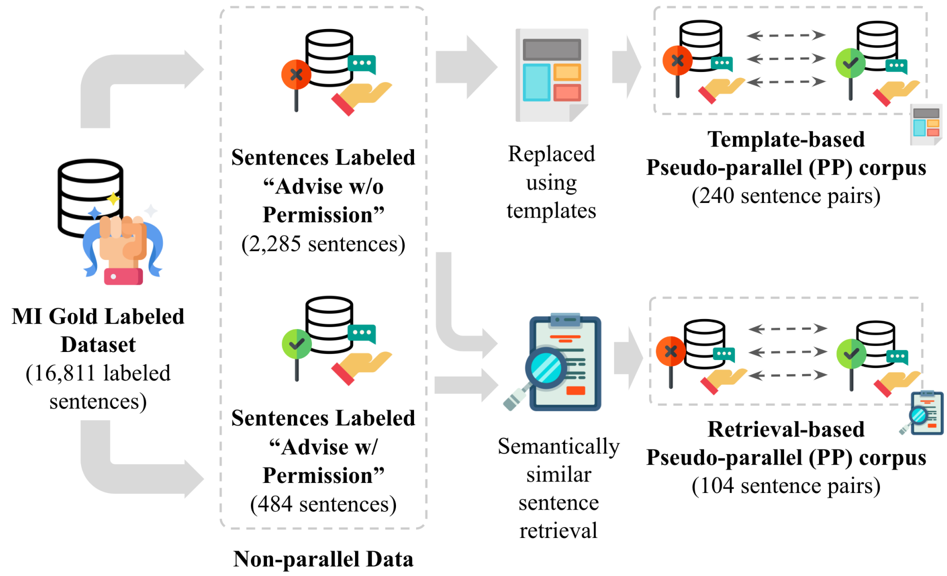
Altogether we developed 10 different rephrasing models by fine-tuning Blender and GPT-3 on: 1) template-based PP and PPA corpora; 2) retrieval-based PP and PPA corpora; 3) combined template-based and retrieval-based PP and PPA corpora; 4) combined template and retrieval based PP and PPA corpora appending generic prompts; 5) combined template and retrieval based PP and PPA corpora appending N-gram prompts. Additional technical details of the rephrasing models are shown in Appendix A.19. Some examples of the rephrased output by these different models are shown in Appendix A.20.

## 7.6 Automatic Evaluation

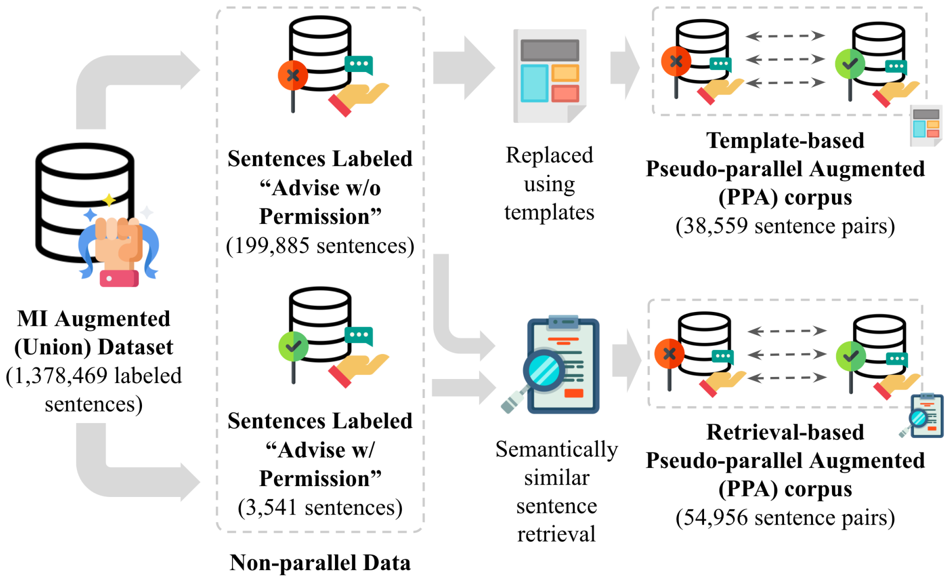
A successful style-transferred output should be able to demonstrate the correct target style and at the same time preserve the semantic content of the original text (Jin et al., 2022; Fu et al., 2018). We refer to the first criterion as *Style Transfer Strength* and the second as *Semantic Similarity*. Automatic metrics used to evaluate text generation methods such as the BLEU score (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), Word Mover Distance (WMD) (Kusner et al., 2015), Character N-gram F-score (chrf) (Popović, 2015), BERTScore (Zhang et al., 2019) and cosine similarity based on sentence embeddings (Reimers and Gurevych, 2019) are used in the literature to evaluate the semantic similarity between the original and the rephrased text. The Part-of-Speech distance (Tian et al., 2018), a metric specific to TST, is also used to measure semantic similarity. Mir et al. (2019) suggest deleting all attribute-related expressions in the text when applying these metrics to evaluate the output of TST tasks. Thus, before evaluation, we removed the style-specific phrases discovered during N-gram analysis from the input and output text. Additional technical details related to the automatic evaluation of the MI rephrasing models are denoted in Appendix A.21.

To evaluate the style transfer strength, most works use a style classifier to predict if the output conforms to the target style (Hu et al., 2017; Li et al., 2018; Prabhumoye et al., 2018). We used the MI classifier trained on the MI Augmented (Union) dataset to compute the style transfer strength. It is calculated as the percentage of samples classified as *Advise with Permission* out of all the test samples.

Table 7.3 shows the results of automatic evaluation of the rephrasers on the combined PP test dataset, which contains data from both template and retrieval-based PP test sets. Accordingly, GPT3-based rephrasers show better performance compared to Blender-based rephrasers




(a) Pseudo-Parallel (PP) Corpus




(b) Pseudo-Parallel Augmented (PPA) Corpus

Figure 7.3: Pseudo-Parallel (PP) and Pseudo-Parallel Augmented (PPA) corpus construction.

in 85% of the time across the metrics. It could also be observed that data augmentation improves the scores across most metrics irrespective of the backbone model used. Combining the pseudo-parallel corpora obtained from template-based and retrieval-based methods could improve the performance scores of Blender-based rephrasers across most automatic metrics. But GPT-3 based rephrasers trained only on template-based pseudo-parallel data seem to achieve better scores across almost all the metrics when compared to those trained on retrieval-based and combined corpora.

Blender-based rephrasers that used generic prompting ranked the best across most metrics over all the other Blender-based rephrasers. With the smaller PP training corpus, the GPT-3-based rephraser that incorporated generic prompting ranked the best across most metrics. But with the larger PPA training corpus, the GPT-3 based rephraser that was trained on simple template-replaced pseudo-parallel corpora ranked the best across most automatic metrics.

## 7.7 Human Evaluation

Similar to automatic evaluation, we used two human evaluation criteria to rate the rephrased sentences. The first is how close the rephrased sentence is to *Advise with permission* (Style transfer strength). The second is to what extent the rephrased sentence preserves the context/meaning of the original sentence (Semantic similarity).

We used the UpWork crowdsourcing platform (www.upwork.com) and recruited four professional counselors to rate the rephrased sentences. Given the original *Advise without Permission* sentence and a list of rephrased sentences generated by the 10 different rephrasers, we asked two questions from the counselors: 1) *Is the rephrased sentence indicative of Advise with permission?*; and 2) *Does the rephrased sentence preserve the original context?* The counselors were asked to answer these questions by indicating a rating on a Likert scale ranging from 0 (*Not at all*) to 4 (*Yes it is*). Along with the rephrased sentences, we also presented them the corresponding *Advise with permission* sentence obtained from the pseudo-parallel corpora in order to gauge the quality of the corpora used for training. The sentences to be rated were presented to them in a random order to reduce bias. User interfaces of the human evaluation task is denoted in Appendix A.22.

As the combined PP test corpus developed on the MI Gold dataset is small (only 34 samples), we used 200 randomly selected samples from the combined PPA test corpus developed on the augmented MI dataset to be rated by the human workers. This was to verify the trend of results reported on the PP test corpus. We bundled 9 randomly selected test cases in one batch and allocated two workers to rate each batch. Results were calculated based on the average rating given by the two workers. Following Adiwardana et al. (2020) we also calculated the average of style transfer strength and semantic similarity ratings to obtain a single score. We computed the inter-rater agreement based on weighted Kappa that uses Fleiss-Cohen weights (Wan et al., 2015) and the scores were 0.5870 (moderate agreement) and 0.6933 (substantial agreement) for style transfer strength and semantic similarity, respectively.

| Criteria | Template | | Retrieval | | Template + Retrieval | | Template + Retrieval (with generic prompting) | | Template + Retrieval (with N-gram prompting) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **BB** | **GPT3** | **BB** | **GPT3** | **BB** | **GPT3** | **BB** | **GPT3** | **BB** | **GPT3** |
| **Training dataset: PP** | | | | | | | | | | |
| BLEU-1 | 0.1315 | **0.3464** | 0.0787 | **0.1308** | 0.1429 | **0.2977** | 0.1763 | **0.3821** | 0.1585 | **0.2751** |
| BLEU-2 | 0.0366 | **0.3225** | 0.0131 | **0.0501** | 0.0496 | **0.2671** | 0.0613 | **0.3556** | 0.0677 | **0.2374** |
| BLEU-3 | 0.0046 | **0.3120** | 0.0046 | **0.0328** | 0.0000 | **0.2543** | 0.0031 | **0.3465** | 0.0000 | **0.2269** |
| BLEU-4 | 0.0033 | **0.2994** | 0.0000 | **0.0326** | 0.0000 | **0.2262** | 0.0000 | **0.3301** | 0.0000 | **0.2164** |
| ROUGE-L | 0.1760 | **0.5333** | 0.1176 | **0.1608** | 0.1843 | **0.4495** | 0.2167 | **0.5450** | 0.2135 | **0.4404** |
| METEOR | 0.1568 | **0.4622** | 0.0994 | **0.1323** | 0.1879 | **0.4210** | 0.2084 | **0.5014** | 0.2108 | **0.3726** |
| WMD ↓ | 1.0311 | **0.7068** | 1.1122 | **1.0800** | 1.0345 | **0.7928** | 1.0073 | **0.6746** | 1.0163 | **0.8447** |
| Chrf Score | 0.2690 | **0.5008** | 0.1678 | **0.2095** | 0.2690 | **0.4737** | 0.3082 | **0.5341** | 0.2955 | **0.4245** |
| BERTScore | 0.8656 | **0.9138** | 0.8382 | **0.8658** | 0.8683 | **0.9048** | 0.8821 | **0.9137** | 0.8693 | **0.9003** |
| POS dist. ↓ | 5.4771 | **2.5523** | 9.8218 | **7.1482** | 5.8271 | **2.7042** | 4.8378 | **2.5830** | 5.8854 | **3.6298** |
| Cos Similarity | 0.6116 | **0.7524** | **0.4429** | 0.4291 | 0.6129 | **0.6516** | 0.6918 | **0.7403** | **0.6571** | 0.6471 |
| Style Strength | 29.41 | **73.53** | 0.00 | **47.06** | 38.24 | **79.41** | **94.12** | 61.76 | 23.53 | **58.82** |
| **Training dataset: PPA** | | | | | | | | | | |
| BLEU-1 | 0.2039 | **0.3751** | **0.2122** | 0.0987 | 0.2308 | **0.3229** | 0.2588 | **0.3688** | 0.2021 | **0.3349** |
| BLEU-2 | 0.0913 | **0.3456** | **0.1468** | 0.0263 | 0.1591 | **0.2836** | 0.1849 | **0.3332** | 0.1455 | **0.3034** |
| BLEU-3 | 0.0031 | **0.3352** | **0.1370** | 0.0172 | 0.1319 | **0.2725** | 0.1536 | **0.3161** | 0.1239 | **0.2922** |
| BLEU-4 | 0.0000 | **0.3217** | **0.1286** | 0.0069 | 0.1213 | **0.2536** | 0.1437 | **0.2987** | 0.1169 | **0.2798** |
| ROUGE-L | 0.2642 | **0.5363** | **0.2419** | 0.1216 | 0.2718 | **0.4467** | 0.3016 | **0.5278** | 0.2352 | **0.5178** |
| METEOR | 0.3081 | **0.4673** | **0.2436** | 0.1063 | 0.2932 | **0.4261** | 0.3102 | **0.4607** | 0.2557 | **0.4381** |
| WMD ↓ | 0.9716 | **0.6849** | **1.0069** | 1.1584 | **0.9451** | 0.9754 | 0.9095 | **0.7258** | 1.0000 | **0.7927** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Chrf Score | 0.3758 | **0.5038** | **0.3550** | 0.1782 | 0.4005 | **0.4648** | 0.4048 | **0.5047** | 0.3672 | **0.4897** |
| BERTScore | 0.8770 | **0.9116** | **0.8748** | 0.8582 | 0.8795 | **0.9021** | 0.8837 | **0.9140** | 0.8700 | **0.9028** |
| POS dist. ↓ | 7.4745 | **1.9593** | 8.0439 | **7.0396** | 6.9338 | **2.8695** | 6.1747 | **2.6637** | 10.1620 | **3.0649** |
| Cos Similarity | 0.6428 | **0.7481** | **0.5910** | 0.4605 | 0.6277 | **0.6501** | 0.6303 | **0.7318** | 0.5717 | **0.6807** |
| Style Strength | 73.53 | **76.47** | **58.82** | 32.35 | **70.59** | 61.76 | **67.65** | 55.88 | **52.94** | 52.94 |

Table 7.3: Automatic evaluation results on PP test set. Under each method (Template, Retrieval etc.), the score of the rephraser that performs the best is made bold. The best score obtained for each of BB and GPT3-based rephrasers along each criteria is highlighted in green. Out of them, the best overall score is highlighted with a darker green.

| Criteria | Template | | Retrieval | | Template + Retrieval | | Template + Retrieval (with generic prompting) | | Template + Retrieval (with N-gram prompting) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | BB | GPT3 | BB | GPT3 | BB | GPT3 | BB | GPT3 | BB | GPT3 |
| **Training dataset: PP; Tested on: PP** | | | | | | | | | | |
| Semantic Similarity (SS) | 1.74 | **3.35** | 0.32 | **1.07** | 1.62 | **2.65** | 2.49 | **2.72** | 1.88 | **2.31** |
| Style Transfer Strength (STS) | 2.78 | **3.88** | 0.44 | **2.16** | 2.72 | **3.47** | **3.99** | 3.21 | 2.47 | **3.21** |
| (Average of SS and STS) | 2.26 | **3.62** | 0.54 | **1.62** | 2.17 | **3.06** | **3.24** | 2.97 | 2.18 | **2.76** |
| **Training dataset: PP; Tested on: PPA** | | | | | | | | | | |
| Semantic Similarity (SS) | **2.07** | 0.69 | 0.79 | **0.94** | 2.22 | **2.60** | 2.82 | **2.87** | 2.10 | **2.50** |
| Style Transfer Strength (STS) | 2.51 | **3.70** | 0.65 | **2.00** | 2.61 | **3.17** | **3.96** | 3.14 | 2.26 | **3.02** |
| (Average of SS and STS) | **2.29** | 2.20 | 0.72 | **1.47** | 2.42 | **2.89** | **3.39** | 3.01 | **3.23** | 2.76 |
| **Training dataset: PPA; Tested on: PP** | | | | | | | | | | |
| Semantic Similarity (SS) | 2.63 | **3.19** | **1.21** | 0.81 | 1.69 | **2.57** | 1.74 | **2.53** | 1.21 | **2.32** |
| Style Transfer Strength (STS) | **3.94** | 3.82 | **2.74** | 1.44 | 3.15 | **3.28** | 3.00 | **3.47** | 2.57 | **2.99** |
| (Average of SS and STS) | 3.29 | **3.51** | **1.98** | 1.13 | 2.42 | **2.93** | 2.37 | **3.00** | 1.89 | **2.66** |
| **Training dataset: PPA; Tested on: PPA** | | | | | | | | | | |
| Semantic Similarity (SS) | 2.78 | **3.26** | **1.40** | 1.00 | 1.70 | **2.31** | 1.71 | **2.36** | 1.22 | **2.31** |
| Style Transfer Strength (STS) | **3.92** | 3.82 | **2.30** | 1.92 | 2.59 | **2.85** | 2.60 | **3.06** | 2.40 | **2.98** |
| (Average of SS and STS) | 3.35 | **3.54** | **1.85** | 1.46 | **2.15** | **2.58** | 2.16 | **2.71** | 1.81 | **2.65** |

Table 7.4: Results of human evaluation. Under each methodology (Template, Retrieval etc.), the score of the rephraser that performs the best is highlighted in bold. The best score obtained for each of BB and GPT3-based rephrasers along each criteria is highlighted in green. Out of them, the best overall score is highlighted with a darker green.

Table 7.4 shows the results of the human evaluation experiment. According to the results, GPT3-based rephrasers win over Blender-based rephrasers 70% and 85% of the time along style transfer and semantic similarity dimensions, respectively. And in the smaller PP training corpus, using generic prompting during training increases the scores across most cases. But in the larger PPA corpus, simply training the rephrasers with template-replaced pseudo-parallel pairs gives the best results irrespective of the underlying backbone model.

It was also observed that in 97.5% of the cases, the average scores obtained for style transfer strength are better than the average scores obtained for semantic similarity. This observation is invariant of the type of backbone model used in training. This implies template-based and retrieval-based methods used in creating pseudo parallel data to train the rephrasers make it easier for the rephrasers to generate rephrased sentences that reflect a particular style (in this case, *Advise with permission*) than preserving the semantic meaning of the original sentence. This is a matter to be further investigated. To improve the scores on semantic similarity, future work can explore ways to take into account the context that precedes the sentence to be rephrased. In this way, though the rephrased version may not reflect exactly what was in the original sentence, it might still be able to generate rephrasings relevant to the preceding context.

The average ratings obtained for *style transfer strength* and *semantic similarity* for sentence pairs in the PP test corpus were 3.21 and 3.16, respectively. The sentence pairs in the PPA test corpus scored 3.12 and 2.69 in the above two dimensions, respectively. The average ratings being close to 3 with most of them being above 3 suggests that the training corpora used are of substantial quality.

## 7.8 Chapter Summary

This chapter presented an example on how distress-consoling responses could be boosted with MI strategy. For this, we first developed a classifier that can identify favourable and unfavourable response types as defined by the MITI code. Then we narrowed our focus to the MI non-adherent response type *Advise without Permission* and developed several rephrasers that can rephrase *Advise without Permission* responses into *Advise with Permission.* As curating human written rephrasings was costly, we used templated-based replacement and retrieval methods to create pseudo-parallel corpora from gold-labeled and augmented-labeled MI datasets that contained responses from Reddit and CounselChat platforms. We used this data to train several Blender and GPT3-based rephrasers. We also used generic and N-gram-based prompts to see if prompting can improve the rephrasers' performance. Automatic as well as human evaluation results suggested fine-tuning on GPT3 gives better results in rephrasing *Advise without permission* responses into *Advise with permission.* Data augmentation techniques we used by expanding the MITI labels using N-gram-based matching and similarity-based retrieval improved the performance of the MI classifier as well as the Blender and GPT3-based rephrasers. The results also suggested when the training datasets are small, the use of generic prompting can enable the rephrasing models to produce better results across style transfer and semantic similarity dimensions. But if you are dealing with large datasets (in our case through data augmentation), pseudo-parallel data generated through simpler methods such as template-based replacement can enable the models to generate substantially good rephrasings closer to the required style and semantically similar to the original sentence.

This work can help to boost distress-support responses generated by chatbots trained on peer-support dialogues to be more compliant with the MI strategy. MI boosting can be applied at two different levels: one at the data level; and the other at the model level. At data level boosting, the MI classifier can be applied to automatically label the responses in the training data itself. By doing so, it is able to rephrase the MI non-adherent responses such as *Advise without Permission* into more MI-adherent responses and omit the other unfavourable responses from the training data. The MI-boosted training data can then be used to train the chatbot. At model-level boosting, a similar methodology can be applied at the level the chatbot is decoding responses (e.g. beam search). Not only generative chatbots but also retrieval-based chatbots could be benefited from this methodology. It should also be noted that the application of this work is not limited to improving chatbot responses for distress consolation. This could also be applied for the development of intelligent writing assistants that can suggest better responses when peers untrained in the practice of counseling attempt to respond to distress-related posts on peer support platforms such as Reddit.

# 8 Conclusion

This chapter summarizes our main contributions and the lessons learned, the limitations, and the ethical implications of our work. We also propose some future directions of research and discuss other potential applications of empathetic conversational agents.

## 8.1 Main Contributions

Our study introduced significant contributions to the development of empathetic chatbots. Specifically, the resources and methodologies developed in this thesis enable chatbots to provide empathetic responses to dialogue prompts involving a broad range of emotions ranging from positive and negative emotions experienced in day-to-day situations to extremely negative and distressing emotions. Compared to existing work, the chatbot models we proposed exhibited higher levels of empathy, reliability, safety, and professionalism, which makes them more trustworthy social companions to discuss distress-related concerns. Figure 8.1 breaks this down illustrating our research process and highlighting our individual contributions to the development of empathetic conversational agents. We further summarize them as follows:

1. **Taxonomy of empathetic response intents:** In Chapter 2, we established a taxonomy of empathetic response intents capable of supporting automatic empathetic communication in social chitchat. The strategies described in this taxonomy are useful for a chatbot to engage in prosocial conversations while expressing empathic concern for its users, and keeping the users engaged. We developed it by manually annotating a subset of the responses from EmpatheticDialogues (Rashkin et al., 2019), a state-of-the-art empathetic dialogue dataset. Further, we employed automatic techniques to label the entire EmpatheticDialogues dataset with fine-grained emotions and empathetic response intents and visualized the frequent empathetic conversation patterns seen among humans when engaged in social chitchat and how they temporally vary over the course of the conversation. These results formed the foundation for the development of more reliable and predictable empathetic chatbots.

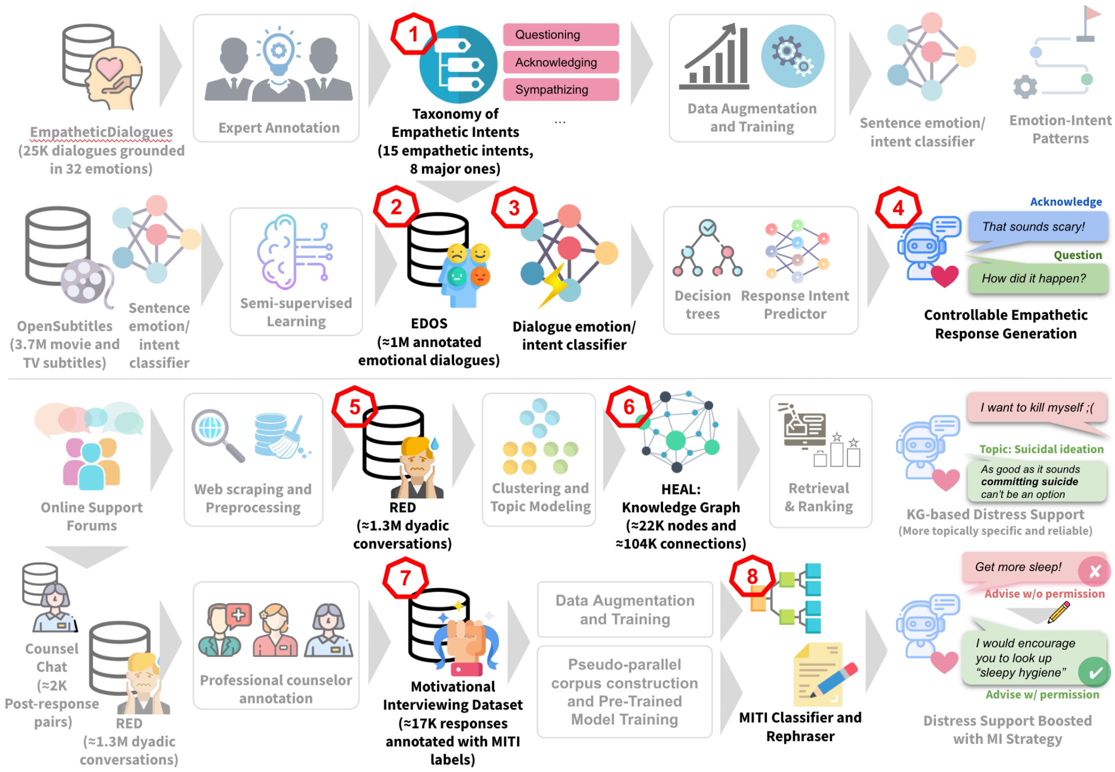



Figure 8.1: Highlights of the research process in generating empathetic responses for distress support. The main contributions along the process are marked by numbers.

2. **EDOS (Empathetic Dialogues in OpenSubtitles) dataset:** In Chapter 3, we developed a large silver-scale emotionally colored dialogue dataset, EDOS, containing ≈1M dialogues annotated with 32 fine-grained emotions, eight empathetic response intents, and the "Neutral" category. We outlined the novel data curation pipeline used to derive this dataset, which involved developing a small seed set of human-annotated data and training a dialogue emotion classifier in a semi-supervised manner to extend the annotations to the entire dataset. This pipeline significantly reduced manual annotation costs and time. Further, we compared the quality of EDOS against the state-of-the-art gold standard EmpatheticDialogues dataset using visual validation and using them in the downstream task of empathetic response generation. Our results showed that the EDOS dataset preserves the conversational dynamics present in the EmpatheticDialogues dataset and it enables conversational models trained on it to generate more diverse empathetic responses.

3. **Fine-grained dialogue emotion classifier:** In the process of curating the EDOS dataset in Chapter 3, we developed a fine-grained dialogue emotion classifier capable of recognizing 32 fine-grained positive and negative emotions, eight empathetic response intent, and the "Neutral" category. It is developed by fine-tuning on RoBERTa (Zhuang et al., 2021). It also takes into account the context preceding the dialogue turn to be classified

and incorporate a special weighting scheme in doing so. It was trained iteratively in a semi-supervised manner by utilizing self-labeling and sentence similarity approaches. It shows a classification accuracy of 65% on the manually labeled EDOS test dataset, which is comparable with the state-of-the-art dialogue emotion classifiers, despite the large set of labels that it can recognize.

4. **Controllable and interpretable empathetic dialogue response generation model:** In Chapter **??**, we made use of the taxonomy of empathetic response intents in addition to fine-grained emotion categories to build a dialogue response generation model capable of generating empathetic responses in a more controllable and interpretable manner compared to neural models trained in an end-to-end fashion. It consisted of two modules: 1) a response emotion/intent prediction module; and 2) a response generation module. We proposed several rule-based and neural approaches to predict the emotion or intent of the next response and used them to condition response generation. Experimental results showed the power of neural response intent prediction over rule-based methods and the usefulness of utilizing the taxonomy of empathetic response intents in generating more diverse and empathetically appropriate responses than end-to-end models.

5. **RED (Reddit Emotional Distress) dataset:** To study human responses catered to dialogue prompts involving extreme negative emotions, psychological distress in particular, in Chapter 5, we curated a large-scale dialogue dataset containing ≈1.3M dyadic conversations scraped from carefully selected eight mental-health related subreddits in Reddit. We described in detail the data curation and the preprocessing process followed to develop the dataset. Due to the large size of the dataset, it can be used to benchmark and test conversational agents developed to respond to psychological distress.

6. **Knowledge-graph HEAL (Healing, Empathy, and Affect Learning):** In Chapter 5, we used automatic clustering and topic modeling applied to the above dataset to develop a knowledge-graph, HEAL, that identifies a multitude of distress-related topics and common responses associated with them, which leads to emotional relief. It consisted 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. We also associated each node with one of 41 affective states identified by our emotion/intent classifier. To show the utility of the knowledge graph, we built a retrieval-based chatbot based on HEAL, and experimental results showed that it can produce responses that are more empathetically appropriate, diverse, and reliable than state-of-the-art empathetic neural response generation models.

7. **Motivational Interviewing (MI) dataset:** To compare the differences between the responses given by peers and trained counselors to dialogue prompts involving psychological distress, in Chapter 6, we developed the MI dataset by recruiting professional counselors to annotate a subset of 1K dialogues from the RED dataset and 1K dialogues

curated from CounselChat (a platform where professional counselors answer distress-related questions posed by users) with labels derived from the Motivational Treatment Integrity (MITI) code, a well-established behavioral coding system used to evaluate therapists' responses. We used a carefully designed annotation pipeline consisting of several stages of quality control to collect these annotations, which enables 97% of label coverage with moderate agreement. Using the annotated data, we made comparisons between the responses of peers and professionals and make recommendations on making chatbot responses trained on peer-support dialogues sound more professional and compliant with the MI strategy.

8. **MI classifier and MI rephraser:** As an attempt to recognize responses that conform or non-conform to the MI strategy, and rephrase the non-conforming responses into a more MI-adherent form, in Chapter 7, we introduced an MI classifier that is able to recognize 15 types of labels adapted from the MITI code and an MI rephraser that can rephrase the MI non-adherent response type *"Advice without permission"* into more MI-adherent *"Advice with permission"*. Our classifier, fine-tuned on RoBERTa (Zhuang et al., 2021) achieved a classification accuracy of ≈73% on the MI test dataset. We built several rephrasers by fine-tuning pre-trained language architectures Blender (Roller et al., 2021) and GPT3 (Brown et al., 2020b) on pseudo-parallel corpora constructed using automatic methods. Through automatic and human evaluation, we showed that in the absence of human-constructed parallel corpora, the construction of pseudo-parallel corpora, and the use of techniques such as prompting and data augmentation can enable text rephrasing models to produce substantially good rephrasings that reflect the intended style and preserve the content of the original text. Finally, we described how the MI classifier and the MI rephraser can be used in a pipeline to boost chatbot responses trained on peer-support dialogues to be more compliant with the MI strategy, both at the data level (data level boosting) and at the model level (model level boosting).

## 8.2 Lessons Learned, and Limitations

Throughout the thesis, we have made significant contributions to the development of empathetic dialogue agents capable of providing support during distressful situations with an emphasis on making these models respond to a broad range of emotions and making them more reliable, safe, and professional. The work that we conducted can be categorized into five main topics: emotion and intent recognition; data resources; dialogue models; framework for advancing distress support; and human computation. We would further like to elaborate on some lessons learned and limitations that exist under each of these topics.

**Emotion and Intent Recognition:** For dialogue response generation models to be able to respond to a wide range of positive and negative emotions, it requires them to identify subtle variations of emotions and empathetic response strategies used to respond to these different types of emotions. Towards this end, we introduced a dialogue emotion classifier that can recognize 32 fine-grained positive and negative emotions, 8 empathetic response intents, and *Neutral*. The accuracy reported by this classifier is on par with the state-of-the-art dialogue emotion classifiers considering the wide range of labels that they can classify and hence can be well utilized in experiments involving empathetic response generation where identification of emotions and specific intents are necessary. From our experiments that utilized this classifier, we learned that explicit recognition of fine-grained emotions and intents present in dialogue utterances makes the responses more empathetically appropriate than traditional end-to-end response generation methods. Further, combined with response emotion and intent prediction methods, they can bring more controllability and interpretability to the dialogue responses generated. However, there is a limitation to this classifier that it is unable to automatically recognize chunks in the response (several sentences combined together) that can be classified into one label. This is a limitation in most of the existing state-of-the-art dialogue emotion recognition models and hence future work can work on further improving this aspect. The development of this classifier further facilitated the automatic curation of emotional dialogue datasets and controllable and interpretable empathetic response generation approaches.

**Data Resources:** The EmpatheticDialogues dataset (Rashkin et al., 2019), which is a gold standard dataset curated by humans for empathetic response generation, served as a benchmark dataset in most of our experiments. In this thesis, we introduced the EDOS, RED, and MI datasets to further facilitate the development of empathetic conversational agents. The EDOS and RED datasets each contain ≈1M empathetic conversations. Due to their large size, they cover a wide variety of scenarios and contain more diverse responses than existing datasets. These datasets are automatically annotated with fine-grained emotions and intents and hence can serve as silver-standard benchmarks in future research involving empathetic and distress support chatbots. We learned that when the datasets become larger in size, covering a variety of different situations and responses, the chatbot models trained on them have the tendency to produce responses richer in diversity than when trained on datasets that are limited in size. We also learned that, even though they are silver standard datasets that use automatic methods for dialogue curation and annotation, with sufficient quality control they can serve equally well in generating effective empathetic responses compared to the state-of-the-art gold standard empathetic dialogues datasets. However, still, some limitations may exist such as certain biases and fairness issues in the data collected in this manner. Further preprocessing and filtering techniques should be developed in the future to make such silver standard data more usable. Overall, the EDOS and RED datasets served in training more controllable and interpretable empathetic response generation models for daily chitchat like conversations and high-stake distress-related conversations, respectively. Further, the MI dataset that we introduced and the differences in empathetic support patterns between professionals and peers recognized from this dataset, served as a foundation for a framework for boosting distress support responses generated by chatbots trained on peer-support data.

**Empathetic Chatbot Models:** We introduced several generative and retrieval-based empathetic and distress-support chatbot models throughout this thesis. They were developed giving a special focus on making the responses more controllable, interpretable, and thus more reliable. We used fine-grained emotions and our taxonomy of empathetic responses to control and interpret the responses generated by these models. We introduced decision tree-based and neural response emotion and intent prediction methods that can be used to steer the response generation process in predictable routes, with the latter outperforming the former. We also explored the ability to represent whole dialogues with relations between context-response pairs in a form of a knowledge-graph so that responses could be reliably retrieved given a dialogue prompt that is strongly associated with a graph node. We learned that these methods for controllability enable chatbot models to generate more empathetically appropriate responses and play a vital role in preventing unfavourable and toxic responses from being generated. This is crucial especially when addressing extremely negative emotions when the user is more sensitive to inappropriate comments. However, there are limitations associated with the methods that we have proposed. For example, the response emotion and intent prediction accuracy in our controllable dialogue models is still below 20%, and traversing the knowledge-graph and retrieve responses, and reranking them requires more efficient methods. Future work can focus on developing more advanced methods for response intent prediction and for the incorporation of knowledge graphs in steering the response generation process.

**Framework for Advancing Distress Support:** The final development in our thesis is the introduction of the MI classifier and the MI rephraser. They are components instrumental in the pipeline that serves as a proof of concept on how distress support responses provided by chatbots, which are most often trained on responses given by laypersons, could be boosted to reflect more the language style of therapists and make these responses more adherent to established norms in the psychological literature. In the development process, we learned that the creation of pseudo-parallel corpora is an effective solution when training rephrasing models in the absence of manually constructed parallel corpora. We also learned that in the presence of less training data, techniques such as prompting and data augmentation can largely boost the performance of the rephrasers. However, there are some limitations to the proposed methodology. Certain parts of it, for example, template-based replacement and n-gram-based prompting are applicable when style-specific linguistic attributes could be identified between the source and the target style. And due to the cost of human labor and the lack of publicly available client-therapist dialogues, the sample sizes drawn in our experiments are small and thus may have an impact on the conclusions drawn. Future work can develop upon this idea and build more sophisticated rephrasing and response-boosting methodologies.

**Human Computation:** Under each chapter, we conducted at least one human computation experiment for the purposes of annotating and evaluating dialogue responses. We utilized crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) and UpWork to recruit human workers for our experiments. In UpWork, the workers are more visible, and communication with them is possible over UpWork's chat platform. In contrast, MTurk workers are only identifiable through an identifier, and communication with them is not as straightforward as UpWork. These reasons may seem to have an impact on the quality of the work received. In MTurk we have to be more explicit in the instructions we provide since later clarifications are almost impossible. In addition, due to the invisibility of the workers, more serious quality control mechanisms and rigorous data filtering methods need to be deployed in MTurk compared to UpWork. These mechanisms were able to detect bots and inattentive workers. We introduced mechanisms such as the inclusion of detailed tutorials with concrete examples, the inclusion of hidden quiz checkpoints, and providing bonuses for those who score well on quiz checkpoints to address the above concerns. Choosing workers based on their percentage of completed and accepted hits and their demographics, and data filtering based on the time spent on the task, and scores obtained in quiz questions further contributed to obtaining high-quality annotations. Still, we observed some inter-rater agreement scores in the lower acceptable range in some of the experiments conducted on MTurk, which indicates that more sophisticated quality control and data filtering methods should be deployed in such cases. Moreover in Chapter 6, we introduced a human computation annotation pipeline that was carefully designed to improve the inter-rater agreement over several stages. The methods developed in this thesis can be readily applied in similar human computation experiments. Since these experiments consume much time and cost, future work can focus on developing means to automate such processes as much as possible.

## 8.3 Ethical Implications

There are ethical implications associated with the development of empathetic conversational agents for distress support in general as well as with some methodologies used in our research.

**Data Curation:** In this thesis, we analyze posts taken from social media websites such as Reddit. In social sciences, analysis of such posts is most likely considered “fair play" because individuals are anonymous, and users most of the time understand their responses remain archived on the site unless they take action explicitly to delete them. The Reddit privacy policy states it allows third parties to access public Reddit content through the Reddit API and other similar technologies and users should consider that before posting.[1] And Reddit data is already widely available in larger dumps such as Pushshift (Baumgartner et al., 2020). We collected only publicly available data on Reddit and the curation process did not involve any intervention or interaction with the Reddit users. However, a study on user perceptions on social media research ethics (Fiesler and Proferes, 2018) highlights some potential harms that can be caused due to social computing research due to the fact that internet users rarely read or could fully understand website terms and conditions and are unaware that the data they share publicly could be used for research and analysis. In particular, our data contains sensitive information such as mental health diagnoses and personal attributes. Hence, as suggested by Benton et al. (2017)'s guidelines for working with social media data in health research, in this thesis, we share only anonymized and paraphrased excerpts from the dataset so that it is not possible to recover usernames through a web search with the verbatim post text.

[1] www.redditinc.com/policies/privacy-policy-october-15-2020

The RED dataset contains the original post texts but with anonymized usernames and post identifiers. In addition, references to usernames as well as URLs are removed from dialogue content for de-identification.

**Use of Pre-trained Language Models:** We used large-scale pre-trained language models such as RoBERTa (Zhuang et al., 2021), Blender (Roller et al., 2021), and GPT-3 (Brown et al., 2020a) when developing empathetic chatbots in most of our experiments. Henderson et al. (2018) highlight several ethical concerns that arise in dialogue systems relying on data-driven approaches such as the use of language models. They mainly emphasize the implicit biases in data-driven systems, the vulnerability of these models to adversarial examples, potential for privacy violations, and concerns related to reproducibility. Weidinger et al. (2022) propose a taxonomy of ethical and social risks associated with language models. For example, there are fairness and toxicity risks associated with language models due to training on corpora that overrepresent certain social identities and include harmful language. Such data can cause models to have stereotypes and social biases, which can hurt minority social groups. Toxic language can also incite hate and can be offending. They may also provide false or misleading information because the underlying statistical methods cannot distinguish between factually correct and incorrect information. Such misinformation can cause harm, specially in sensitive domains such as mental health support. Hence real-world deployment of agents based on these models should be done with caution and with human supervision. We refer to the study by Weidinger et al. (2021) for a more detailed analysis of the ethical implications associated with large language models.

**Making Chatbots Human-like:** Embedding emotion recognition and empathetic capabilities in chatbots contribute to making chatbots more human-like (Svikhnushina and Pu, 2022). However, there is much speculation regarding human-like chatbots. Human likenesses can be deceptive and there is a risk that users may get emotionally attached to them over time. This may make humans open up to chatbots, which may cause harm if they are designed in manipulative ways or to collect users' personal information. To address these concerns, the IEEE Standards Association (2016) has taken the initiative to provide guidelines to ensure the safety and well-being of humans, for example, being transparent about the system's affective nature, and being considerate about the risk of causing emotional harm through manipulation, dampening, or amplifying users' emotions. Another study conducted by Vanderlyn et al. (2021), talks about the tension between making a chatbot seem more human and the ethical implications associated with doing so. By making a chatbot seem empathetic, there is an ethical concern that the users may assume that it is capable of actually understanding their emotional needs. The study evaluates the users' perceptions of human-like chatbots and concludes that such human-like traits are not inherently problematic so long as transparency about the artificial nature of the dialogue system is ensured. Thus, it is advisable to always follow such recommendations when these agents are deployed in the real-world.

The current literature on Fairness, Accountability, Transparency, and Explainability (FATE) takes rather a condemnatory position regarding the attribution of human-like characteristics in conversational agents, commonly referred to by the term *anthromorphism*. Abercrombie et al. (2023) states that anthromorphism in dialogue systems can lead to a breach of transparency and trust, leading to over-reliance on the outputs produced by these systems. These consequences can be particularly true when considering vulnerable population groups such as children, the elderly, and people with illnesses or disabilities. Simulation of empathy and emotive language as we describe in this thesis further anthropomorphise dialogue systems. Thus, we emphasize that while there is significant motivation to develop human-like dialogue systems, practitioners should reflect on how this technology can affect people's understanding and behaviour.

**Chatbots for Distress Support:** The idea of making use of chatbots for distress support is not a new concept. Chatbots such as SimSensei (DeVault et al., 2014), Dipsy (Xie, 2017), Emma (Ghandeharioun et al., 2019), Woebot (woebothealth.com), and Wysa (www.wysa.io) are some examples. As Czerwinski et al. (2018) state, *About 1 billion people globally are affected by mental disorders; a scalable solution such as an AI therapist could be a huge boon.* Thus, even though empathetic and distress support chatbots may encompass certain ethical implications as pointed out by several researchers (Lanteigne, 2019; Montemayor et al., 2021; Tatman, 2022), based on previous studies, we already can acknowledge that the use of chatbots has the potential to improve mental health services notably in relation to accessibility and anonymity. However, there are many potential risks associated with such agents. For example, in the responses produced by these agents, there can be risks of revealing sensitive information, though anonymized, related to someone's personal struggles that are present in or can be inferred from the training data. With respect to the distress support responses that we generate in this thesis, there is still the risk of providing misinformation since the models are trained on data scraped from online support forums without any validation of the correctness of the information present in them. Not mitigating gender stereotypes present in the training data can further exacerbate concerns with queer communities undergoing distress.

It is also important to note that chatbots are not a replacement for therapy, and their capabilities and limitations should be taken into consideration when deploying such agents in the real-world. Kretzschmar et al. (2019), investigate social and ethical concerns that arise from conversational agents for mental health support and outline several recommendations when deploying them in the real-world. They recommend continuous assessment of harms and benefits associated with them to allow a timely response to unpredicted issues or concerns. They emphasize that such chatbots should be as transparent as possible about what they are currently able to offer and should have systems built in place to encourage users to prevent over-reliance and deal with emergency situations. Chatbots can provide helpful support, but they are not able to provide the same level of in-depth understanding, empathy, and human connection that a therapist can provide. Therefore, it is important to encourage users to seek professional help when needed and to use chatbots as a supplement to, rather than a replacement for, therapy.

## 8.4 Future Directions

Some promising future directions can be outlined in the area of empathetic and distress-support conversational agents. We present here only general perspectives for future work. Specific suggestions for continuing our research are presented at the end of each chapter.

**Prompt-based empathetic response generation:** Research about pre-trained language models (OpenAI, 2023; Touvron et al., 2023; Chowdhery et al., 2022; Zhang et al., 2022) has advanced task performance related to dialogue response generation. So, instead of training neural models on large datasets or fine-tuning them on smaller domain-specific datasets, nowadays research is more focussed on designing intelligent prompts to instruct such models to respond in specific ways. The advantage here is that it requires only a few training examples or in certain cases none, allowing few-shot and zero-shot learning. Recent work has successfully managed to use prompting for dialogue response generation (Madotto et al., 2021; Gu et al., 2021; Zheng and Huang, 2021). It is also seen that large language models such as GPT-3.5 and GPT-4 have empathetic capabilities such as cognitive empathy. Kosinski (2023) states that such capabilities may have emerged spontaneously as a byproduct of training the language models to achieve other goals instead of explicitly engineering them into these models. Future work can explore effective prompt designs that can navigate the empathetic capabilities of these large language models and instruct them to carry out supportive and engaging empathetic conversations with users.

**Fairness and debiasing:** Incorporating fairness and debiasing into empathetic and distress support chatbot models is another promising direction for future research. Since empathetic chatbots are deisgned to provide emotional assistance and support for the users, it is vital that they are designed in a way that their responses are fair and free from bias. Since chatbot models are trained on large corpora that overly represent certain social groups and certain social norms, it is natural that these biases are present in the responses generated by these chatbot models. Thus, explicit steps should be taken to avoid biases and discrimination against particular user groups. Standardized methods should be introduced to ensure that chatbots provide the same level of support and empathy towards all users regardless of their demographic characteristics, such as gender, age, race, or socio-economic status.

**Cultural sensitivity:** There are certain socio-cultural differences when it comes to empathetic responding. People having different cultural backgrounds may have different ways of expressing emotions and different expectations of receiving empathy. For example, some cultures may be more inclined to self-disclose emotional experiences, while some other cultures may be more emotionally restrained and value privacy. However, these intercultural differences are largely understudied, and current empathetic and distress-support chatbot models do not take this into consideration. Thus, it is important to incorporate cultural understanding of empathy into the design of chatbot models. It is also important to design and develop evaluation methods that include a diverse range of cultural perspectives when testing these models to ensure that they are effective across different cultures.

**Privacy and safety:** Another important future direction would be to devise methodologies to address privacy and safety concerns associated with empathetic conversational agents. Even though these agents have great potential to benefit society, privacy, and safety concerns create speculations when deployed in the real world. One main concern is that these agents could be used in a way to manipulate or influence vulnerable individuals with mental health issues or cognitive impairments for various things such as the revelation of their personal information. Thus, it is important to ensure that these agents are not designed in a way to take advantage of such individuals. Also, conversations with empathetic agents may contain personal and sensitive information, which should be handled in a secure and responsible manner. Thus, data protection mechanisms should be developed to ensure that the data disclosed by users are protected from unauthorized access or misuse.

## 8.5 Applications

The application of empathetic chatbots incorporating MI techniques that we explored in this thesis holds significant potential beyond distress support. MI is about empowering and influencing people to make positive changes in their behavior that improve their overall well-being. It uses certain principles and techniques to build rapport and influence one's thoughts toward making positive change taking a more client-centered and empathetic approach. Understanding and modeling social influence dynamics like MI are important when designing AI systems that can effectively engage with and influence people toward making meaningful behavioral changes. Psychological research suggests that MI can be applied in various other settings such as substance use (Baker et al., 2002; Smedslund et al., 2011), behavioral addiction (e.g., gambling, shopping, internet, video games, risky behaviors etc.) (Weegmann, 2002; Baer et al., 2004), and procrastination (Rozental et al., 2014; Hosseini et al., 2020). Thus, these chatbots can be tuned as motivational coaches that can help people overcome such issues and make positive changes in their lives. These chatbots can employ non-intrusive persuasive strategies such as giving advice with permission along with tailored interventions to motivate individuals towards positive behavior change. Also, empathetic conversational agents in general can be used for various other applications such as enhancing customer service experiences by understanding and responding to customers' emotions and needs, acting as social companions for people who suffer from loneliness, and in education settings adapting their responses to individual students' emotional states. Overall, such applications leveraging empathetic responding and motivating capabilities can empower and positively influence individuals, promoting overall well-being and personal growth.

# A Appendices

## A.1 Words and phrases indicative of the empathetic response intents

Table A.1 shows the words and phrases most indicative of the empathetic response intents that were used to extract more example listener utterances from the EmpatheticDialogues dataset for training the BERT transformer-based classifier.

| **Response intent** | **Words and phrases most indicative of the intent** |
|---|---|
| Agreeing | *100%, exactly, absolutely, definitely, agree, i know, me either, me neither, i understand, i completely understand, me too, that's right, you're right, correct* |
| Acknowledging | *it sucks, that sucks, i'd ... too, i would ... too, i feel you, that's splendid, i bet ... was, that's great", that's a good idea, i bet ... can't, that's pretty, i see, it's pretty, can understand, sounds, that would, i would have, must've, cool, nice, awesome* |
| Encouraging | *hopefully ... will, i hope ... will, works out for you, i bet ... will, i bet ... 'll, i bet ... can* |
| Consoling | *there you go, hopefully ... will, i hope ... will, cheer up, get better, will pass quickly* |
| Sympathizing | *i'm sorry, sorry to hear, oh no, bless you, deepest sympathy* |
| Suggesting | *maybe, i think ... should, perhaps, why don't you, you could always, what if* |
| Questioning | *what ... ?, why ... ?, when ... ?, where ... ?, how ... ?, are ... ?, is ... ?, did ... ?, do ... ?, does ... ?, have ... ?, has ... ?, had ... ?* |
| Wishing | *congratulations, happy birthday, happy anniversary, i wish you, wish you ... !, all the best, good luck* |

Table A.1: Words and phrases that are most indicative of the empathetic response intents.

## A.2 Example speaker-listener utterance pairs corresponding to the taxonomy of emotion/intent exchanges

| **Speaker's emotion** | **Listener's response emotion/intent** | **Example utterance-response pairs** |
|---|---|---|
| Anticipating | Questioning | S: *When tax season came I was in a hurry to get mine done. I was looking forward to a big refund.* (Anticipating)<br>L: *really? why is that?* (Questioning) |
| | Acknowledging | S: *I cannot wait for the newest Pokemon game, it looks amazing to me!* (Anticipating)<br>L: *Those games do seem fun* (Acknowledging) |
| Joyful | Questioning | S: *i was happy to see that i was able to get a new pet the other day* (Joyful)<br>L: *What pet did you get?* (Questioning) |
| | Acknowledging | S: *I jumped for joy when my baby was born.* (Joyful)<br>L: *wow that must have been a huge moment for you* (Acknowledging) |
| Trusting | Questioning | S: *Man, I let one of my friends take my Benz one day to run some errands. I really thought she would be careful with it.* (Trusting)<br>L: *Oh, no! Did she damage your car?* (Questioning) |
| | Acknowledging | S: *My therapist was so kind to me, I had to tell her a lot.* (Trusting)<br>L: *That's good you have someone that you can talk to about your problems and feelings. I'm sure it helps!* (Acknowledging) |
| Surprised | Questioning | S: *I was shocked when i got invited on a random trip* (Surprised)<br>L: *Was a happy shocked feeling or a bad one?* (Questioning) |
| | Acknowledging | S: *The other day I found out that my sister is having twins!* (Surprised)<br>L: *Oh that's wonderful twins seem really cool.* (Acknowledging) |
| | Neutral | S: *No one even knew she was dating anyone until the announcement, so I was very surprised.* (Surprised)<br>L: *I guess she wanted to keep it a secret for some reason.* (Neutral) |

| | | |
|---|---|---|
| Afraid | Questioning | S: *It's so dark and creepy down there.* (Afraid)<br>L: *lol. Do you think there are monsters down there?* (Questioning) |
| | Acknowledging | S: *It was only off for a little over 2 hours, but I could not find a flashlight and it was so scary.* (Afraid)<br>L: *That sounds awful!* (Acknowledging) |
| Sad | Questioning | S: *I feel bad I don't always get to go through bad things and full get healed.* (Sad)<br>L: *Do you mean you feel bad that you don't get to go through bad things or that you don't get to be healed?* (Questioning) |
| | Sympathizing | S: *I was extremely emotional when my dog passed away* (Sad)<br>L: *Aww man sorry for your loss, those are the worst.* (Sympathizing) |
| | Acknowledging | S: *My favorite donut shop went out of business.* (Sad)<br>L: *Ah that's a pity. It really sucks to lose favorite shops.* (Acknowledging) |
| | Agreeing | S: *I'm sad. My youngest son starts kindergarten tomorrow!* (Sad)<br>L: *I am sure it is a bittersweet moment. I can relate myself.* (Agreeing) |
| Disgusted | Questioning | S: *I am disgusted that so many people voted in favour of Brexit in the UK.* (Disgusted)<br>L: *Why is that?* (Questioning) |
| | Acknowledging | S: *It was a brand new box of Rice crispies. When I opened it and poured it in my bowl, there were several live bugs.* (Disgusted)<br>L: *Well that sounds disgusting* (Acknowledging) |
| | Disgusted | S: *I was at Mcdonalds and was given a rotten cheese burger. I almost puked after I ate it.* (Disgusted)<br>L: *Oh gross, makes me never want McDonalds again.* (Disgusted) |
| | Agreeing | S: *Everytime I see my cat vomit on floor it makes me sick.* (Disgusted)<br>L: *i think you have the same attitude like me.* (Agreeing) |
| Angry | Questioning | S: *i was upset when i saw someone put a dent in my door* (Angry)<br>L: *Was this a parking lot?* (Questioning) |
| | Acknowledging | S: *My grandma didn't make my oatmeal right yesterday. I was so mad.* (Angry)<br>L: *Oh wow! You were pretty angry* (Acknowledging) |

Table A.2: Example speaker and listener utterances corresponding to the most common emotion exchanges between speakers and listeners, when the speaker's emotion is one of the Plutchik's 8 basic emotions.

## A.3 Computing the readability of OS dialogues

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

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

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

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

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

## A.4 AMT task interface for curating EDOS

The user interface used to collect labels from the AMT workers is denoted in Figure A.1.

## A.5 Choice of hyper-parameters and additional training details regarding the dialogue emotion classifier used to annotate the EDOS dataset

We used the same hyper-parameter setting used in RoBERTa (Liu et al., 2019) when training the dialogue emotion classifier used for annotation. We used the Adam optimizer with $\beta_1$ of 0.9, $\beta_2$ of 0.98, an $\epsilon$ value of $1 \times 10^{-6}$, and a learning rate of $2 \times 10^{-5}$. A dropout of 0.1 was used on all layers and attention weights, and a GELU activation function (Hendrycks and Gimpel,

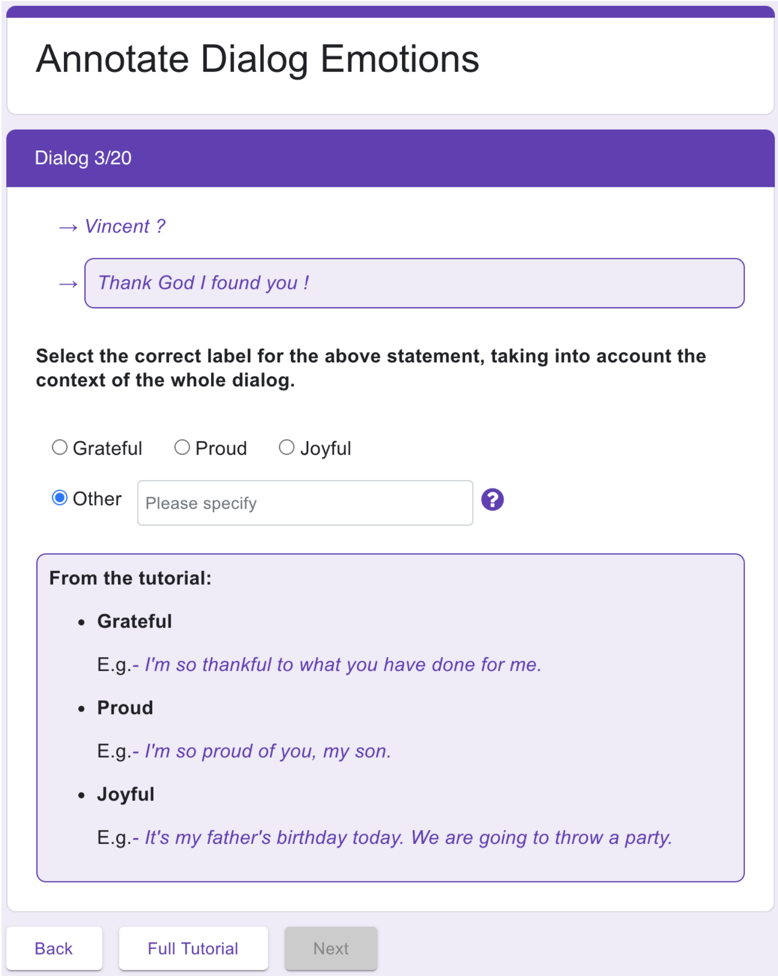



Figure A.1: The user interface of the AMT crowd-annotation task.

2016). We limited the maximum number of input tokens to 100, and used a batch size of 256. All the experiments were conducted on a machine with 2x12cores@2.5GHz, 256 GB RAM, 2x240 GB SSD, and 2xGPU (NVIDIA Titan X Maxwell). In total, 546.84 seconds were taken to train the final emotion classifier. The optimal model was selected based on the average cross entropy loss calculated between the ground-truth and predicted labels of the validation set.

## A.6 Detailed statistics of the EDOS dataset

Table A.3 shows more descriptive statistics of the EDOS dataset: the number of dialogues; and the number of dialogues turns per emotion and intent category. A dialogue is counted under an emotion or an intent if the beginning dialogue prompt is annotated with that emotion or intent.

| Emotion or Intent | No. of dialogues | No. of turns |
|---|---|---|

| | | |
|---|---|---|
| Prepared | 21,178 | 48,883 |
| Anticipating | 27,256 | 100,433 |
| Hopeful | 21,328 | 54,012 |
| Proud | 13,910 | 33,365 |
| Excited | 22,118 | 53,756 |
| Joyful | 6,586 | 24,282 |
| Content | 20,688 | 64,569 |
| Caring | 13,599 | 42,806 |
| Grateful | 15,416 | 42,222 |
| Trusting | 41,650 | 134,197 |
| Confident | 26,199 | 84,918 |
| Faithful | 8,095 | 25,029 |
| Impressed | 12,867 | 25,045 |
| Surprised | 16,658 | 46,022 |
| Terrified | 9,449 | 28,730 |
| Afraid | 15,964 | 49,285 |
| Apprehensive | 8,634 | 46,727 |
| Anxious | 2,376 | 8,578 |
| Embarrassed | 11,541 | 32,338 |
| Ashamed | 3,401 | 14,797 |
| Devastated | 6,245 | 17,539 |
| Sad | 23,023 | 66,262 |
| Disappointed | 5,234 | 18,298 |
| Lonely | 3,662 | 16,396 |
| Sentimental | 7,104 | 20,715 |
| Nostalgic | 7,880 | 20,461 |
| Guilty | 9,632 | 30,043 |
| Disgusted | 5,546 | 15,070 |
| Furious | 54,647 | 169,917 |
| Angry | 13,228 | 34,924 |
| Annoyed | 6,637 | 30,072 |
| Jealous | 5,766 | 20,902 |
| Agreeing | 20,173 | 96,562 |
| Acknowledging | 39,781 | 138,165 |
| Encouraging | 3,024 | 10,329 |
| Consoling | 3,785 | 17,256 |
| Sympathizing | 15,557 | 38,774 |
| Suggesting | 42,470 | 101,591 |
| Questioning | 357,255 | 841,556 |
| Wishing | 42,789 | 108,668 |
| Neutral | 7,649 | 55,932 |
| Total | 1,000,000 | 2,829,426 |

Table A.3: Descriptive statistics of the EDOS dataset pertaining to each emotion and intent category.

## A.7 Additional training details about the experimental baselines used to test the EDOS dataset

Here we summarize some of the parameters of the model implementation. We used the RoBERTa tokenizer to tokenize the input utterances, and the vocabulary size is 50,265. We allow a maximum number of 100 tokens as the input to the model. We used 4 sub-layers in the encoder and decoder, with 6 heads in the multi-head attention. The dimension of the hidden units is 300, and the dimension of the pointwise feed-forward layers is 1200. We use a dropout rate of 0.1, and the GELU (Hendrycks and Gimpel, 2016) activation function for the hidden layers. The loss function was optimized with the Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 5 × 105 For inference, we use beam search with a beam size of 32. To prevent the models from generating repetitive tokens or n-grams, we modified the beam search algorithm so that at each time step, if any of the branches contains repetitive 4-grams, we set the log probability of this branch to infinitely negative, to stop it from being further expanded. All the models were trained with a batch size of 512, on machines with 4 Nvidia Titan X Pascal GPUs, 2 Intel Xeon E5-2680 v3 CPUs, and 256GB RAM. Table A.4 lists the training details as well as the validation performance for all the models.

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

Table A.4: Training details and validation performance of each model configuration.

## A.8 Hyper-parameters used and additional training details of the two modules in the controllable and interpretable empathetic response generation architecture

The BERT-base architecture with 12 layers, 768 dimensions, 12 heads, and 110M parameters is used as the representation network in both the neural response intent predictor and the response generation model. It is initialized with weights from the pre-trained language model RoBERTa (Zhuang et al., 2021), which is proven to perform better than pre-trained BERT. We used the same hyperparameter setting used in RoBERTa. We used the Adam optimizer with $\beta_1$ of 0.9, $\beta_2$ of 0.98, an $\epsilon$ value of $1 \times 10^{-6}$, and a learning rate of $5 \times 10^{-5}$. A dropout of 0.1 was used on all layers and attention weights, and a GELU activation function (Hendrycks and Gimpel, 2016a). We limited the maximum number of input tokens to 100, and used a batch size of 256. All the models were first trained on the OS dialogues dataset for 50 epochs and fine-tuned on the EmpatheticDialogues and OSED datasets for 10 epochs each. The best model was chosen based on the minimum loss computed on the validation set after each epoch. The number of training epochs taken for each of the models to converge is denoted in Table A.5. All the experiments were conducted on a machine with 2x12cores@2.5GHz, 256 GB RAM, 2x240 GB SSD, and 2xGPU (NVIDIA Titan X Maxwell).

| Model setting | No. of training epochs to converge |
|---|---|
| **Neural response emotion and intent predictor** | |
| OS | 2 epochs (pre-training) |
| OS + ED | 1 epoch (fine-tuning) |
| OS + EDOS | 1 epoch (fine-tuning) |
| **Response Generator** | |
| End-to-end (OS) | 50 epochs (pre-training) |
| End-to-end (OS + ED) | 9 epochs (fine-tuning) |
| End-to-end (OS + EDOS) | 3 epochs (fine-tuning) |
| End-to-end + Neural Predictor (OS) | 50 epochs (pre-training) |
| End-to-end + Neural Predictor (OS + ED) | 9 epochs (fine-tuning) |
| End-to-end + Neural Predictor (OS + EDOS) | 3 epochs (fine-tuning) |

Table A.5: The number of training epochs taken for the models to converge

## A.9 Additional details of the AMT human evaluation experiment conducted to evaluate the responses generated by the controllable and interpretable empathetic response generation model

Table A.6 indicates the statistics of the Amazon Mechanical Turk experiment conducted to evaluate responses generated by different models. As further measures to control the quality of the workers, we monitored the total time spent on a particular task and reject work that was completed in less than two minutes. And to prevent workers from monopolizing the task by accepting a large number of HITs, we warned and blocked the workers who accepted more than 30 HITs.

| Description | Statistics |
|---|---|
| No. of HITs | 200<br>(10 dialogues each) |
| No. of assignments | 400 (200 × 2) |
| No. of workers | 160 |
| Percentage of bonuses earned | 91.67% |
| No. of blocked workers (Either due to completing more than 30 HITs or due to submitting an assignment in less than 2 minutes | 9 |

| | |
|---|---|
| No. of rejected assignments (due to time being < 2 min.s) | 9 |
| Percentage of 2 out of 3 worker agreements on ratings | 88.43% |
| Interrater agreement (Fleiss' kappa) | 0.2294 (fair agreement) |

Table A.6: Statistics of the AMT human evaluation experiment.

Figures A.2 and A.3 show the guidelines provided to the crowd workers and the AMT task interface design, respectively.

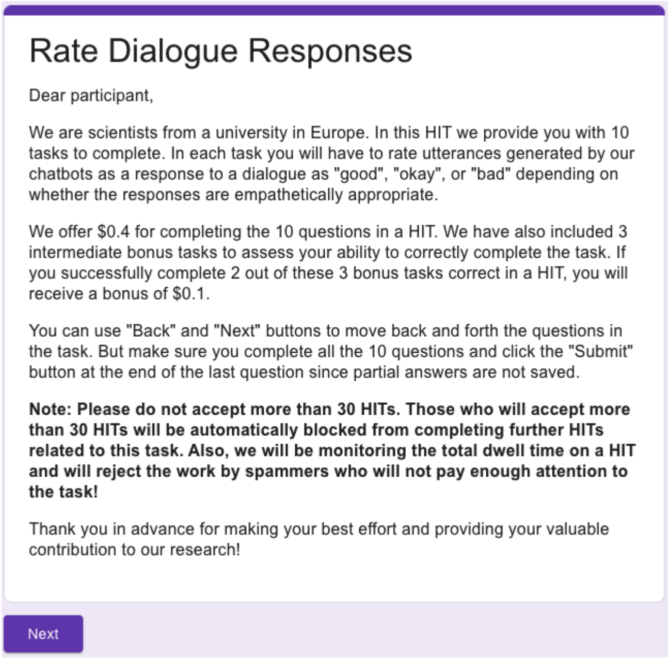



Figure A.2: Guidelines of the AMT task for dialogue response evaluation.

## A.10 Cluster quality metrics computed to determine the optimal clustering thresholds for identifying different nodes in HEAL

One of the crucial decisions to make during clustering was the selection of the clustering threshold. For this purpose, we computed a number of statistical and cluster quality metrics such as the number of resulting clusters, the Silhoutte coefficient Rousseeuw (1987), Dunn index Misuraca et al. (2019), and average point-to-centroid cosine distance for each threshold. For example, Table A.7 shows these values for clustering distress narratives. Figure A.4 plots the cluster quality metrics computed over different similarity thresholds to cluster distress narratives.

As two of the cluster quality metrics, the Silhouette coefficient and average point-to-centroid

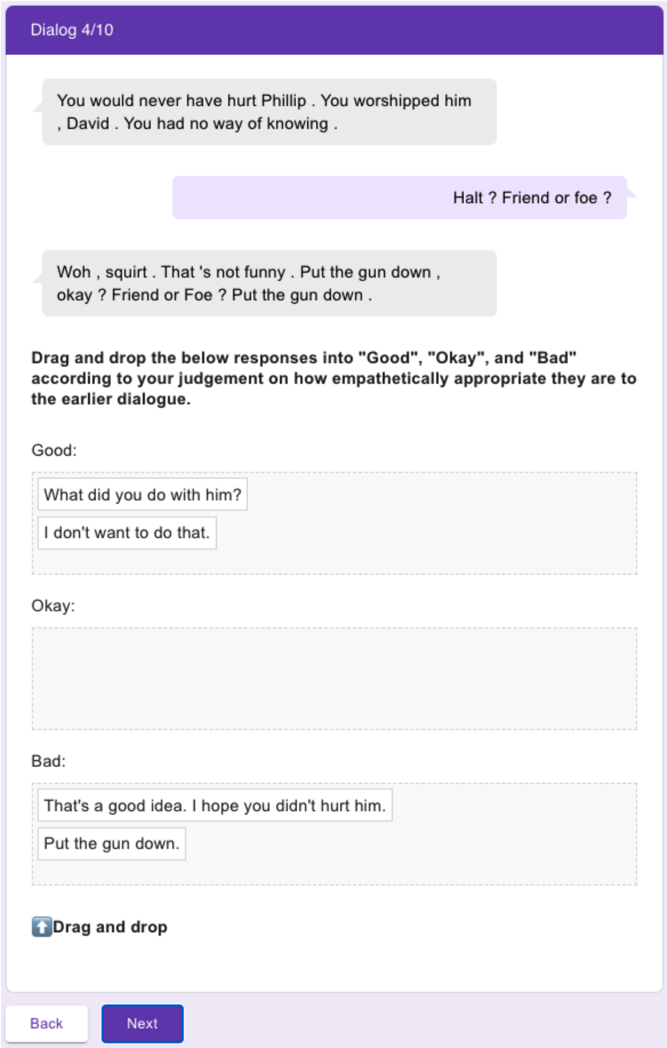



Figure A.3: AMT user interface designed for evaluating dialogue responses.

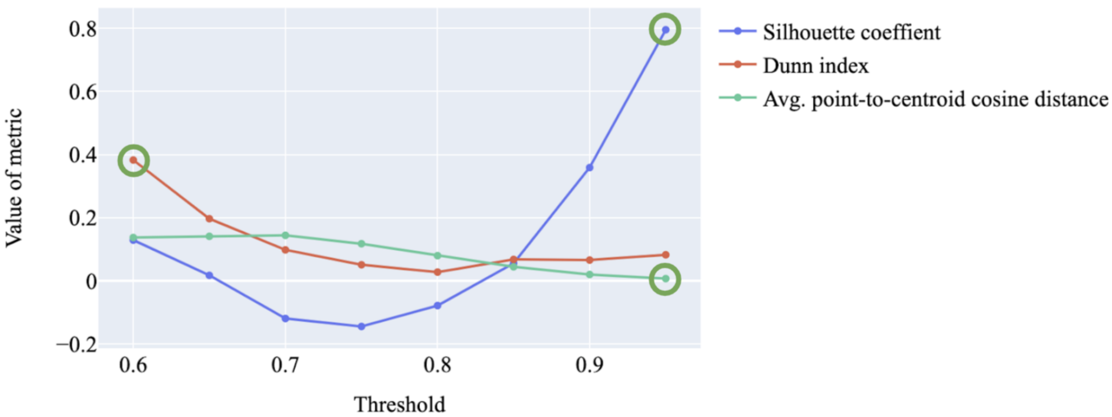



Figure A.4: Plot of cluster quality metrics computed over different similarity thresholds for clustering distress narratives.

| Threshold | No. of clusters | Largest cluster size | Total no. of narratives clustered | % of narratives clustered | Silhouette coefficient | Dunn-Index (cosine) | Avg. point-to-centroid cosine distance |
|---|---|---|---|---|---|---|---|
| 0.95 | 958 | 238 | 2,382 | 0.25% | **0.7952** | 0.0824 | **0.0069** |
| 0.9 | 2,218 | 2,155 | 9,497 | 0.99% | 0.3587 | 0.0657 | 0.0199 |
| 0.85 | **4,363** | 11,856 | 47,109 | 4.93% | 0.0554 | 0.0677 | 0.0443 |
| 0.8 | 3,912 | 49,546 | 127,966 | 13.38% | -0.0785 | 0.0276 | 0.0796 |
| 0.75 | 1,950 | 128,593 | 197,426 | 20.64% | -0.1447 | 0.0508 | 0.1174 |
| 0.7 | 519 | 235,956 | 250,400 | 26.19% | -0.1194 | 0.0979 | 0.1440 |
| 0.65 | 108 | 310,609 | 312,027 | 32.64% | 0.0172 | 0.1967 | 0.1409 |
| 0.6 | 35 | 426,181 | **426,509** | **44.61%** | 0.1289 | **0.3826** | 0.1371 |

Table A.7: Statistics and cluster quality metrics computed for different cosine similarity thresholds used during clustering distress narratives (minimum cluster size is considered to be 2).

cosine distance suggest, 0.95 seems to be the best threshold that maximized the cluster quality. However, at this threshold only 0.25% of the entire dataset were assigned to clusters and the rest remained as outliers. The above metrics are also known to work best only for datasets having convex shaped clusters and the results of manually inspecting a subset of 100 distress narratives (10 narratives sampled uniformly at random from 10 random clusters) confirmed that these metrics do not work best for this dataset. 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. Topics discovered at different thresholds through manual inspection are are shown in Table A.9. This resulted in selecting an optimal threshold of 0.85, using which 4,363 distinguishable clusters were identified.

## A.11 Example dialogue responses generated by HEAL-ranked and other state-of-the-art empathetic response generation models

Table A.10 shows some examples of dialogue responses generated by the models used for comparison for distress prompts in the RED dialogue dataset.

## A.12 Additional details of the AMT experiment to compare responses generated by HEAL-ranked against state-of-the-art empathetic response generation models

In the human evaluation task conducted using Amazon Mechanical Turk (AMT) to compare the responses generated by HEAL-ranked, 10 dialogues were bundled into one HIT (Human Intelligence Task) such that one worker works on at least 10 such cases to avoid too much bias between answers. To evaluate the quality of the work generated, we included three quiz questions equally spaced in a HIT. In these, we included the ground-truth response among the other responses produced by the models. If a worker rated the ground-truth response either as *Good* or *Okay*, then a bonus point was added. To encourage attentiveness to the task, for those who obtained at least two out of three quiz questions correctly, we gave a bonus of 0.1$. The instructions given and the user interface of the task are similar to the human evaluation task conducted in Chapter 4. Workers who accepted more than 10 HITs were blocked automatically to avoid one worker monopolizing the task. We monitored the total dwell time on a task and the assignments that were completed in less than two minutes were automatically rejected to avoid work by inattentive workers and spammers. Table A.8 shows the statistics of the experiment.

| Description | Statistics |
|---|---|
| No. of HITs | 20 (10 dialogues each) |
| No. of assignments | 60 (20 × 3) |
| No. of workers | 38 |

| | |
|---|---|
| No. of blocked workers | 1 |
| Percentage of bonuses earned | 81.67% |
| Average time per HIT | 7.19 min. |
| Percentage of 2 out of 3 worker agreements on ratings | 85% |
| Inter rater agreement (Fleiss' kappa) | 0.23 (fair agreement) |

Table A.8: Statistics of the AMT human evaluation experiment.

## A.13 User interfaces of the MITI annotation experiment

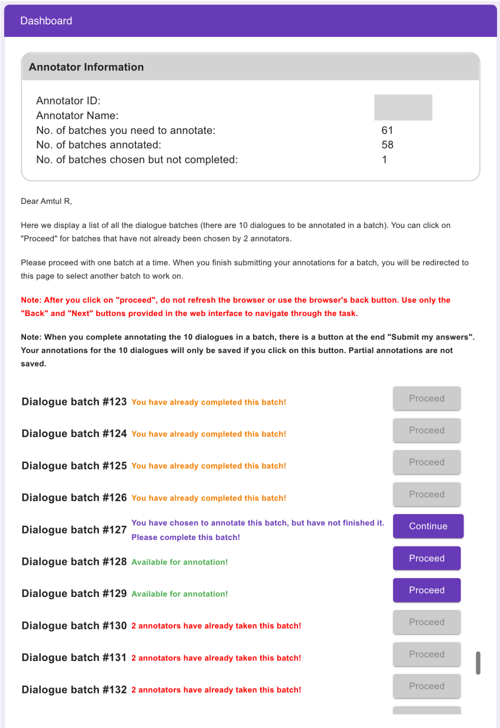
Figures A.5 and A.6 shows the user interfaces of the first and second stages of the MITI annotation experiment conducted in UpWork. The first stage is when two workers from UpWork were asked to annotate each sentence contained in the listener utterances of the dialogues from CounselChat and RED datasets and the second stage is when a high quality worker was asked to act as a judge to resolve the disagreements occured in the first stage. Interfaces similar to the second stage were used in the third stage as well. To educate the worker on the MITI coding scheme and the labels we derived out of it, a detailed tutorial was shown to the worker at the beginning of the task. This is shown in Figure A.5c. A practice task to self-evaluate their competence in annotating responses with the labels derived from the MITI code followed next. Figure A.6c depicts this.

## A.14 Confusion matrices between the annotators in the three stages of the MITI annotation experiment

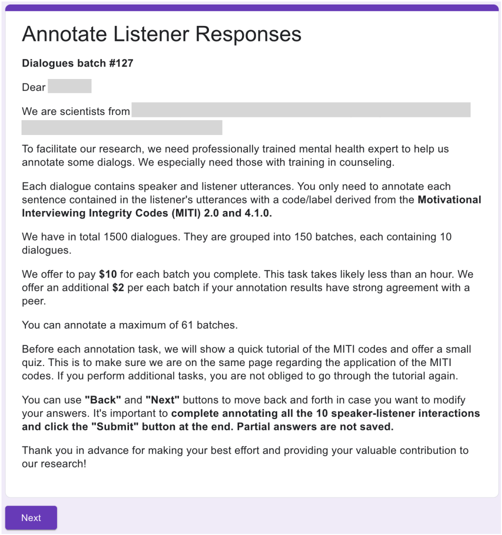
Figure A.7 shows the confusion matrices in stage 1 of the experiment between the two annotators for the CounselChat and RED datasets separately. Labels such as *Give Information*, *Advise without Permission*, and *Closed Question* had the highest agreement between the two workers in the CounselChat dataset, whereas in RED, the highest agreed labels were *Self-Disclose*, *Give Information*, and *Support*.

Figure A.8 shows the confusion matrices between the two annotators for sentences for which the label was unresolved in stage 1 and between each of these annotators and the judge in stage 2 of the annotation process. From the second and third confusion matrices corresponding to each dataset, it could be seen how the judge's annotations aligned with annotations from each annotator from stage 1.

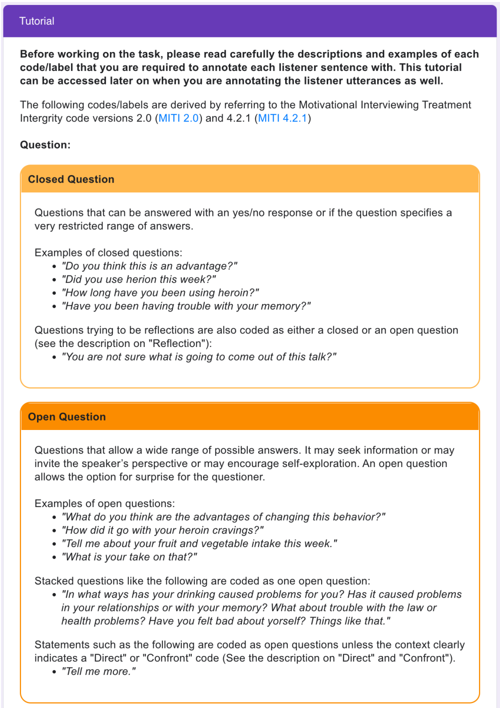
Figure A.9 shows the confusion matrices between the two remaining annotators after the annotations from the poorly performed worker are removed and between each of these annotators and the second judge in stage 2 of the annotation process. Note that in the remaining two annotations, the first one comes from a relatively better-performed worker from stage 1 and the second one comes from the first judge from stage 2. By observing



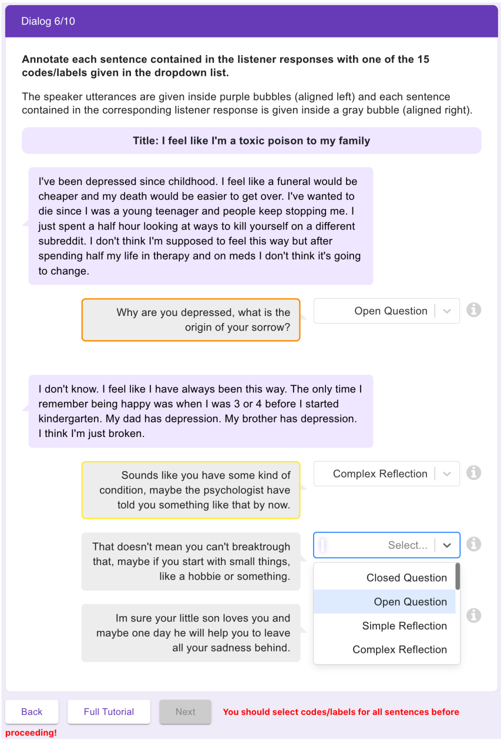

(a) The dashboard interface




(b) Instructions



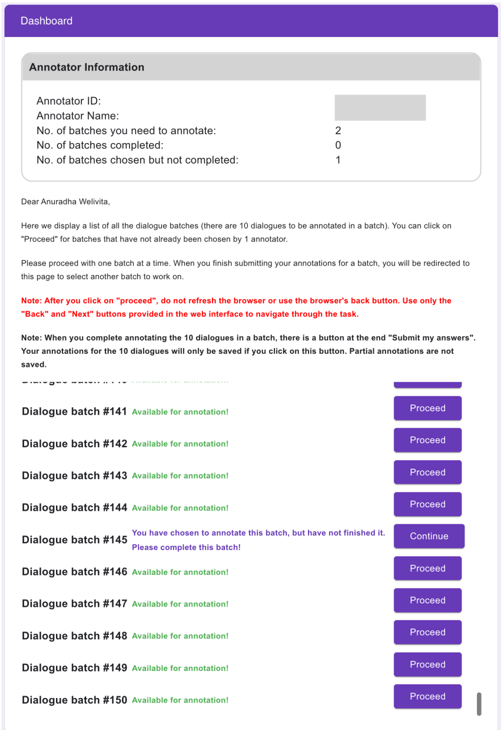

(c) The tutorial



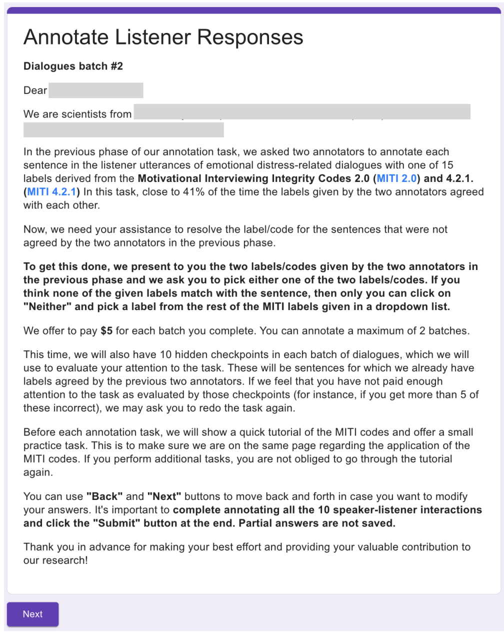

(d) The annotation task interface

Figure A.5: User interfaces of the first stage of the MITI annotation experiment.

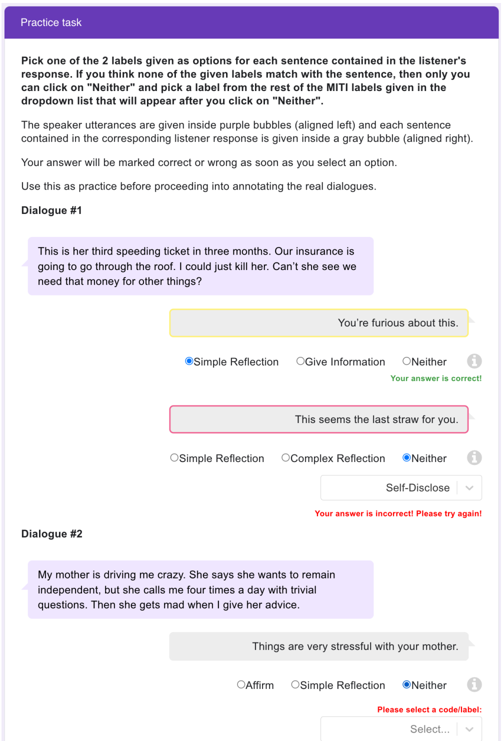
## A.14 Confusion matrices between the annotators in the three stages of the MITI annotation experiment



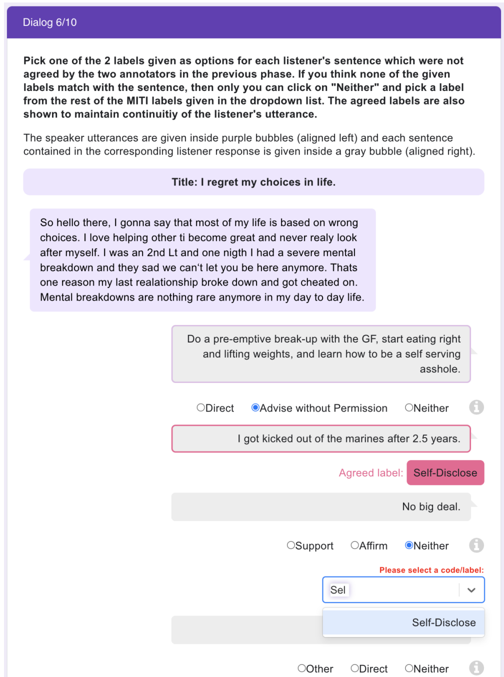

(a) The dashboard interface




(b) Instructions




(c) The practice task




(d) The task interface for resolving labels

Figure A.6: User interfaces of the second stage of the MITI annotation experiment.

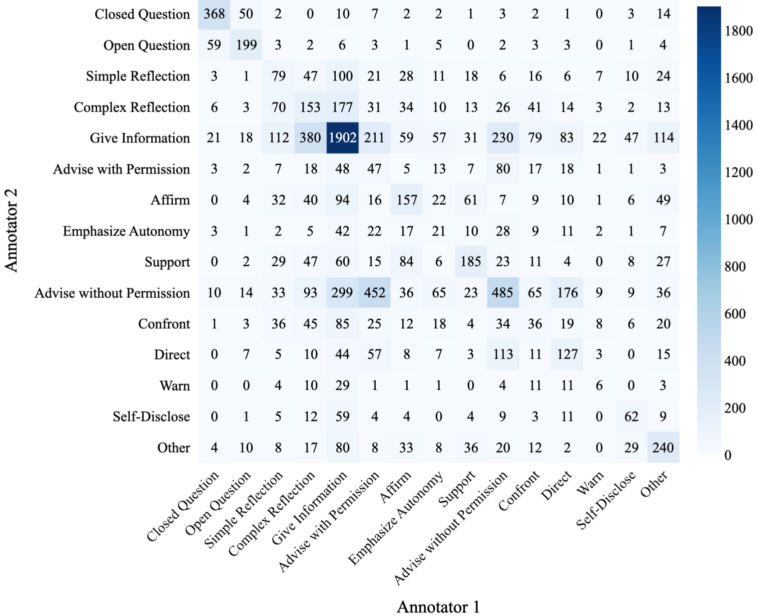
| **Topics identified at different thresholds when clustering distress narratives:** | |
|---|---|
| 0.95 | Inquiring about methods to commit suicide, Going to suicide soon, Thinking about suicide, Need someone to talk to, How to make suicide look like an accident, Cannot stop crying, Emojis, Want to fucking die, How to stop being suicidal, Is suicide selfish |
| 0.9 | Suicide, How to handle depression, Losing weight, Is suicide bad, I wish I could die, How to stop suicidal thoughts, Questioning about what depression really is, How to control my emotions, Cutting myself, Suicide by shooting with a gun |
| 0.85 | Suicide, Depression, Antidepressants, Missing one's dog, Therapy, Eating disorder, Feeling lonely, College stress, Dating/Relationships, Can't live anymore! |
| 0.8 | ??, Feeling lonely, Pets/Dogs, Stressed with Christmas/Holiday season, ??, Dating/Sexual relationships, Therapy, Need to talk to someone, Addicted to video games, Hanging oneself |
| 0.75 | Suicidal thoughts (But vague), Interpersonal conflicts, Conflicts with friends/Difficulty making friends/Loneliness, Suicidal thoughts due to interpersonal conflicts (But vague), Need to talk to someone, Loved ones dying, Giving up dog/puppy, ??, Worried to have kids (but vague), Got refused by girlfriend (But vague) |
| 0.7 | ??, Feeling less of oneself (But vague), Workplace conflicts, Thankful for support/Offering support, Mexico/Spanish, Loved ones dying, Relationship conflicts (But vague), Teachers, Want someone to listen (But vague), ?? |
| 0.65 | ??, ??, Winter depression, Estonian language, I hate living in America, LGBT pride month/protests, Canada, Hate being Filipino, Belgium, Brazil |
| 0.6 | Suicide, I hate myself, ??, Antidepressants (But vague), Canadian cities related cluster, Treatments for anxiety and depression, Russian language, ??, ??, ?? |

Table A.9: Topics identified through manually inspecting randomly sampled 10 distress narratives from 10 random clusters. "??" is shown when the cluster topic was too vague to identify by inspecting the narratives. Topics highlighted in red include clusters that were too vague to identify a topic, and topics that were too broad or did not make sense with respect to identifying stressors.

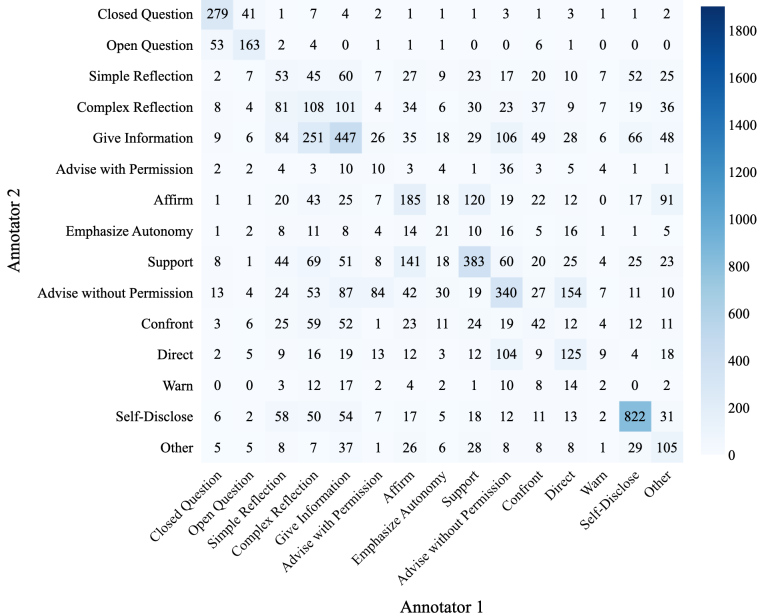
| Dialogue prompt | Responses | | |
|---|---|---|---|
| | Xie et al. (2020) | Blender | HEAL-ranked |
| It is a lame time to have my birthday. I'm not being so social as everyone else does :( | Why is that? | i ' m sorry to hear that . what did you do on your birthday ? did you have a good time ? | Don't like having birthday parties either. |
| How do you face anxiety? I feel it almost everyday. My stomach, throat, and my muscles become tight and sore. I can't concentrate. I always feel scared and confused. | Oh my gosh, that must be nerve wracking. Are you going to be okay? | i don ' t know how to deal with it . i think it ' s a combination of stress and anxiety . | Meditation is something that could help with that - if it resonates with you definitely check it out! |
| I'm just tired of feeling like this. I just want this to end. I don't want to get out from the bed and eat. I'm so tired and tired is an understatement. | I can understand that. I don't want to get out of bed anymore. | i ' m sorry to hear that . i hope you feel better soon . what are you going to eat ? | I'm really tired too.... |
| I cannot carry on like this. I have no reason to hold on anymore. I feel I'm a total failure. I don't have any friends and I feel so lonely. | I'm sorry to hear that | i am so sorry to hear that . i am sure you will find some new friends one day . | I'm so sorry this had to happen to you, if you need someone to talk to I'll be there for you. |
| I have been taking anti-depression medicine since two weeks. But I still cry constantly and feel hurt. | I don't think I will ever be able to get over it. | i ' m sorry to hear that . i hope you feel better soon . i know how you feel . | What antidepressants are you on? |

| | | | |
|---|---|---|---|
| I constantly think about suicide. But I never do it. Sometimes the thought becomes very aggressive and I cry. I know I won't do it so there's no use calling a hotline. | Have you thought about getting a new one? | i ' m so sorry to hear that . i hope you ' re able to get through it . | I'm sorry you are struggling with all this. |
| I like to be reborn in a better world. In a world where I matter to people :( | What's that? | That ' s a good goal to have . i ' m sure you ' ll have a great life . | I'm sorry that you feel this way, I can't imagine what you're going through. |
| What is life? What is a good life? What is a bad life? I think life without bad is worthless. But a life without good is worthless equally. | I agree with you | That ' s a great way of looking at it . it ' s hard to say what is good or bad . | All the experiences, whether good or bad, make life worth living. |
| What do you think will happen if I tell my therapist I'm suicidal? I'm 18 though my parents accompany me when going to the therapist. But they don't come to the room when I talk with him. | Have you talked to them about it? | i don ' t think you should tell your therapist about your thoughts about suicide . you should talk to your parents about it . | I know it sucks and I'm so sorry you're going through this! |
| I'm experiencing extreme highs and lows and having memory loss sometimes. Can you please help? | I don't know what you mean. | i ' m not sure what kind of help you ' re looking for , but i know it ' s important to be aware of your surroundings . | I find it's more the depression itself that causes memory loss and zoning out. |

Table A.10: Examples of responses generated by the three models for distress prompts from the RED dataset. To protect the anonymity of the actual authors, only paraphrased dialogue prompts are shown here.

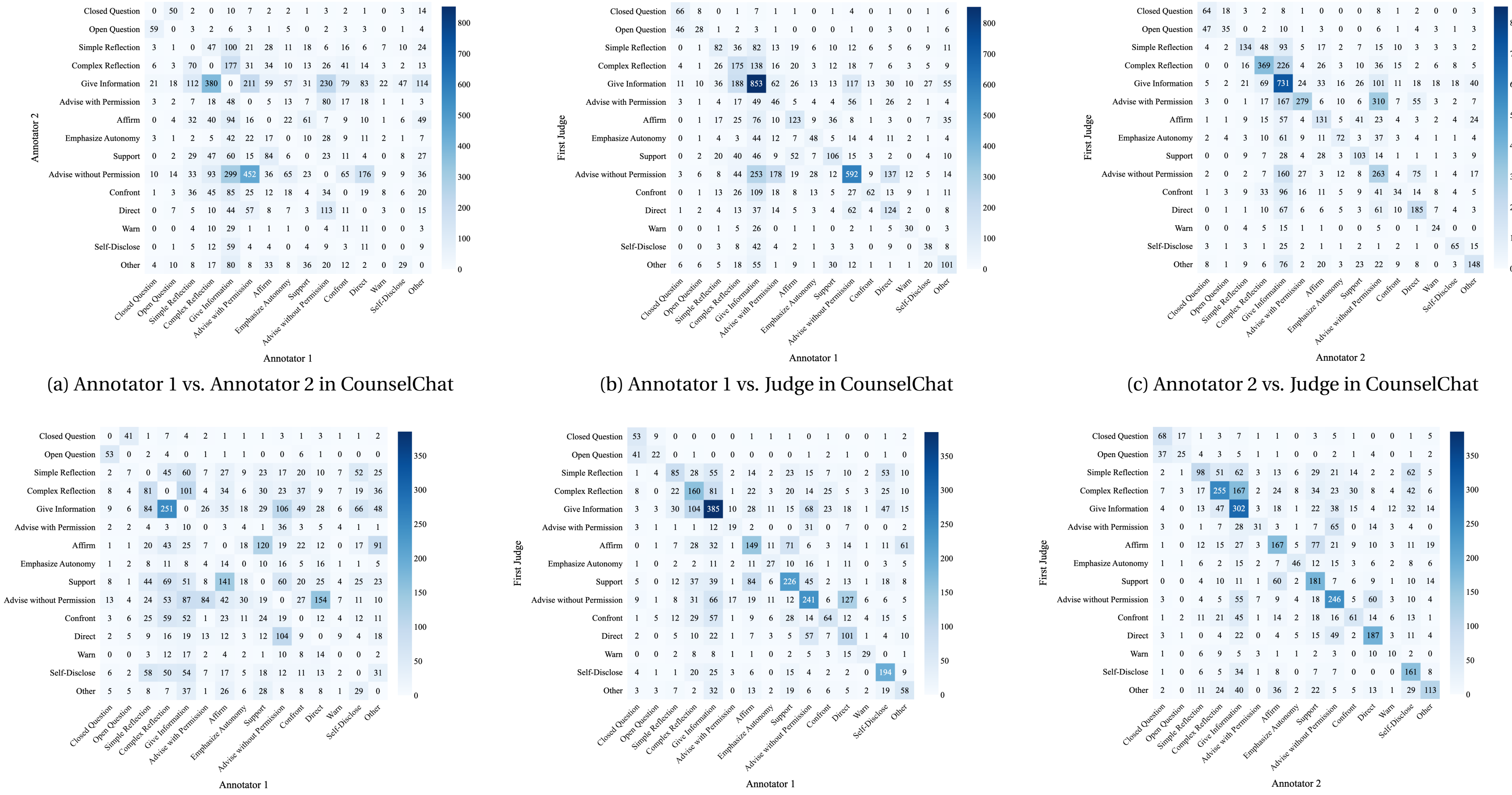



(a) CounselChat dataset




(b) RED dataset

Figure A.7: Confusion matrices between the two annotators for responses in the CounselChat and RED datasets during stage 1 of the annotation process.



(a) Annotator 1 vs. Annotator 2 in CounselChat (b) Annotator 1 vs. Judge in CounselChat (c) Annotator 2 vs. Judge in CounselChat

(d) Annotator 1 vs. Annotator 2 in RED (e) Annotator 1 vs. Judge in RED (f) Annotator 2 vs. Judge in RED

Figure A.8: Confusion matrices between the two annotators for sentences for which the label was unresolved in stage 1 and between each of these annotators and the judge in stage 2 of the annotation process. From the second and third confusion matrices corresponding to each dataset, it could be seen how the judge's annotations aligned with annotations from each annotator from stage 1.

the confusion matrices, it was noted that 73.34% times, the second judge agreed with the annotation provided by the first judge in stage 2. This further validated the quality of the judges selected.

## A.15 Examples of most frequent n-grams discovered corresponding to each MITI label

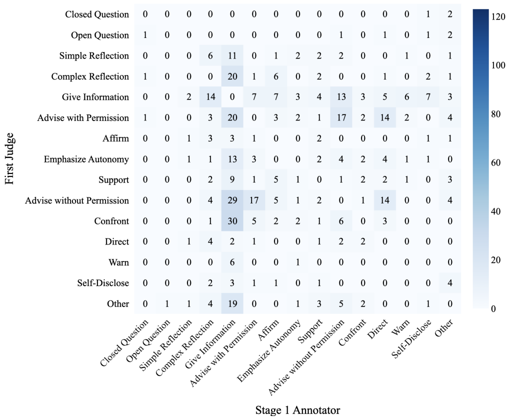
We denote examples of the most frequent N-grams corresponding to each label in Table A.12. For simplicity, we list only some of them along with their corresponding frequencies. For data augmentation, we used all four-grams and five-grams, which had a frequency of above 5.

## A.16 Statistics of the MITI labels extended through n-gram based matching and similarity-based retrieval methods

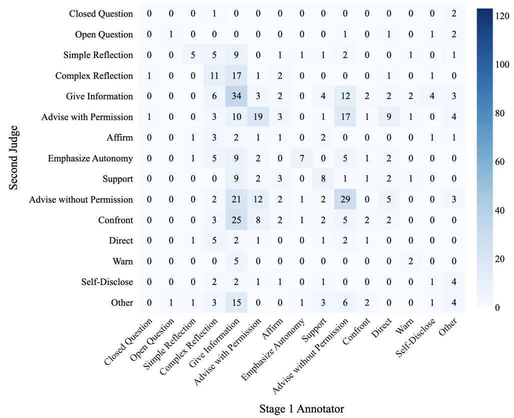
Table A.11 shows the statistics of the labels extended through N-gram based matching in CC and RED datasets. We also encountered 518 and 53,196 sentences in CounselChat and RED datasets respectively that had overlapping labels, which were discarded due to ambiguity.

| Label | N-gram based matching | | | Similarity-based retrieval | | |
|---|---|---|---|---|---|---|
| | # Labels in CC | # Labels in RED | Total | # Labels in CC | # Labels in RED | Total |
| Closed Question | 75 | 17,190 | 17,265 | 132 | 71,505 | 61,637 |
| Open Question | 29 | 12,242 | 12,271 | 49 | 36,107 | 36,156 |
| Simple Reflection | 71 | 9,674 | 9,745 | 43 | 21,827 | 21,870 |
| Complex Reflection | 110 | 20,539 | 20,649 | 20 | 17,243 | 17,263 |
| Give Information | 571 | 71,996 | 72,567 | 893 | 166,586 | 167,479 |
| Advise w/ Permission | 161 | 5,979 | 6,140 | 5 | 3,728 | 3,733 |
| Affirm | 136 | 16,407 | 16,543 | 187 | 106,066 | 106,253 |
| Emphasize Autonomy | 0 | 0 | 0 | 3 | 2,839 | 2,842 |
| Support | 213 | 94,670 | 94,883 | 482 | 528,469 | 528,951 |
| Advise w/o Permission | 520 | 58,857 | 59,377 | 969 | 171,502 | 172,471 |
| Confront | 0 | 0 | 0 | 1 | 2,581 | 2,582 |
| Direct | 0 | 0 | 0 | 16 | 21,058 | 21,074 |
| Warn | 0 | 0 | 0 | 6 | 2,342 | 2,348 |
| Self-Disclose | 5 | 28,309 | 28,314 | 8 | 14,702 | 14,710 |
| Other | 27 | 4,498 | 4,525 | 67 | 29,457 | 28,524 |
| Total | 1,918 | 340,361 | 342,279 | 2,881 | 1,196,012 | 1,198,893 |

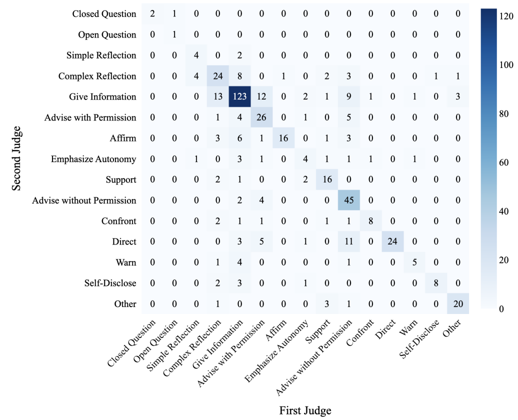
Table A.11: Statistics of the labels extended through N- gram-based matching and similarity-based retrieval in CC and RED datsets.



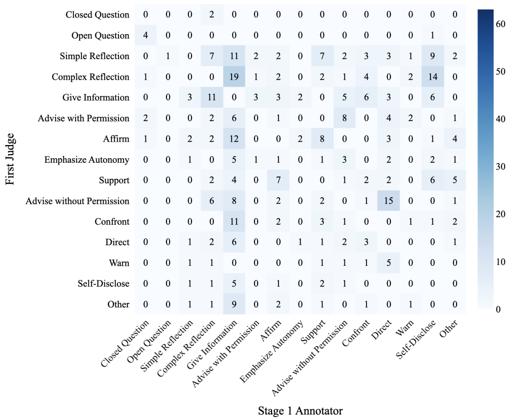

(a) Stage 1 annotator vs. First judge in CounselChat



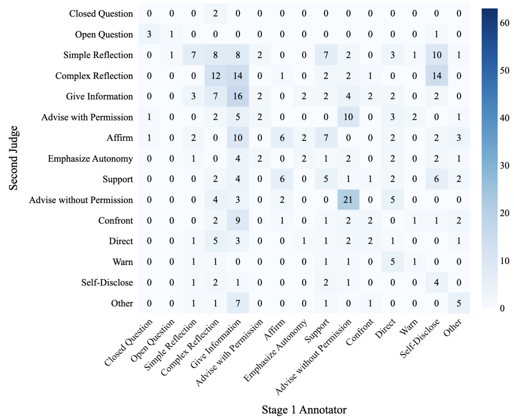

(b) Stage 1 annotator vs. Second judge in CounselChat



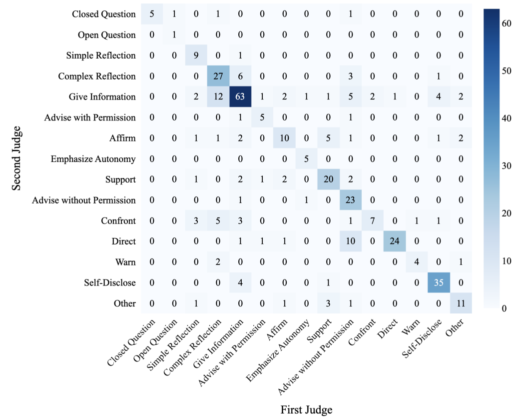

(c) First judge vs. Second judge in CounselChat




(d) Stage 1 annotator vs. First judge in RED




(e) Stage 1 annotator vs. Second judge in RED




(f) First judge vs. Second judge in RED

Figure A.9: Confusion matrices between different annotators for sentences which were still unresolved after stage 2 that contained at least one annotation from a poorly performed worker. It could be observed that the second judge's annotations in stage 3 aligned mostly with the first judge's annotations in stage 2.

| **Label** | **Examples of most frequent four-grams** | **Examples of most frequent five-grams** |
|---|---|---|
| Closed Question | *Do you have any* (11), *Do you have a* (7), *Do you want to* (7), *Have you talked to* (5), *Do you think you* (5) | - |
| Open Question | *Do you want to* (10), *you want to be* (8), *How do you feel* (5), *Why do you feel* (5), *What is the evidence* (5) | *Do you want to be* (6) |
| Simple Reflection | *It sounds like you* (16), *sounds like you have* (9), *sounds like you are* (8) | *It sounds like you are* (7), *It sounds like you have* (6) |
| Complex Reflection | *It sounds like you* (26), *My guess is that* (5), *The fact that you* (5), *why you might feel* (5) | *It sounds like you are* (7), *It sounds like you have* (6) |
| Give Information | *may be able to* (11), *who you are and* (8), *For example , if* (8), *A lot of people* (7), *A good therapist will* (6) | *who you are and what* (6), *you are and what you* (6), *be able to help you* (6), *it is important to* (5), *a higher level of care* (5) |
| Advise w/ Permission | *It may be helpful* (8), *would be a good* (7), *you would like to* (6), *a good idea to* (5), *I would encourage you* (5) | *It may be helpful to* (6), *I would encourage you to* (5) |
| Affirm | *I 'm glad you* (19), *wish you the best* (7), *I 'm glad that* (7), *I wish you the* (6), *you 're doing better* (5) | *I 'm glad you 're* (9), *I wish you the best* (6) |
| Emphasize Autonomy | - | - |
| Support | *I 'm so sorry* (12), *sorry to hear about* (12), *I hope you find* (10), *you are not alone* (9), *m here for you* (8) | *I 'm sorry to hear* (11), *I 'm here for you* (8), *I know how you feel* (8), *if you wan na talk* (6), *I hope you can find* (5) |
| Advise w/o Permission | *Reach out to a* (6), *I would suggest that* (6), *I think you should* (5), *I urge you to* (5), *I think you need* (5) | *, you may want to* (5), *I would suggest that you* (5) |
| Confront | - | - |
| Direct | - | - |
| Warn | - | - |

| | | |
|---|---|---|
| Self-Disclose | *I feel the same* (9), *I 've been in* (8), *the same way .* (7), *do n't know what* (6), *I feel like it* (5) | *I feel the same way* (5), *I do n't know what* (5) |
| Other | *you for your question* (12), *Hello , and thank* (9), *thank you for your* (9) | *Hello , and thank you* (9), *you for your question .* (12) |

Table A.12: Examples of most frequent four-grams and five-grams corresponding to each label. Their frequencies are denoted within brackets.

## A.17 Statistics corresponding to each label in the augmented MI datasets

Table A.14 shows the statistics corresponding to each label in the MI Augmented (Union) and MI Augmented (Intersection) datasets developed by taking the union and the intersection of the sentences automatically annotated by N-gram based matching and similarity based retrieval methods.

## A.18 Additional technical details related to the MI classifier

We used the same hyper-parameter setting used in RoBERTa Zhuang et al. (2021) when training the MI classifier. We used the Adam optimizer with $\beta_1$ of 0.9, $\beta_2$ of 0.98, an $\epsilon$ value of $1 \times 10^{-6}$, and a learning rate of $2 \times 10^{-5}$. A dropout of 0.1 was used on all layers and attention weights, and a GELU activation function Hendrycks and Gimpel (2016b). We limited the maximum number of input tokens to 100, and used a batch size of 32. All models were trained for 20 epochs. In all cases, the optimal epoch was selected based on the average cross entropy loss calculated between the ground-truth and predicted labels of the human-annotated (MI Gold) validation set. All the experiments were conducted on a machine with 2x12cores@2.5GHz, 256 GB RAM, 2x200 GB SSD, and 4xGPU (NVIDIA Titan X Pascal). Experiments were also done using GPT3 as the pre-trained language model, however, RoBERTa was seen to outperform GPT3 in this classification task. Table A.13 shows the performance scores of the MI classifier when trained on gold-labeled and augmented MI datasets.

| **Dataset** | **Size** | | **Optimal Epoch** | **Train Loss** | **Valid Acc. (%)** | **Test Acc. (%)** |
|---|---|---|---|---|---|---|
| MI Gold | Train: | 13,449 | 7 | 0.3002 | 67.08 | 68.31 |
| | Valid (Gold): | 1,681 | | | | |
| | Test (Gold): | 1,681 | | | | |
| MI Augmented (Intersection) | Train: | 80,690 | 2 | 0.2277 | 64.07 | 67.13 |
| | Valid (Gold): | 1,681 | | | | |
| | Test (Gold): | 1,681 | | | | |
| MI Augmented (Union) | Train: | 1,375,107 | 13 | 0.1324 | **72.67** | **73.44** |
| | Valid (Gold): | 1,681 | | | | |
| | Test (Gold): | 1,681 | | | | |

Table A.13: The performance scores of the MI classifier when trained on gold-labeled and augmented MI datasets. All scores are reported on the human-annotated validation and test sets. All scores are reported for a single run.

| Label | MI Augmented (Intersection) | | | | MI Augmented (Union) | | | |
|---|---|---|---|---|---|---|---|---|
| | **# Labels in CC** | **# Labels in RED** | **Total** | **Total + MI Gold** | **# Labels in CC** | **# Labels in RED** | **Total** | **Total + MI Gold** |
| Closed Question | 9 | 5,598 | 5,607 | **6,512** | 135 | 78,932 | 79,067 | **79,972** |
| Open Question | 1 | 2,353 | 2,354 | **2,830** | 60 | 40,805 | 40,865 | **41,341** |
| Simple Reflection | 1 | 185 | 186 | **742** | 41 | 19,961 | 20,002 | **20,558** |
| Complex Reflection | 2 | 201 | 203 | **1,497** | 44 | 21,247 | 21,291 | **22,585** |
| Give Information | 77 | 3,379 | 3,456 | **8,312** | 1083 | 203,110 | 204,193 | **209,049** |
| Advise w/ Permission | 0 | 28 | 28 | **512** | 5 | 3,052 | 3,057 | **3,541** |
| Affirm | 48 | 898 | 946 | **1,891** | 208 | 106,575 | 106,783 | **107,728** |
| Emphasize Autonomy | 0 | 0 | 0 | **253** | 3 | 2,700 | 2,703 | **2,956** |
| Support | 76 | 44,635 | 44,711 | **45,944** | 551 | 592,220 | 592,771 | **594,004** |
| Advise w/o Permission | 144 | 8,872 | 9,016 | **11,301** | 1,029 | 196,571 | 197,600 | **199,885** |
| Confront | 0 | 0 | 0 | **318** | 0 | 2,468 | 2,468 | **2,786** |
| Direct | 0 | 0 | 0 | **898** | 15 | 20,690 | 20,705 | **21,603** |
| Warn | 0 | 0 | 0 | **113** | 6 | 2,278 | 2,284 | **2,397** |
| Self-Disclose | 0 | 729 | 729 | **2,119** | 12 | 36,522 | 36,534 | **37,924** |
| Other | 0 | 5 | 5 | **810** | 67 | 31,268 | 31,335 | **32,140** |
| Total | 358 | 66,883 | 67,241 | **84,052** | 3,259 | 1,358,399 | 1,361,658 | **1,378,469** |

Table A.14: Statistics of the annotated responses in MI Augmented (Intersection) and MI Augmented (Union) datasets.

## A.19 Additional technical details of the MI rephrasing models

For developing rephrasing models, we used the 90M parameter version of Blender Roller et al. (2021). It contains an 8 layer encoder, an 8-layer decoder with 512-dimensional embeddings, and 16 attention heads. It has a maximum input length of 1024 tokens. All code for fine-tuning is available in ParlAI Miller et al. (2017b). All the models were fine-tuned for 200 epochs, with a batch size of 8, and a learning rate of $1 \times 10^{-6}$. For other hyperparameters, we used the default values defined in their documentation at https://parl.ai/projects/recipes. Fine-tuning the models was conducted in a machine with 2x12cores@2.5GHz, 256 GB RAM, 2x200 GB SSD, and 4xGPU (NVIDIA Titan X Pascal).

We also used GPT3 pretrained language model having 175 billion parameters. The smallest but fastest version of GPT3, Ada was used in our experiments. Fine-tuning of GPT3 models were done through the paid API provided by OpenAI (www.openai.com) following API guide at https://beta.openai.com/docs/guides/fine-tuning. We used the default set of hyperparameters for fine-tuning all GPT3 based models. These hyperparameters are tested to work well across a range of use cases. All the models were fine-tuned for 4 epochs, with a batch size ≈0.2% of the number of examples in the training set (capped at 256), and a learning rate of 0.05.

## A.20 Examples of rephrased sentences by the MI rephrasing models

Table A.15 shows some examples of rephrased sentences by the different rephraser models we fine-tuned.

**Advise without permission:** *try to learn from your mistakes and meet some new people .*
**Advise with permission (ground-truth):** *It may be important to try to learn from your mistakes and meet some new people .*

**Template-based:**

**Blender:** *it would be a good idea to try to learned from your mistakes and meet some new people like me .*
**GPT3:** *I encourage you to try to learn from your mistakes and meet some new people .*

**Retrieval-based:**

**Blender:** *In addition to being open to meeting new friends , involve yourself in projects from which you ' ll feel satisfied by your own results .*
**GPT3:** *I think it would be valuable for you to start mingling with other young people .*

**Template+Retrieval-based:**

**Blender:** *i would encourage you to look up ' ' toxic relationships ' ' and try to make your relationships stronger .*
**GPT3:** *I would encourage you to try to learn from your mistakes and meet some new people.*

**Template+Retrieval-based (with generic prompting):**

**Blender:** *in addition to being open to meeting new friends , involve yourself in projects from which you ' ll feel satisfied by your own results .*
**GPT3:** *I would encourage you to try to learn from your mistakes and meet some new people.*

**Template+Retrieval-based (with N-gram prompting):**

**Prompt:** *it may be important to*

**Blender:** *it may be important to try to learn from your mistakes and meet a new person . : ) it may also be important for you to start engaging in things outside of the home that make you happy .*
**GPT3:** *It may be important to try to learn from your mistakes and meet some new people*

---

**Advise without permission:** *find something you enjoy and work towards it .*
**Advise with permission (ground-truth):** *It maybe helpful to find something you enjoy and work towards it .*

**Template-based:**

**Blender:** *it would be a good idea to find something yourself and work towards it . . . i wonder if you have heard of this ?*
**GPT3:** *It may be important to find something you enjoy and work towards it .*

**Retrieval-based:**

**Blender:** *i think it would be valuable for you to start engaging in things outside of the home that make you happy .*
**GPT3:** *Give yourself some credit for being awesome and tell yourself you can build new friendships .*

**Template+Retrieval-based:**

**Blender:** *i think it would be valuable for you to start engaging in things outside of the home that make you happy .*
**GPT3:** *It would be a good idea to find something you enjoy and work towards it .*

**Template+Retrieval-based (with generic prompting):**

**Blender:** *i think it would be valuable for you to start engaging in things outside of the home that make you happy .*
**GPT3:** *An option would be to find something you enjoy and work towards it .*

**Template+Retrieval-based (with N-gram prompting):**

**Prompt:** *it maybe helpful to*

**Blender:** *it's perhaps helpful to find something yourself and work towards it .*
**GPT3:** *It maybe helpful to find something you enjoy and work towards it .*

---

**Advise without permission:** *Talk to a therapist .*
**Advise with permission (ground-truth):** *I encourage you to meet with a therapist and ask them how they might be able to help you .*

**Template-based:**

**Blender:** *it would be a good idea to talk to one of the therapists . i wonder if you can talk to an therapist .*

**GPT3:** *I encourage you to talk to a therapist .*

**Retrieval-based:**

**Blender:** *I encourage you to meet with a therapist and ask them how they might be able to help you .*

**GPT3:** *I encourage you to meet with a therapist and ask them how they might be able to help you*

**Template+Retrieval-based:**

**Blender:** *i encourage you to meet with a therapist and ask them how they might be able to help you .*

**GPT3:** *I encourage you to meet with a therapist and ask them how they might be able to help*

**Template+Retrieval-based (with generic prompting):**

**Blender:** *i encourage you to meet with a therapist and ask them how they might be able to help you .*

**GPT3:** *I encourage you to meet with a therapist and ask them how they might be able to help*

**Template+Retrieval-based (with N-gram prompting):**

**Prompt:** *able to help you*

**Blender:** *i encourage you to meet with a therapist . they ' ll help you find some strategies or tools to make it easier .*

**GPT3:** *I encourage you to meet with a therapist and ask them how they might be able to help*

Table A.15: Examples of rephrased sentences by the different rephraser models we developed

## A.21 Additional technical details related to the automatic evaluation of the MI rephrasing models

We used the the NLTK package to compute the BLEU [1], METEOR [2], and chrf [3] scores. The ROUGE score and the BERTscore were computed using the rouge [4] and bert_score [5] python libraries, respectively. The POS distance was calculated as mentioned in the work by Tian et al. pos following the code released by the authors on github. [6] For computing the Word Mover Distance (WMD), we used Gensim's implementation of the WMD. [7] We used sentence embeddings generated using Sentence-BERT Reimers and Gurevych (2019) to compute the cosine similarity between the original and rephrased text. Among the models the authors have proposed, we used the *roberta-base-nli-stsb-mean-tokens* model, fine-tuned on the NLI

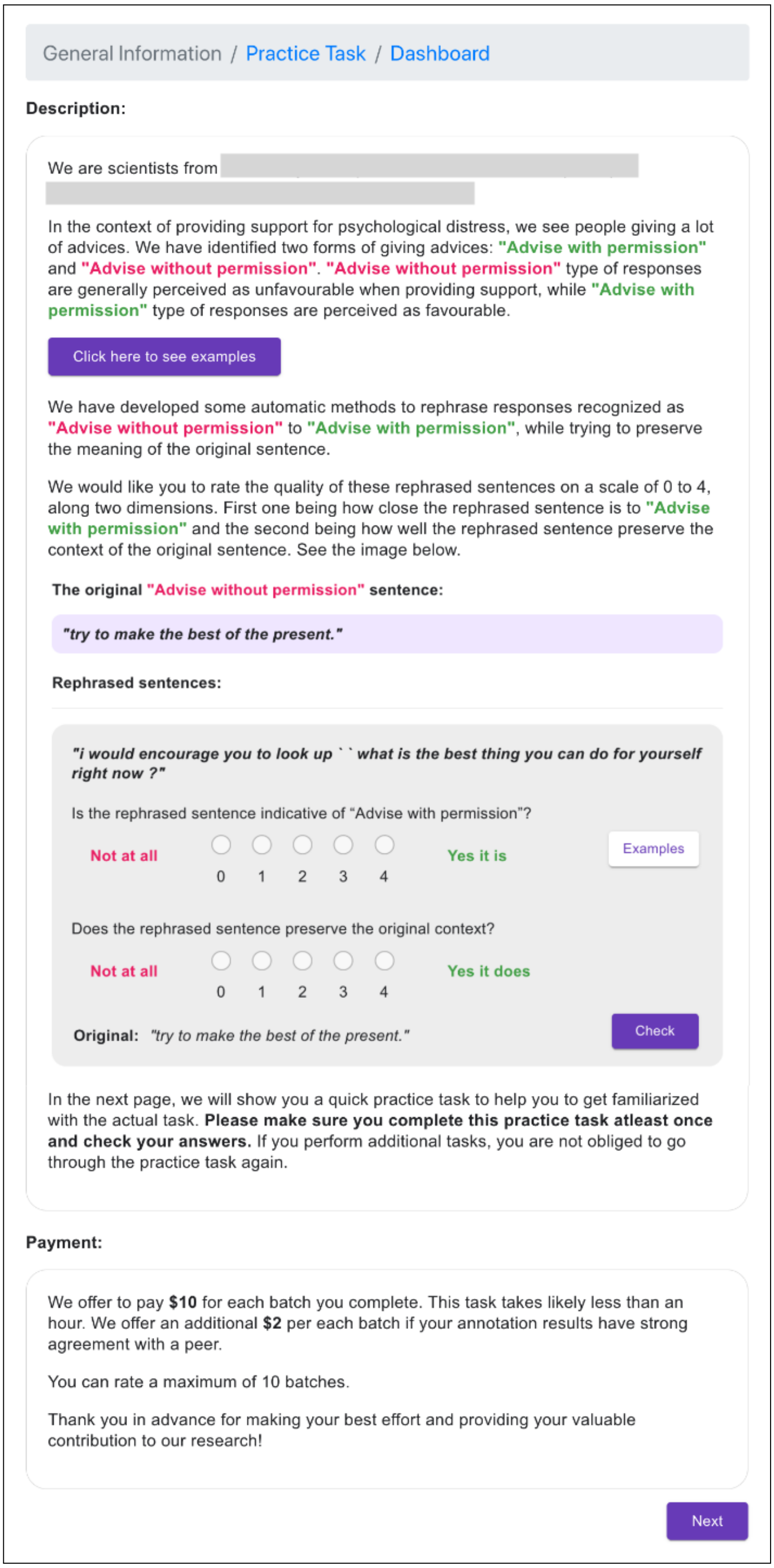
[1] https://www.nltk.org/_modules/nltk/translate/bleu_score.html
[2] https://www.nltk.org/_modules/nltk/translate/meteor_score.html
[3] https://www.nltk.org/_modules/nltk/translate/chrf_score.html
[4] https://pypi.org/project/rouge/
[5] https://pypi.org/project/bert-score/
[6] https://github.com/YouzhiTian/Structured-Content-Preservation-for-Unsupervised-Text-Style-Transfer/blob/master/POS_distance.py
[7] https://radimrehurek.com/gensim/auto_examples/tutorials/run_wmd.html

Bowman et al. (2015) and STS benchmark (STSb) Cer et al. (2017) datasets to generate the embeddings. All the automatic evaluation scores are reported for a single run.

## A.22 User interfaces of the human evaluation task carried out to evaluate the MI rephrasing models

Figures A.10, A.11, and A.12 shows the user interfaces developed for the human evaluation task. The first one shows the task description, the second one shows the self-evaluating practice task designed to get the counselors familiarized with the rating task, and the last one shows the actual human evaluation task itself.

## A.22 User interfaces of the human evaluation task carried out to evaluate the MI rephrasing models



General Information / Practice Task / Dashboard

**Description:**

We are scientists from

In the context of providing support for psychological distress, we see people giving a lot of advices. We have identified two forms of giving advices: **"Advise with permission"** and **"Advise without permission"**. **"Advise without permission"** type of responses are generally perceived as unfavourable when providing support, while **"Advise with permission"** type of responses are perceived as favourable.

Click here to see examples

We have developed some automatic methods to rephrase responses recognized as **"Advise without permission"** to **"Advise with permission"**, while trying to preserve the meaning of the original sentence.

We would like you to rate the quality of these rephrased sentences on a scale of 0 to 4, along two dimensions. First one being how close the rephrased sentence is to **"Advise with permission"** and the second being how well the rephrased sentence preserve the context of the original sentence. See the image below.

**The original "Advise without permission" sentence:**

***"try to make the best of the present."***

**Rephrased sentences:**

***"i would encourage you to look up ` ` what is the best thing you can do for yourself right now ?"***

Is the rephrased sentence indicative of “Advise with permission”?

**Not at all** 0 1 2 3 4 **Yes it is** Examples

Does the rephrased sentence preserve the original context?

**Not at all** 0 1 2 3 4 **Yes it does**

**Original:** *"try to make the best of the present."* Check

In the next page, we will show you a quick practice task to help you to get familiarized with the actual task. **Please make sure you complete this practice task atleast once and check your answers.** If you perform additional tasks, you are not obliged to go through the practice task again.

**Payment:**

We offer to pay **$10** for each batch you complete. This task takes likely less than an hour. We offer an additional **$2** per each batch if your annotation results have strong agreement with a peer.

You can rate a maximum of 10 batches.

Thank you in advance for making your best effort and providing your valuable contribution to our research!

Next

Figure A.10: Human evaluation task description.

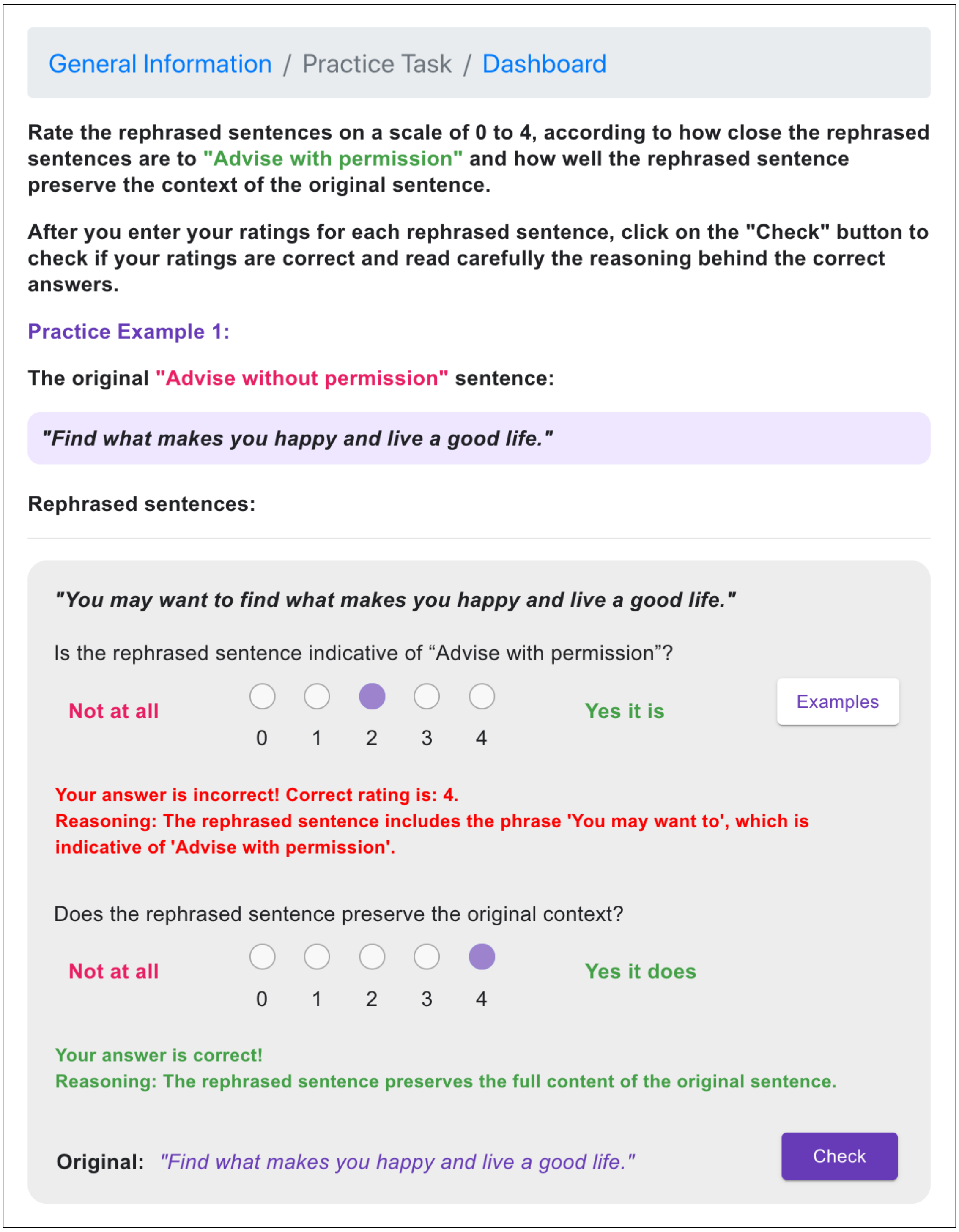



Figure A.11: Self-evaluating practice task offered to the counselors to get familiarized with the rating task.

### A.22 User interfaces of the human evaluation task carried out to evaluate the MI rephrasing models

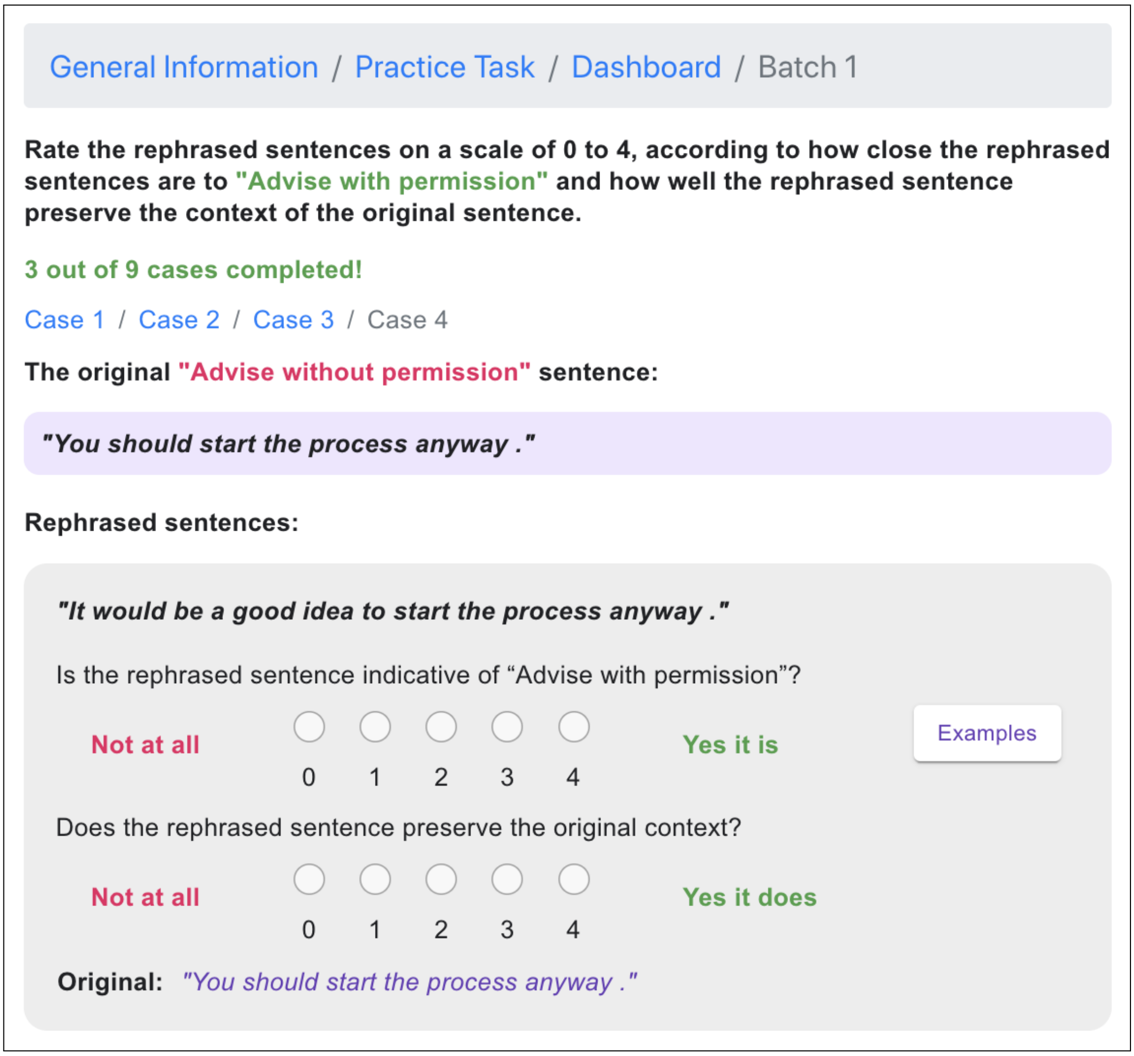



Figure A.12: The human evaluation task interface.

# Bibliography


Abercrombie, G., Curry, A. C., Dinkar, T., and Talat, Z. (2023). Mirages: On anthropomorphism in dialogue systems. *arXiv preprint arXiv:2305.09800*.

Adiwardana, D., Luong, M.-T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., et al. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Alambo, A., Gaur, M., Lokala, U., Kursuncu, U., Thirunarayan, K., Gyrard, A., Sheth, A., Welton, R. S., and Pathak, J. (2019). Question answering for suicide risk assessment using reddit. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 468–473. IEEE.

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.

Asghar, N., Poupart, P., Hoey, J., Jiang, X., and Mou, L. (2018). Affective neural response generation. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings 40*, pages 154–166. Springer.

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*, pages 722–735.

Baer, J. S., Rosengren, D. B., Dunn, C. W., Wells, E. A., Ogle, R. L., and Hartzler, B. (2004). An evaluation of workshop training in motivational interviewing for addiction and mental health clinicians. *Drug and alcohol dependence*, 73(1):99–106.

Baker, A., Lewin, T., Reichler, H., Clancy, R., Carr, V., Garrett, R., Sly, K., Devir, H., and Terry, M. (2002). Motivational interviewing among psychiatric in-patients with substance use disorders. *Acta Psychiatrica Scandinavica*, 106(3):233–240.

Ball, P. (2011). How movies mirror our mimicry.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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.

Bellet, P. S. and Maloney, M. J. (1991). The importance of empathy as an interviewing skill in medicine. *Jama*, 266(13):1831–1832.

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*, pages 94–102.

Bickmore, T. W. (2003). *Relational agents: Effecting change through human-computer relationships.* PhD thesis, Massachusetts Institute of Technology.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

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*, pages 1247–1250.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., et al. (2020a). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020b). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Buechel, S., Buffone, A., Slaff, B., Ungar, L., and Sedoc, J. (2018). Modeling empathy and distress in reaction to news stories. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4758–4765, Brussels, Belgium. Association for Computational Linguistics.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Can, D., Georgiou, P. G., Atkins, D. C., and Narayanan, S. S. (2012). A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *Thirteenth Annual Conference of the International Speech Communication Association*.

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

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

Chatterjee, A., Narahari, K. N., Joshi, M., and Agrawal, P. (2019b). Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 39–48.

Chen, Z., Song, R., Xie, X., Nie, J.-Y., Wang, X., Zhang, F., and Chen, E. (2019a). Neural response generation with relevant emotions for short text conversation. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, pages 117–129. Springer.

Chen, Z., Song, R., Xie, X., Nie, J.-Y., Wang, X., Zhang, F., and Chen, E. (2019b). Neural response generation with relevant emotions for short text conversation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 117–129. Springer.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. (2022). Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. *Artificial intelligence*, 2(1):1–25.

Colombo, P., Witon, W., Modi, A., Kennedy, J., and Kapadia, M. (2019). Affect-driven dialog generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3734–3743, Minneapolis, Minnesota. Association for Computational Linguistics.

Core, M. G. and Allen, J. (1997). Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.

De Choudhury, M. and De, S. (2014). Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*.

Decety, J. (2010). The neurodevelopment of empathy in humans. *Developmental neuroscience*, 32(4):257–267.

Design, E. A. (2016). A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems, version 1. *IEEE Standards Assoc.*

DeVault, D., Artstein, R., Benn, G., Dey, T., Fast, E., Gainer, A., Georgila, K., Gratch, J., Hartholt, A., Lhommet, M., et al. (2014). Simsensei kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

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

Dhillon, I. S. and Sra, S. (2005). Generalized nonnegative matrix approximations with bregman divergences. In *NIPS*, volume 18. Citeseer.

Eisenberg, N. and Eggum, N. D. (2009). Empathic responding: Sympathy and personal distress. *The social neuroscience of empathy*, 6(2009):71–830.

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*, pages 697–706.

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark. Association for Computational Linguistics.

Fiesler, C. and Proferes, N. (2018). “participant” perceptions of twitter research ethics. *Social Media+ Society*, 4(1).

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

Fitzpatrick, K. K., Darcy, A., and Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e19.

Foltz, P. W., Kintsch, W., and Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Forgues, G., Pineau, J., Larchevêque, J.-M., and Tremblay, R. (2014). Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2, page 168.

Fu, Z., Tan, X., Peng, N., Zhao, D., and Yan, R. (2018). Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Gan, C., Gan, Z., He, X., Gao, J., and Deng, L. (2017). Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Garcez, A. d. and Lamb, L. C. (2023). Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, pages 1–20.

Gaume, J., Gmel, G., Faouzi, M., and Daeppen, J.-B. (2009). Counselor skill influences outcomes of brief motivational interventions. *Journal of substance abuse treatment*, 37(2):151–159.

Ghandeharioun, A., McDuff, D., Czerwinski, M., and Rowan, K. (2019). Emma: An emotion-aware wellbeing chatbot. In *International Conference on Affective Computing and Intelligent Interaction*.

Ghosal, D., Majumder, N., Poria, S., Chhaya, N., and Gelbukh, A. (2019). DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. 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 154–164, Hong Kong, China. Association for Computational Linguistics.

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., and Scherer, S. (2017). Affect-LM: A neural language model for customizable affective text generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 634–642, Vancouver, Canada. Association for Computational Linguistics.

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

Gilbert, P. (2009). Introducing compassion-focused therapy. *Advances in psychiatric treatment*, 15(3):199–208.

Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society.

Goetz, J., Kiesler, S., and Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. In *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.*, pages 55–60.

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

Gopnik, A. and Wellman, H. M. (1992). Why the child's theory of mind really is a theory.

Gordon, R. M. (1992). The simulation theory: Objections and misconceptions. *Mind & Language*.

Gu, X., Yoo, K. M., and Lee, S.-W. (2021). Response generation with context-aware prompt learning. *arXiv preprint arXiv:2111.02643*.

Gupta, P., Bigham, J., Tsvetkov, Y., and Pavel, A. (2021). Controlling dialogue generation with semantic exemplars. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3018–3029, Online. Association for Computational Linguistics.

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*, pages 129–133.

Hatfield, E., Cacioppo, J. T., and Rapson, R. L. (1993). Emotional contagion. *Current directions in psychological science*, 2(3):96–100.

Hayes, S. C. and Pierson, H. (2005). Acceptance and commitment therapy. *Encyclopedia of cognitive behavior therapy*, pages 1–4.

Hazarika, D., Poria, S., Mihalcea, R., Cambria, E., and Zimmermann, R. (2018a). Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604.

Hazarika, D., Poria, S., Zadeh, A., Cambria, E., Morency, L.-P., and Zimmermann, R. (2018b). Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, page 2122. NIH Public Access.

Hedayatnia, B., Gopalakrishnan, K., Kim, S., Liu, Y., Eric, M., and Hakkani-Tur, D. (2020). Policy-driven neural response generation for knowledge-grounded dialog systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 412–421, Dublin, Ireland. Association for Computational Linguistics.

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., and Pineau, J. (2018). Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 123–129.

Hendrycks, D. and Gimpel, K. (2016a). Adjusting for dropout variance in batch normalization and weight initialization. *arXiv preprint arXiv:1607.02488*.

Hendrycks, D. and Gimpel, K. (2016b). Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Herzig, J., Feigenblat, G., Shmueli-Scheuer, M., Konopnicki, D., Rafaeli, A., Altman, D., and Spivak, D. (2016). Classifying emotions in customer support dialogues in social media. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 64–73.

Hettema, J., Steele, J., and Miller, W. R. (2005). Motivational interviewing. *Annu. Rev. Clin. Psychol.*, 1:91–111.

Hill, C. E., Carter, J. A., and O'Farrell, M. K. (1983). A case study of the process and outcome of time-limited counseling. *Journal of Counseling Psychology*, 30(1):3.

Hodges, S. D., Kiel, K. J., Kramer, A. D., Veach, D., and Villanueva, B. R. (2010). Giving birth to empathy: The effects of similar experience on empathic accuracy, empathic concern, and perceived empathy. *Personality and Social Psychology Bulletin*, 36(3):398–409.

Hosseini, S., Rezaei, A., Kazemi, S., and Samani, S. (2020). The effectiveness of motivational interviewing on academic procrastination in adolescents. *Psychological Methods and Models*, 11(39):81–94.

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)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., Luo, J., and Akkiraju, R. (2018). Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–12.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., and Xing, E. P. (2017). Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Huang, Y., Zhu, W., Xiong, D., Zhang, Y., Hu, C., and Xu, F. (2020). Cycle-consistent adversarial autoencoders for unsupervised text style transfer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2213–2223, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Inkster, B., Sarda, S., and Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.

Jeremy Sutton, P. (2022). Empathy in counseling: How to show empathetic understanding.

Jin, D., Jin, Z., Hu, Z., Vechtomova, O., and Mihalcea, R. (2022). Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Jin, Z., Jin, D., Mueller, J., Matthews, N., and Santus, E. (2019). IMaT: Unsupervised text attribute transfer via iterative matching and translation. 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 3097–3109, Hong Kong, China. Association for Computational Linguistics.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Kayhani, A. K., Meziane, F., and Chiky, R. (2020). Movies emotional analysis using textual contents. In *International Conference on Applications of Natural Language to Information Systems*, pages 205–212. Springer.

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

Khosla, S. (2018). Emotionx-ar: Cnn-dcnn autoencoder based emotion classifier. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 37–44.

Klonek, F. E., Quera, V., and Kauffeld, S. (2015). Coding interactions in motivational interviewing with computer-software: What are the advantages for process researchers? *Computers in Human Behavior*, 44:284–292.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*.

Kretzschmar, K., Tyroll, H., Pavarini, G., Manzini, A., Singh, I., and Group, N. Y. P. A. (2019). Can your phone be your therapist? young people's ethical perspectives on the use of fully automated conversational agents (chatbots) in mental health support. *Biomedical informatics insights*, 11:1178222619829083.

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

Kusner, M., Sun, Y., Kolkin, N., and Weinberger, K. (2015). From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Lahnala, A., Zhao, Y., Welch, C., Kummerfeld, J. K., An, L. C., Resnicow, K., Mihalcea, R., and Pérez-Rosas, V. (2021). Exploring self-identified counseling expertise in online support forums. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4467–4480, Online. Association for Computational Lingfgfggftzr666757tl.uistics.

Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211.

Lanteigne, C. (2019). Social robots and empathy: The harmful effects of always getting what we want.

Lee, D. (2016). Tay: Microsoft issues apology over racist chatbot fiasco.

Lee, H.-y., Ho, C.-H., Lin, C.-F., Chang, C.-C., Lee, C.-W., Wang, Y.-S., Hsu, T.-Y., and Chen, K.-Y. (2020). Investigation of sentiment controllable chatbot. *arXiv preprint arXiv:2007.07196*.

Levitt, H. M., Minami, T., Greenspan, S. B., Puckett, J. A., Henretty, J. R., Reich, C. M., and Berman, J. S. (2016). How therapist self-disclosure relates to alliance and outcomes: A naturalistic study. *Counselling Psychology Quarterly*, 29(1):7–28.

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*, pages 7871–7880.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2015). A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016a). 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*, pages 110–119.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016b). A diversity-promoting objective function for neural conversation models. In Knight, K., Nenkova, A., and Rambow, O., editors, *Proceedings of NAACL-HLT 2016*, pages 110–119.

Li, J., Jia, R., He, H., and Liang, P. (2018). Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017a). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017b). DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Lin, Z., Xu, P., Winata, G. I., Siddique, F. B., Liu, Z., Shin, J., and Fung, P. (2020). Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13622–13623.

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

Lison, P., Tiedemann, J., and Kouylekov, M. (2018). Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

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*. European Language Resources Association (ELRA).

Liu, B. and Sundar, S. S. (2018). Should machines express sympathy and empathy? experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, 21(10):625–636.

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*, pages 2122–2132.

Liu, R., Gao, C., Jia, C., Xu, G., and Vosoughi, S. (2022). Non-parallel text style transfer with self-parallel supervision. *arXiv preprint arXiv:2204.08123*.

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)*, pages 1489–1498.

Liu, Z., Niu, Z.-Y., Wu, H., and Wang, H. (2019). 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)*, pages 1782–1792.

Liu-Thompkins, Y., Okazaki, S., and Li, H. (2022). Artificial empathy in marketing interactions: Bridging the human-ai gap in affective and social customer experience. *Journal of the Academy of Marketing Science*, 50(6):1198–1218.

Madaan, A., Setlur, A., Parekh, T., Poczos, B., Neubig, G., Yang, Y., Salakhutdinov, R., Black, A. W., and Prabhumoye, S. (2020). Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Madotto, A., Lin, Z., Winata, G. I., and Fung, P. (2021). Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Mairesse, F. and Walker, M. A. (2011). Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics*, 37(3):455–488.

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

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

Merdivan, E., Singh, D., Hanke, S., Kropf, J., Holzinger, A., and Geist, M. (2020). Human annotated dialogues dataset for natural conversational agents. *Applied Sciences*, 10(3):762.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.

Miller, A., Feng, W., Batra, D., Bordes, A., Fisch, A., Lu, J., Parikh, D., and Weston, J. (2017a). ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Miller, A., Feng, W., Batra, D., Bordes, A., Fisch, A., Lu, J., Parikh, D., and Weston, J. (2017b). ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Mir, R., Felbo, B., Obradovich, N., and Rahwan, I. (2019). Evaluating style transfer for text. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 495–504, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Mitchell, J. and Lapata, M. (2008). Vector-based models of semantic composition. In *proceedings of ACL-08: HLT*, pages 236–244.

Montemayor, C., Halpern, J., and Fairweather, A. (2021). In principle obstacles for empathic ai: why we can't replace human empathy in healthcare. *AI & society*, pages 1–7.

Montenegro, C., López Zorrilla, A., Mikel Olaso, J., Santana, R., Justo, R., Lozano, J. A., and Torres, M. I. (2019). A dialogue-act taxonomy for a virtual coach designed to improve the life of elderly. *Multimodal Technologies and Interaction*, 3(3):52.

Mousavi, S. M., Cervone, A., Danieli, M., and Riccardi, G. (2021). Would you like to tell me more? generating a corpus of psychotherapy dialogues. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 1–9.

Moyers, T., Manuel, J., Ernst, D., Moyers, T., Manuel, J., Ernst, D., and Fortini, C. (2014). Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1). *Unpublished manual*.

Moyers, T. B., Martin, T., Manuel, J. K., Miller, W. R., and Ernst, D. (2003). The motivational interviewing treatment integrity (miti) code: Version 2.0. *Retrieved from Verfübar unter: www. casaa. unm. edu [01.03. 2005]*.

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.

Nambisan, P. (2011). Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association*, 18(3):298–304.

OpenAI (2023). Gpt-4 technical report.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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*, pages 1–12.

Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., and An, L. (2016). Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51.

Pérez-Rosas, V., Sun, X., Li, C., Wang, Y., Resnicow, K., and Mihalcea, R. (2018). Analyzing the quality of counseling conversations: the tell-tale signs of high-quality counseling. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Pérez-Rosas, V., Wu, X., Resnicow, K., and Mihalcea, R. (2019). What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Poskiparta, M., Kettunen, T., and Liimatainen, L. (2000). Questioning and advising in health counselling: results from a study of finnish nurse counsellors. *Health Education Journal*, 59(1):69–89.

Prabhumoye, S., Tsvetkov, Y., Salakhutdinov, R., and Black, A. W. (2018). Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

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.

Rao, S. and Tetreault, J. (2018). Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140. Association for Computational Linguistics.

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*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Raskin, N. J. and Rogers, C. R. (2005). Person-centered therapy.

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)*, pages 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*, pages 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.

Rozental, A., Carlbring, P., et al. (2014). Understanding and treating procrastination: A review of a common self-regulatory failure. *Psychology*, 5(13):1488.

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*, pages 157–162.

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

Santhanam, S., Cheng, Z., Mather, B., Dorr, B., Bhatia, A., Hebenstreit, B., Zemel, A., Dalton, A., Strzalkowski, T., and Shaikh, S. (2020). Learning to plan and realize separately for open-ended dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2736–2750, Online. Association for Computational Linguistics.

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, pages 3027–3035.

Schuller, B., Valster, M., Eyben, F., Cowie, R., and Pantic, M. (2012). Avec 2012: the continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 449–456.

Schwartz, R. (2021). The big reveal | ethical implications of therapist self-disclosure.

See, A., Roller, S., Kiela, D., and Weston, J. (2019). What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

*Technologies, Volume 1 (Long and Short Papers)*, pages 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Serban, I., Sordoni, A., Bengio, Y., Courville, A., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. (2017a). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., Kim, T., Pieper, M., Chandar, S., Ke, N. R., et al. (2017b). A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Shang, L., Lu, Z., and Li, H. (2015). Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Shang, L., Sakai, T., Lu, Z., Li, H., Higashinaka, R., and Miyao, Y. (2016). Overview of the ntcir-12 short text conversation task. In *NTCIR*.

Shang, M., Li, P., Fu, Z., Bing, L., Zhao, D., Shi, S., and Yan, R. (2019). Semi-supervised text style transfer: Cross projection in latent space. 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 4937–4946. Association for Computational Linguistics.

Sharma, A., Choudhury, M., Althoff, T., and Sharma, A. (2020a). Engagement patterns of peer-to-peer interactions on mental health platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 614–625.

Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., and Althoff, T. (2021). Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Sharma, A., Miner, A., Atkins, D., and Althoff, T. (2020b). A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Sharma, R., Wigginton, B., Meurk, C., Ford, P., and Gartner, C. E. (2017). Motivations and limitations associated with vaping among people with mental illness: A qualitative analysis of reddit discussions. *International journal of environmental research and public health*, 14(1):7.

Sheikha, F. A. and Inkpen, D. (2011). Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193.

Shen, T., Lei, T., Barzilay, R., and Jaakkola, T. (2017). Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Singer, T. and Klimecki, O. M. (2014). Empathy and compassion. *Current Biology*, 24(18):R875–R878.

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

Smedslund, G., Berg, R. C., Hammerstrøm, K. T., Steiro, A., Leiknes, K. A., Dahl, H. M., and Karlsen, K. (2011). Motivational interviewing for substance abuse. *Campbell Systematic Reviews*, 7(1):1–126.

Smith-Merry, J., Goggin, G., Campbell, A., McKenzie, K., Ridout, B., Baylosis, C., et al. (2019). Social connection and online engagement: insights from interviews with users of a mental health online forum. *JMIR mental health*, 6(3):e11084.

Song, Z., Zheng, X., Liu, L., Xu, M., and Huang, X.-J. (2019). Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3685–3695.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J.-Y., Gao, J., and Dolan, B. (2015). A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 196–205, Denver, Colorado. Association for Computational Linguistics.

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.

Spring, T., Casas, J., Daher, K., Mugellini, E., and Abou Khaled, O. (2019). Empathic response generation in chatbots. In *Proceedings of 4th Swiss Text Analytics Conference (SwissText 2019), 18-19 June 2019, Wintherthur, Switzerland*, number CONFERENCE. 18-19 June 2019.

Steel, Z., Marnane, C., Iranpour, C., Chey, T., Jackson, J. W., Patel, V., and Silove, D. (2014). The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, 43(2):476–493.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C. V., and Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Stroessner, S. J. and Benitez, J. (2019). The social perception of humanoid and non-humanoid robots: Effects of gendered and machinelike features. *International Journal of Social Robotics*, 11:305–315.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Svikhnushina, E. and Pu, P. (2022). Peace: A model of key social and emotional qualities of conversational chatbots. *ACM Transactions on Interactive Intelligent Systems*, 12(4):1–29.

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

Tatman, R. (2022). Large language models cannot replace mental health professionals.

Tian, Y., Hu, Z., and Yu, Z. (2018). Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Truong, C., Gallo, J., Roter, D., and Joo, J. (2019). The role of self-disclosure by peer mentors: Using personal narratives in depression care. *Patient education and counseling*, 102(7):1273–1279.

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., and Torous, J. B. (2019). Chatbots and conversational agents in mental health: a review of the psychiatric landscape. *The Canadian Journal of Psychiatry*, 64(7):456–464.

Vanderlyn, L., Weber, G., Neumann, M., Väth, D., Meyer, S., and Vu, N. T. (2021). "it seemed like an annoying woman": On the perception and ethical considerations of affective language in text-based conversational agents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 44–57.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017a). Attention is all you need. *Advances in neural information processing systems*, 30.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017b). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Proceedings of NeurIPS 2017*, pages 5998–6008.

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

Vinyals, O. and Le, Q. (2015b). A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning, Deep Learning Workshop*.

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

Wan, T., Jun, H., Zhang, H., Pan, W., and Hua, H. (2015). Kappa coefficient: a popular measure of rater agreement. *Shanghai archives of psychiatry*, 27(1):62.

Weegmann, M. (2002). Motivational interviewing and addiction-a psychodynamic appreciation. *Psychodynamic Practice*, 8(2):179–195.

Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.-S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., et al. (2021). Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., et al. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.

Welch, C., Lahnala, A., Perez-Rosas, V., Shen, S., Seraj, S., An, L., Resnicow, K., Pennebaker, J., and Mihalcea, R. (2020). Expressive interviewing: A conversational system for coping with COVID-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.

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*, pages 4886–4899.

Welivita, A. and Pu, P. (2022a). Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Welivita, A. and Pu, P. (2022b). Heal: A knowledge graph for distress management conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11459–11467.

Welivita, A. and Pu, P. (2023a). Boosting distress support dialogue responses with motivational interviewing strategy. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Welivita, A. and Pu, P. (2023b). Use of a taxonomy of empathetic response intents to control and interpret empathy in neural chatbots. *arXiv preprint arXiv:2305.10096*.

Welivita, A., Xie, Y., and Pu, P. (2021). A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1251–1264, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wen, T.-H., Gašić, M., Mrkšić, N., Su, P.-H., Vandyke, D., and Young, S. (2015). Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721, Lisbon, Portugal. Association for Computational Linguistics.

WHO (2022a). Mental health.

WHO (2022b). Who highlights urgent need to transform mental health and mental health care.

Wu, W. and Yan, R. (2018). Deep chit-chat: Deep learning for ChatBots. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, Melbourne, Australia. Association for Computational Linguistics.

Wu, Z., Galley, M., Brockett, C., Zhang, Y., Gao, X., Quirk, C., Koncel-Kedziorski, R., Gao, J., Hajishirzi, H., Ostendorf, M., and Dolan, B. (2021). A controllable model of grounded response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14085–14093.

Xie, X. (2017). Dipsy: A digital psychologist.

Xie, Y., Svikhnushina, E., and Pu, P. (2020). A multi-turn emotionally engaging dialog model. In *In Companion Proceedings of the 25th International Conference on Intelligent User Interfaces: 2nd workshop on user-aware conversational agents (user2agent)*.

Xing, C., Wu, Y., Wu, W., Huang, Y., and Zhou, M. (2018). Hierarchical recurrent attention network for response generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Xu, C., Wu, W., and Wu, Y. (2018). Towards explainable and controllable open domain dialogue generation with dialogue acts. *arXiv preprint arXiv:1807.07255*.

Xu, R., Ge, T., and Wei, F. (2019). Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv:1903.06353*.

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.

Yeh, S.-L., Lin, Y.-S., and Lee, C.-C. (2019). An interaction-aware attention network for speech emotion recognition in spoken dialogs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689. IEEE.

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. (2020a). Aser: A large-scale eventuality knowledge graph. *Proceedings of The Web Conference 2020*, page 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*, pages 5276–5289.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022). Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y., Sun, S., Galley, M., Chen, Y.-C., Brockett, C., Gao, X., Gao, J., Liu, J., and Dolan, B. (2020b). DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhao, T., Zhao, R., and Eskenazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664, Vancouver, Canada. Association for Computational Linguistics.

Zheng, C. and Huang, M. (2021). Exploring prompt-based few-shot learning for grounded dialog generation. *arXiv preprint arXiv:2109.06513*.

Zhong, P., Wang, D., and Miao, C. (2019). An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7492–7500.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018a). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018b). Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4623–4629.

Zhou, L., Gao, J., Li, D., and Shum, H.-Y. (2020a). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Zhou, V., Mistriotis, D., and Shestopalov, V. (2020b). profanity-check.

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

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*.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). A robustly optimized BERT pre-training approach with post-training. In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

Zillmann, D. (2008). Empathy theory. *The International Encyclopedia of Communication*.

// Kalpani Anuradha Welivita

Rue Centrale 44, Chavannes-près-Renens, 1022, Switzerland.

(+41) 78 350 08 30 | kalpani.welivita@epfl.ch | anuradhawelivita | anuradha1992

EDUCATION

**Doctoral Degree in Computer and Communication Sciences** *Sep 2018-Present*
*School of Computer and Communication Sciences, EPFL, Switzerland*
- Advised by Dr. Pearl Pu, the head of the HCI research group at EPFL.
- Doctoral thesis research areas : Human-Computer Interaction, Natural Language Processing, Conversational Agents, Empathetic Response Generation, Chatbots for Distress Consolation.

**Honours Degree of Bachelor of the Science of Engineering** *Mar 2013-Apr 2017*
*Department of Computer Science and Engineering, University of Moratuwa, Sri Lanka*
- **GPA : 4.12/4.20**
- Ranked **1st in the Department of Computer Science and Engineering** out of 125 students.
- **Batch topper in the Faculty of Engineering** out of 930 students.

WORK EXPERIENCE

- **Research and Teaching Assistant** *Feb 2019-Present*
  *HCI Group, EPFL, Switzerland*
  - Analyzed human-human conversations to discover strategies used by humans to empathize with others and utilize them in developing empathetic response generation models.
  - Curated a large-scale empathetic dialogues dataset out of movie subtitles and used it in training a dialogue-emotion classifier in a semi-supervised manner.
  - Constructed a knowledge-graph for distress consolation by analyzing Reddit dialogues to help retrieval based chatbots to generate responses that are more empathetic, diverse, and reliable.
  - Investigated psychotherapeutic discourse to make distress-consoling chatbot responses more compliant with strategies used in therapy.
  - Assisted in the courses **"Introduction to database systems"**, **"Information, Computation, Communication"**, **"Practice of object-oriented programming"**, **"Introduction to natural language processing"**, and **"Interaction Design"**. Engaged in helping students in exercise sessions, preparing quiz questions, and grading exams.
- **Research Assistant** *Sep 2018-Jan 2019*
  *Salathe Group, EPFL, Switzerland*
  - Developed a crowdsourcing experiment to label and trace boundaries of food images. Built on top of Microsoft COCO (Common Objects in Context) UI.
  - Constructed a pipeline for food image segmentation and classification using Mask R-CNN and Inception V3.
- **Lecturer (Contract)** *Mar 2017-Aug 2018*
  *University of Moratuwa, Sri Lanka.*
  - Engaged in co-teaching and evaluating courses **"Database Systems"**, **"Modular Software Development"**, and **"Communication Skills"**.
  - Lab coordinator of the course Programming Fundamentals with over 800 registered students.
- **Trainee Software Engineer** *Oct 2015-Mar 2016*
  *99X Technology, Sri Lanka.*
  - Constructed a machine learning platform to generate adaptive user interfaces.
  - Worked in a Scrum based Agile software development team and developed a web based form management module.
- **Google Summer of Code Student** *May 2015-Aug 2015*
  *The Fedora Project.*
  - Worked on a complete user-experience and functionality overhaul of the AskFedora website.

SELECTED PUBLICATIONS

**Welivita, A.** and Pu, P., "Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy," In *Findings of the Association for Computational Linguistics (ACL)*, 2023. [PDF]

**Welivita, A.**, Yeh, C., and Pu, P., "Empathetic Response Generation for Distress Support," Accepted in *24th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, 2023.

**Welivita, A.** and Pu, P., "Curating a Large-Scale Motivational Interviewing Dataset Using Peer Support Forums," in *Proceedings of the 29th International Conference on Computational Linguistics (COLING)*, 2022. [PDF]

Svikhnushina, E., Voinea, I., **Welivita, A.**, and Pu, P. "A Taxonomy of Empathetic Questions in Social Dialogs," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022. [PDF]

**Welivita, A.** and Pu, P., "HEAL : A Knowledge Graph for Distress Management Conversations," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, 2021. [PDF]

**Welivita, A.**, Xie, Y., and Pu, P., "A Large-Scale Dataset for Empathetic Response Generation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. [PDF]

**Welivita, A.** and Pu, P., "A Taxonomy of Empathetic Response Intents in Human Social Conversations," in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020. [PDF]

**Welivita, A.**, Perera, I., Meedeniya, D., Wickramarachchi, A. and Mallawaarachchi, V., "Managing Complex Workflows in Bioinformatics-An Interactive Toolkit with GPU Acceleration," in *IEEE Transactions on NanoBioscience*, 2018. [PDF]

**Welivita, A.**, Perera, I. and Meedeniya, D., "An Interactive Workflow Generator to Support Bioinformatics Analysis through GPU Acceleration," in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2017. [PDF]

## Academic Activities

- Reviewer for **Conference on Empirical Methods in Natural Language Processing (EMNLP) 2022**.
- Presenting author, and virtual attendee at the **COLING 2022, AAAI 2021, EMNLP 2021, and COLING 2020** conferences.
- Presenting author, volunteer, and attendee at the **11th IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 2017**, Kansas, USA.
- Presenter at the student research consortium of the **16th International Conference on Human-Computer Interaction (INTERACT) 2017**, Mumbai, India.

## Honours and Awards

- **EDIC Doctoral Fellowship** *2018*
  *EPFL, Switzerland*
  Selected as one of the top 9% out of 550 international applicants.
- **Gold Medalist - Department of Computer Science and Engineering** *2017*
  *University of Moratuwa*
  Awarded for the Computer Science and Engineering graduand who obtains the highest overall GPA of 3.8 or above at the B.Sc. Engineering Honours Degree examinations.
- **Valedictorian at the General Convocation of University of Moratuwa** *2017*
  *University of Moratuwa*
  For obtaining the highest score at the evaluation of the "Most Outstanding Student of the Year" award from among four departments at the Faculty of Engineering, University of Moratuwa.

## Projects

**Empathetic Dialogue Agent for Consumer Health Question Answering** *Feb 2019-Jun 2019*
*Semester Project, EPFL*

- Development of a dialogue agent that can succintly answer consumer health questions in an empathetic manner.
- Research areas : Natural Language Processing, Conversational agents, Machine Learning
- Technologies used : Python, TensorFlow, Keras

**Instance Segmentation and Classification of Food Images** *Sep 2018-Jan 2019*
*Semester Project, EPFL*

- Training and evaluation of neural network models for instance segmentation and classification of food images, as a part of the MyFoodRepo project (www.myfoodrepo.org).
- Research areas : Computer Vision, Machine Learning
- Technologies used : Python, TensorFlow, Keras, Amazon Mechanical Turk

**Interactive Workflow Generator to Support Bioinformatics Analyses** *Mar 2017-Aug 2018*
*University of Moratuwa*

- Development of a GUI based, interactive software solution with GPU support for simple construction and efficient execution of bioinformatics workflows.
- Research areas : Bioinformatics Software
- Technologies used : AngularJS, NodeJS, Amazon Web Services

**Real-time Face Tracking and Head Pose Estimation for Augmented Reality in Mobile Platforms** *May 2016-Mar 2017*
*Final Year Project, University of Moratuwa*

- Development of a face tracking and head pose estimation plugin for the Unity3D game engine, which can be used in building mobile augmented reality applications.
- Research areas : Computer Vision, Mobile Augmented Reality
- Technologies used : MATLAB, C#, Unity3D, OpenCV

SKILLS

**Programming languages :** Python, Java, C, JavaScript, MATLAB, R, SQL, HTML, CSS

**Programming tools :** PyCharm, Microsoft Visual Studio, Netbeans IDE, IntelliJ IDEA, Git

**Frameworks & Libraries :** Pandas, Keras, TensorFlow, React Native

**Platforms :** AWS, Linux, Windows, MacOS

**Languages :** Sinhalese (Native), English (Advanced - C1), French (Beginner - A1/A2)

**Other :** Quick learning, Presentation, Communication, Time management, Teamwork