Boosting Distress Support Dialogue Responses with Motivational Interviewing Strategy

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

Abstract

AI-driven chatbots have become an emerging solution to address psychological distress. Due to the lack of psychotherapeutic data, researchers use dialogues scraped from online peer support forums to train them. But since the responses in such platforms are not given by professionals, they contain both conforming and non-conforming responses. In this work, we attempt to recognize these conforming and non-conforming response types present in online distress-support dialogues using labels adapted from a well-established behavioral coding scheme named Motivational Interviewing Treatment Integrity (MITI) code and show how some response types could be rephrased into a more MI adherent form that can, in turn, enable chatbot responses to be more compliant with the MI strategy. As a proof of concept, we build several rephrasers by fine-tuning Blender and GPT3 to rephrase MI non-adherent *Advise without permission* responses into *Advise with permission*. We show how this can be achieved with the construction of pseudo-parallel corpora avoiding costs for human labor. Through automatic and human evaluation we show that in the presence of less training data, techniques such as prompting and data augmentation can be used to produce substantially good rephrasings that reflect the intended style and preserve the content of the original text.

1 Introduction

Demands of the modern world are increasingly responsible for causing severe psychological distress in people. World Health Organization estimates psychological distress affects 29% of people in their lifetime (Steel et al., 2014). The shortage of mental health workers and the stigma associated with mental health further demotivates people from actively seeking help. With the expansion of the internet, many people are seen resorting to peer support platforms such as Reddit and Talklife to vent their distress.[1] The anonymity associated with these platforms makes it easier for people to discuss their concerns without being affected by the stigma. Distress consolation through AI-driven chatbots has also become an emerging solution (Fitzpatrick et al., 2017; Inkster et al., 2018; Mousavi et al., 2021). Due to the lack of availability of large-scale psycho-therapeutic conversations, researchers are using data scraped from online peer support forums to train such chatbots (Alambo et al., 2019; Welivita and Pu, 2022). High levels of perceived empathy and information richness make them good candidates for training (Nambisan, 2011; De Choudhury and De, 2014; Sharma et al., 2020a,b). 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

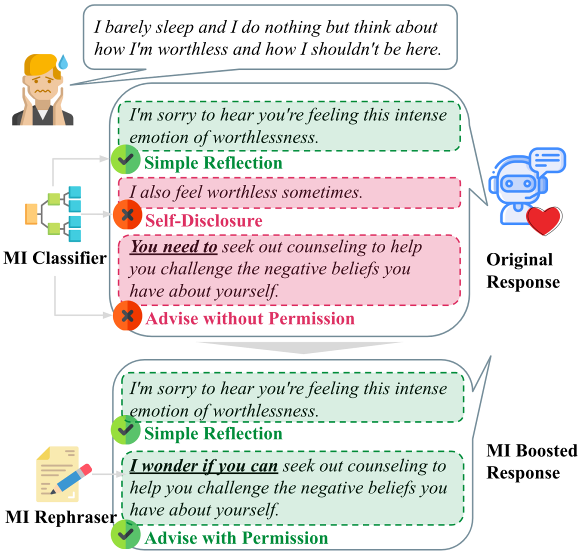

I barely sleep and I do nothing but think about how I'm worthless and how I shouldn't be here.
I'm sorry to hear you're feeling this intense emotion of worthlessness.
Simple Reflection
I also feel worthless sometimes.
Self-Disclosure
You need to seek out counseling to help you challenge the negative beliefs you have about yourself.
Advise without Permission
MI Classifier
Original Response
I'm sorry to hear you're feeling this intense emotion of worthlessness.
Simple Reflection
I wonder if you can seek out counseling to help you challenge the negative beliefs you have about yourself.
Advise with Permission
MI Rephraser
MI Boosted Response

Figure 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.

[1] www.reddit.com; www.talklife.com

Findings of the Association for Computational Linguistics: ACL 2023, pages 5411–5432
July 9-14, 2023 ©2023 Association for Computational Linguistics

that appear in distress support dialogues and developing automatic means that can propose omission or rephrasing of such unfavourable response types. Figure 1 shows an example.

To analyze the types of responses in distress support dialogues, we use labels adapted from a well-established behavioral coding system named Motivational Interviewing Treatment Integrity (MITI) code (Moyers et al., 2014). It is used in psychology to evaluate how well a mental health provider responds. 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). It defines favourable response types such as *Questioning*, *Reflecting*, and *Advising with permission* and unfavourable response types such as *Advising without permission*, *Confronting*, and *Self-Disclosing (extra-session)*. In our previous work, we developed a dataset called the MI dataset, to have a comparative understanding of the differences between online support provided by peers and trained counselors. For this, we hired 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. 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., 2020). 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. Our code and the datasets can be found at `https://github.com/anuradha1992/Boosting-with-MI-Strategy`

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 subtask and releasing an 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.

3 Datasets

For this work, we used dialogues curated from two online support platforms. The first one is CounselChat (counselchat.com), in which verified counselors respond to distress-related posts. The CounselChat dataset available publicly [2] contains 2,129 post-response pairs spanning 31 distress-related topics. We also curated dialogues from a carefully selected set of 8 subreddits: *mentalhealthsupport*; *offmychest*; *sad*; *suicidewatch*; *anxietyhelp*; *depression*; *depressed*; and *depression_help*, which are popular among Reddit users to vent their distress. This dataset, which we call RED (Reddit Emotional Distress), contains 1,275,486 dyadic conversations having on average of 2.66 turns per dialogue.

In our previous work, we recruited professional counselors to annotate a subset of 1,000 dialogues each from CounselChat and RED datasets with labels adapted from the MITI code 2.0 (Moyers et al., 2003) and 4.2.1 (Moyers et al., 2014). We call this the MI dataset. We used 15 labels for annotation. They are elaborated in the appendices. Out of them, 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 the appendices. 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 vot-

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

ing 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. Using this method, we were able to automatically annotate 2,881 and 1,196,012 sentences in CounselChat and RED datasets, respectively.

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 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.

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 (Liu et al., 2019). 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% on 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.

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 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*.

5.1 Pseudo-Parallel Corpora

As discussed in Section 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.

5.1.1 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 11 shows some 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.*

Table 1: Examples of templates corresponding to *Advise without Permission* and *Advise with Permission* responses. The full list is included in the appendices.

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.

5.1.2 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, we gauge the quality of the above corpora using human ratings.

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., 2020). 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 $\approx$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., 2020). 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 2 shows training examples with generic and N-gram-based prompts.

Altogether we developed 10 different rephrasing models by fine-tuning Blender and GPT-3 on: 1)

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 2: Examples with generic and N-gram prompts.

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. Some examples of the rephrased output by these different models are shown in the appendices.

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 and Och, 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.

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 test samples.

Table 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 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 incorporated 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 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

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 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 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.

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.

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.

Table 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 when it comes to the smaller PP training corpus, using generic prompting during training increases the scores across most cases. But when it comes to the larger PPA corpus, simply training the rephrasers with template-replaced pseudo-parallel pairs gives the best results irrespective of the underlying backbone model.

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.

8 Discussion

In this paper, we 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 MI adherent response type *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.

In the future, we hope to develop a chatbot that can respond to psychological distress using the RED dataset that contain dialogues curated from several mental health-related subreddits. Then we hope to improve the responses generated by this chatbot by applying MI boosting at two different levels: one at the data level; and the other at the model level. At data level boosting, we hope to apply the MI classifier and automatically label the responses in the training data itself. By doing so, we will be 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.

9 Limitations

Certain parts of our proposed methodology, for example, template-based replacement and n-gram-based prompting are applicable only when style-specific linguistic attributes could be identified between the source and the target text. And due to the cost of human labor and the lack of publicly available client-therapist dialogues, the sample size drawn in the study is small and thus may have an impact on the conclusions drawn. Our methods have only been tested for the English language. But we believe similar methods could be applied to other languages given they have unparallel corpora tagged with *Advise without Permission* and *Advise with Permission* labels. The rephrasing methods described in this paper are tested for short sentences with a maximum sentence length of 98 tokens. Thus, the scalability of these methods for long text still remains to be tested.

When testing the rephrasers, there are some combinations that could be tried other than the ones already tested. For example, more models can be fine-tuned and tested separately on template-replaced and retrieval-based PP and PPA corpora but incorporating generic and N-gram prompting. In this work, we first combined these two types of corpora before attempting prompting since we could observe better performance on Blender when the corpora were combined.

In order to have more data, we combined the *Advise with Permission* and *Advise without Permission* responses present in CounselChat and RED datasets. But studies show that there are differences in the language used by counselors and peers (Lahnala et al., 2021; Mousavi et al., 2021). So, there can be linguistic differences between the same type of response in CounselChat and RED datasets. Future work should attempt to identify these differences and ideally rephrase the responses given by peers to reflect the language of the counselors.

10 Ethics Statement

Data Curation: Only publicly available data in Reddit and CounselChat websites were used in this work. Analysis of posts on websites such as Reddit is considered "fair play" since individuals are anonymous and users are aware their responses remain archived on the site unless explicitly deleted. It is also stated in Reddit's privacy policy that it allows third parties to access public Reddit content. [3] Also, Reddit's data is already widely available in larger dumps such as Pushshift (Baumgartner et al., 2020). Even though the policies allow it, it should be thoroughly noted that this data contains sensitive information. Thus, we adhere to the guidelines suggested by Benton et al. (2017) for working with social media data in health research, and 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. In addition, references to usernames as well as URLs are removed from dialogue content for de-identification.

Human Evaluation: The human raters recruited from the crowdsourcing platform, UpWork, were all trained in the practice of counseling. Since the methods were tested on English-only text, we recruited workers who had professional competency in the English language. We paid them $10 for evaluating each batch of rephrased sentences that required on average ≈30 minutes to complete. Thus, the amount paid to the human raters was ≈2.75 times above the US minimum wage of $7.25 per hour. We also paid an extra $2 as a bonus per each batch for workers who obtained an above-average agreement with the other worker who rated the same batch.

Chatbots for Distress-Consolation: One of the main applications of the proposed methodology is boosting chatbot responses for distress consolation with motivational interviewing strategy. Using chatbots for distress consolation or other mental health interventions has raised ethical concerns among many (Lanteigne, 2019; Montemayor et al., 2021; Tatman, 2022). However, chatbots that intervene in mental health-related matters have already been developed and have been quite popular for a while. Some examples are SimSensei (DeVault et al., 2014), Dipsy (Xie, 2017), Woebot (`woebothealth.com`), and Wysa (`www.wysa.io`). Czerwinski et al. (2021) state, *About 1 billion people globally are affected by mental disorders; a scalable solution such as an AI therapist could be a huge boon.* The current technology to develop such chatbots rely heavily on deep learning and pre-trained language models. But due to the inherently unpredictable nature of these models, they

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

pose a threat of delivering unfavourable responses when such chatbots are used for distress consolation. We believe the methodology we suggest in this work can help them become more reliable and fail-safe by adhering to the motivational interviewing strategy, a guiding style of communication heavily practiced in psychotherapy. However, since the unfavourable response detection and rephrasing methods still rely on neural network models, the artifacts produced in this paper should be used for research purposes only and real-world deployment of them should be done under human supervision.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Amanuel Alambo, Manas Gaur, Usha Lokala, Ugur Kursuncu, Krishnaprasad Thirunarayan, Amelie Gyrard, Amit Sheth, Randon S Welton, and Jyotishman Pathak. 2019. Question answering for suicide risk assessment using reddit. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 468–473. IEEE.

Satanjeev Banerjee and Alon Lavie. 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.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 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.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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

Mary Czerwinski, Javier Hernandez, and Daniel McDuff. 2021. Building an ai that feels: Ai systems with emotional intelligence could learn faster and be more helpful. *IEEE Spectrum*, 58(5):32–38.

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

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, 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.

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

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 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):e7785.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3137–3146.

Jacques Gaume, Gerhard Gmel, Mohamed Faouzi, and Jean-Bernard Daeppen. 2009. Counselor skill influences outcomes of brief motivational interventions. *Journal of substance abuse treatment*, 37(2):151–159.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Yufang Huang, Wentao Zhu, Deyi Xiong, Yiye Zhang, Changjian Hu, and Feiyu Xu. 2020. Cycle-consistent adversarial autoencoders for unsupervised text style transfer. *arXiv preprint arXiv:2010.00735*.

Becky Inkster, Shubhankar Sarda, Vinod Subramanian, et al. 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.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Zhijing Jin, Di Jin, Jonas Mueller, Nicholas Matthews, and Enrico Santus. 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.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR.

Allison Lahnala, Yuntian Zhao, Charles Welch, Jonathan K. Kummerfeld, Lawrence C An, Kenneth Resnicow, Rada Mihalcea, and Verónica Pérez-Rosas. 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.

Camylle Lanteigne. 2019. Social robots and empathy: The harmful effects of always getting what we want.

Juncen Li, Robin Jia, He He, and Percy Liang. 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.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Ruibo Liu, Chongyang Gao, Chenyan Jia, Guangxuan Xu, and Soroush Vosoughi. 2022. Non-parallel text style transfer with self-parallel supervision. *arXiv preprint arXiv:2204.08123*.

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

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 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.

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

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84.

Remi Mir, Bjarke Felbo, Nick Obradovich, and Iyad Rahwan. 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.

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

Seyed Mahed Mousavi, Alessandra Cervone, Morena Danieli, and Giuseppe Riccardi. 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.

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

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

Priya Nambisan. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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.

Verónica Pérez-Rosas, Xuetong Sun, Christy Li, Yuchen Wang, Kenneth Resnicow, and Rada Mihalcea. 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)*.

Maja Popović. 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.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. *arXiv preprint arXiv:1804.09000*.

Sudha Rao and Joel Tetreault. 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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Metho0ds in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 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, Online. Association for Computational Linguistics.

Robert Schwartz. 2021. The big reveal | ethical implications of therapist self-disclosure.

Mingyue Shang, Piji Li, Zhenxin Fu, Lidong Bing, Dongyan Zhao, Shuming Shi, and Rui Yan. 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.

Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. 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.

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

Fadi Abu Sheikha and Diana Inkpen. 2011. Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 187–193.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. *Advances in neural information processing systems*, 30.

Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel, and Derrick Silove. 2014. The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013. *International journal of epidemiology*, 43(2):476–493.

Rachael Tatman. 2022. [link].

Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. Structured content preservation for unsupervised text style transfer. *arXiv preprint arXiv:1810.06526*.

TANG Wan, HU Jun, Hui Zhang, WU Pan, and HE Hua. 2015. Kappa coefficient: a popular measure of rater agreement. *Shanghai archives of psychiatry*, 27(1):62.

Anuradha Welivita and Pearl Pu. 2022. Heal: A knowledge graph for distress management conversations.

Xing Xie. 2017. Dipsy: A digital psychologist.

Ruochen Xu, Tao Ge, and Furu Wei. 2019. Formality style transfer with hybrid textual annotations. *arXiv preprint arXiv:1903.06353*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

A Datasets

A.1 The RED (Reddit Emotional Distress) Dataset

The RED dataset is curated from carefully selected 8 mental health-related subreddits in Reddit. According to the latest statistics, 61% of Reddit users are male. Of the users, 48% are from the United States. People aged 18-29 make up Reddit's largest user base (64%). The second biggest age group is 30-49 (29%). Only 7% of Reddit users are over 50. It should be noted that these demographic biases can subtly skew our data and models from representing average human behavior. The data we curated were English-only and they may perpetuate an English bias in NLP systems.

A.2 The MI Dataset

Altogether, 15 labels adapted from the MITI code 2.0 (Moyers et al., 2003) and 4.2.1 (Moyers et al., 2014) were used for annotation. They included *Closed Question*, *Open Question*, *Simple Reflection*, *Complex Reflection*, and *Give Information*, which are generally considered favourable. They also included labels recognized specifically as MI adherent, which are *Advise with Permission*, *Affirm*, *Emphasize Autonomy*, and *Support*. There are another four labels recognized as MI non-adherent, which are *Advise without Permission*, *Confront*, *Direct*, and *Warn*. We also included two other labels *Self-Disclose* and *Other*, which are not included in the MITI code. The label *Self-Disclose* was included because, in peer support conversations, peers are mostly seen to share their lived experiences. Though it is believed that *Self-Disclosure* contributes in building rapport between the speaker and listener, 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. Thus, it is important to be able to recognize this response type.

Table 5 shows the full list of labels we adapted from the MITI code along with descriptions and examples. Table 6 shows the statistics of the annotated responses in the MI dataset, corresponding to each label.

A.3 Data Augmentation: N-gram Based Matching

We denote examples of the most frequent N-grams corresponding to each label in Table 7. 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.

Table 8 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.

A.4 Data Augmentation: Similarity Based Retrieval

To derive semantically meaningful sentence embeddings that can be compared using cosine-similarity, we used Sentence-BERT (SBERT) proposed by Reimers and Gurevych (2019), which uses siamese and triplet network structures to compute sentence embeddings. Among several models the authors have proposed, we used the *roberta-base-nli-stsb-mean-tokens* model, fine-tuned on the NLI (Bowman et al., 2015) and STS benchmark (STSb) (Cer et al., 2017) datasets, since it has reported a high Spearman's rank correlation of 84.79 ± 0.38 between the cosine-similarity of the sentence embeddings and the gold labels in the STS benchmark test set outperforming the existing state-of-the-art. It is also more efficient to use than *roberta-large*.

As described in Section 3, we used majority voting followed by computing the average similarity of retrieved sentences with the same label (in case of ties) to choose the final label for an unlabeled sentence. In Figure 2, we show an example elaborating this procedure.

Table 8 shows the statistics of the labels extended through similarity-based retrieval in CC and RED datasets.

A.5 Augmented MI Datasets

Table 9 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.

B MI Classifier

We used the same hyper-parameter setting used in RoBERTa (Liu et al., 2019) when training the MI classifier. We used the Adam optimizer with β_1 of 0.9, β_2 of 0.98, an ϵ value of 1×10^{-6}, and a learning rate of 2×10^{-5}. A dropout of 0.1 was used

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?* *Did you use herion this week?*
2. Open Question	Questions that allow a wide range of possible answers. It may seek information or may invite the speaker's perspective or may encourage self-exploration.	*What do you think are the advantages of changing this behavior?* *What is your take on that?*
3. Simple Reflection	Simple reflections include repetition, rephrasing, or paraphrasing of speaker's previous statement. It conveys understanding or facilitate speaker-listener exchanges.	*It seems that you are not sure what is going to come out of this talk.* *It sounds like you're feeling worried.*
4. Complex Reflection	Complex reflections include repeating or rephrasing the previous statement of the speaker but adding substantial meaning or emphasis to it. It serves the purpose of conveying a deeper or more complex picture of what the speaker has said.	**Speaker:** *Mostly, I would change for future generations. If we waste everything, then there will be nothing left.* **Listener:** *It sounds like you have a strong feeling of responsibility.*
5. Give Information	The listener gives information, educates, provides feedback, or gives an opinion without advising.	*This assignment on logging your cravings is important because we know that cravings often lead to relapses.*
MI Adherent Behaviour Codes:		
6. Advise with Permission	Advising when the speaker asks directly for the information or advice. Indirect forms of permission can also occur, such as when the listener invites the speaker to disregard the advice as appropriate.	*If you agree with it, we could try to brainstorm some ideas that might help you.*
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.	*Yes, you're right. No one can force you to stop drinking.* *It is really up to you to decide.*
9. Support	Supporting the client with statements of compassion or sympathy.	*I'm here to help you with this* *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 turn the computer off during breaks.*
11. Confront	Directly and unambiguously disagreeing, arguing, correcting, shaming, blaming, criticizing, labeling, moralizing, ridiculing, or questioning the speaker's honesty.	*You think that is any way to treat people you love?* *Yes, you are an alcoholic. You might not think so, but you are.*
12. Direct	Giving the speaker orders, commands, or imperatives.	*Don't do that!* *Keep track of your cravings, using this log, and bring it in next week to review with me.*
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 but after maturing some I got over that.*
15. Other	All other statements that are not classified under any of the above codes	*Good morning.* *Hi there.*

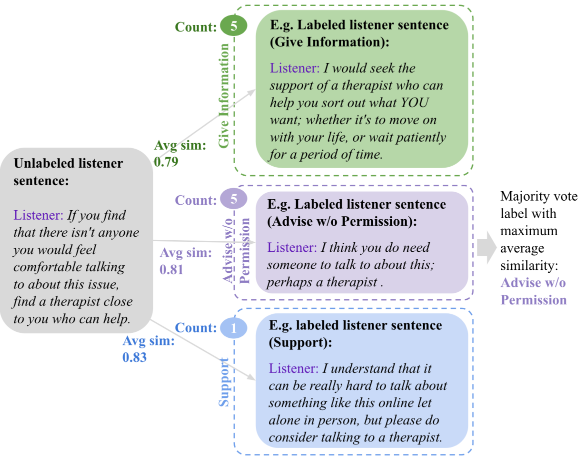
Table 5: The set of labels adapted from the MITI code that the MI classifier is able to recognize.

on all layers and attention weights, and a GELU activation function (Hendrycks and Gimpel, 2016). We limited the maximum number of input tokens to 100, and used a batch size of 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.

Figure 3 shows the architectural diagram of the

Label	# Labels in CC	# Labels in RED	Total
Closed Question	500	405	905
Open Question	264	212	476
Simple Reflection	304	252	556
Complex Reflection	732	562	1,294
Give Information	3,643	1213	4,856
MI Adherent Behavior Codes:			
Advise w/ Permission	417	67	484
Affirm	428	517	945
Emphasize Autonomy	152	101	253
Support	418	815	1,233
MI Non-Adherent Behavior Codes:			
Advise w/o Permission	1,414	871	2,285
Confront	142	176	318
Direct	460	438	898
Warn	67	46	113
Other:			
Self-Disclose	174	1216	1,390
Other	513	292	805
Total	9,628	7,183	16,811

Table 6: Statistics of human annotated MITI labels in CounselChat (CC) and RED datasets.



Unlabeled listener sentence:

Listener: *If you find that there isn't anyone you would feel comfortable talking to about this issue, find a therapist close to you who can help.*

E.g. Labeled listener sentence (Give Information):

Listener: *I would seek the support of a therapist who can help you sort out what YOU want; whether it's to move on with your life, or wait patiently for a period of time.*

E.g. Labeled listener sentence (Advise w/o Permission):

Listener: *I think you do need someone to talk to about this; perhaps a therapist .*

E.g. labeled listener sentence (Support):

Listener: *I understand that it can be really hard to talk about something like this online let alone in person, but please do consider talking to a therapist.*

Majority vote label with maximum average similarity: **Advise w/o Permission**

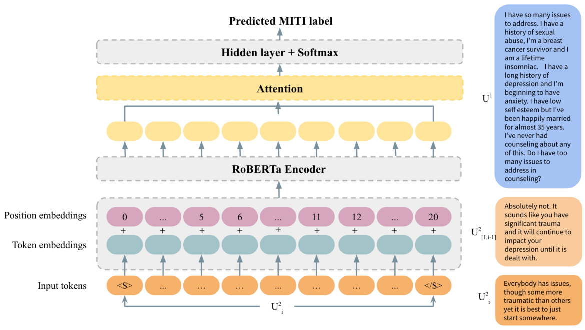
Figure 2: An example of automatically labeling an unlabeled sentence by computing the cosine-similarity with labeled sentences. The label is chosen based on majority voting. But this example shows a tie. Thus, we compute the average similarity of the sentence clusters that hold a tie and select the label of the sentence cluster with the maximum average similarity.

MI classifier used for annotation. Table 10 shows the performance scores of the MI classifier when trained on gold-labeled and augmented MI datasets.

C MI Rephraser

C.1 Construction of pseudo-parallel corpora

Table 11 denotes the full list of templates corresponding to *Advise without Permission* and *Advise with Permission* responses that were used in the process of creating pseudo-parallel corpora using the template-based replacement method.



I have so many issues to address. I have a history of sexual abuse, I'm a breast cancer survivor and I am a lifetime insomniac. I have a long history of depression and I'm beginning to have anxiety. I have low self esteem but I've been happily married for almost 35 years. I've never had counseling about any of this. Do I have too many issues to address in counseling?

Absolutely not. It sounds like you have significant trauma and it will continue to impact your depression until it is dealt with.

Everybody has issues, though some more traumatic than others yet it is best to just start somewhere.

Figure 3: The architecture of the MI classifier.

In Figure 4, we visualize the process of creating Pseudo-Parallel (PP) and Pseudo-Parallel Augmented (PPA) corpora along with statistics corresponding to each dataset.

C.2 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., 2017). All the models were fine-tuned for 200 epochs, with a batch size of 8, and a learning rate of 1 × 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.

Table 12 shows some examples of rephrased sen-

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

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

tences by the different rephraser models we fine- tuned.

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/ Per.	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 Per.	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 9: Statistics of the annotated responses in MI Augmented (Intersection) and MI Augmented (Union) datasets.

Dataset	Size		Optimal Epoch	Train Loss	Valid Acc. (%)	Test	
						Acc. (%)	F1-score (weighted avg.)

MI Gold	Train:
	Valid (Gold):
	Test (Gold):
		13,449
		1,681
		1,681
			7	0.3002	67.08	68.31	68.07

MI Augmented (Intersection)	Train:
	Valid (Gold):
	Test (Gold):
		80,690
		1,681
		1,681
			2	0.2277	64.07	67.13	65.85

MI Augmented (Union)	Train:
	Valid (Gold):
	Test (Gold):
		1,375,107
		1,681
		1,681
			13	0.1324	**72.67**	**73.44**	**72.92**

Table 10: 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.

Advise without Permission	Advise with Permission
- *You can* (verb) ___ - *You could* (verb) ___ - *You need to* (verb) ___ - *You should* (verb) ___ - (Verb) ___ - *You can try to* (verb) ___ - *I think you should* (verb) ___ - *I suggest that you* (verb) ___ - *I suggest you* (verb) ___ - *Maybe you can* (verb) ___ - *Maybe you could* (verb) ___	- *It maybe helpful to* (verb) ___ - *You may want to* (verb) ___ - *I encourage you to* (verb) ___ - *Perhaps you can* (verb) ___ - ___*, if you would like.* - *It would be good idea to* (verb) ___ - *It may be important to* (verb) ___ - *I would encourage you to* (verb) ___ - *I wonder if you can* (verb) ___ - *Maybe it is important to* (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 11: Linguistic templates corresponding to *Advise without Permission* and *Advise with Permission* responses.

D Automatic Evaluation

We used the the NLTK package to compute the BLEU [4], METEOR [5], and chrf [6] scores. The

[4] https://www.nltk.org/_modules/nltk/translate bleu_score.html

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*

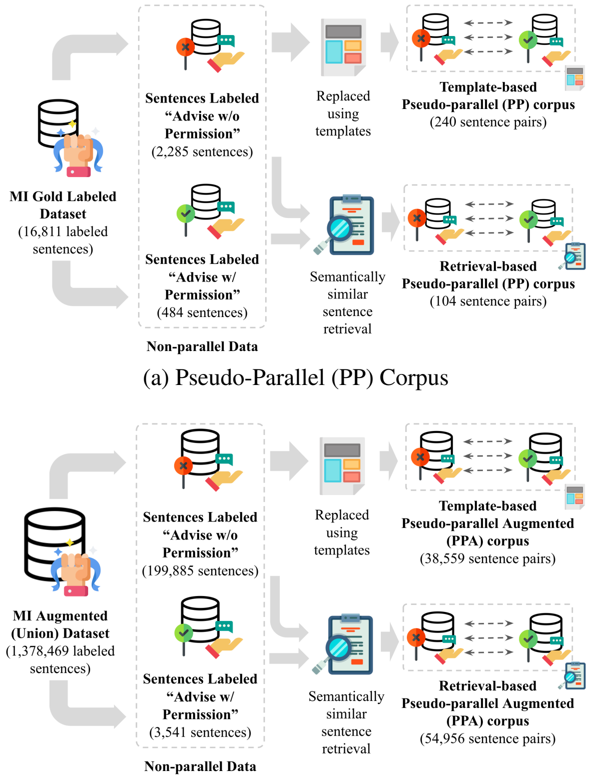
Template+Retrieval-based (with N-gram prompting):
Prompt: *able to help you*

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 12: Examples of rephrased sentences by the different rephraser models we developed

[5] https://www.nltk.org/_modules/nltk/translate/meteor_score.html

[6] https://www.nltk.org/_modules/nltk/translate/chrf_score.html


MI Gold Labeled Dataset (16,811 labeled sentences)
Sentences Labeled "Advise w/o Permission" (2,285 sentences)
Sentences Labeled "Advise w/ Permission" (484 sentences)
Non-parallel Data
Replaced using templates
Semantically similar sentence retrieval
Template-based Pseudo-parallel (PP) corpus (240 sentence pairs)
Retrieval-based Pseudo-parallel (PP) corpus (104 sentence pairs)

(a) Pseudo-Parallel (PP) Corpus

MI Augmented (Union) Dataset (1,378,469 labeled sentences)
Sentences Labeled "Advise w/o Permission" (199,885 sentences)
Sentences Labeled "Advise w/ Permission" (3,541 sentences)
Non-parallel Data
Replaced using templates
Semantically similar sentence retrieval
Template-based Pseudo-parallel Augmented (PPA) corpus (38,559 sentence pairs)
Retrieval-based Pseudo-parallel Augmented (PPA) corpus (54,956 sentence pairs)

(b) Pseudo-Parallel Augmented (PPA) Corpus

Figure 4: Pseudo-Parallel (PP) and Pseudo-Parallel Augmented (PPA) corpus construction.

ROUGE score and the BERTscore were computed using the rouge [7] and bert_score [8] python libraries, respectively. The POS distance was calculated as mentioned in the work by Tian et al. (2018) following the code released by the authors on github.[9] For computing the Word Mover Distance (WMD), we used Gensim's implementation of the WMD. [10] 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 (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.

E Human Evaluation

Figures 5, 6, and 7 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.

F Other Remarks

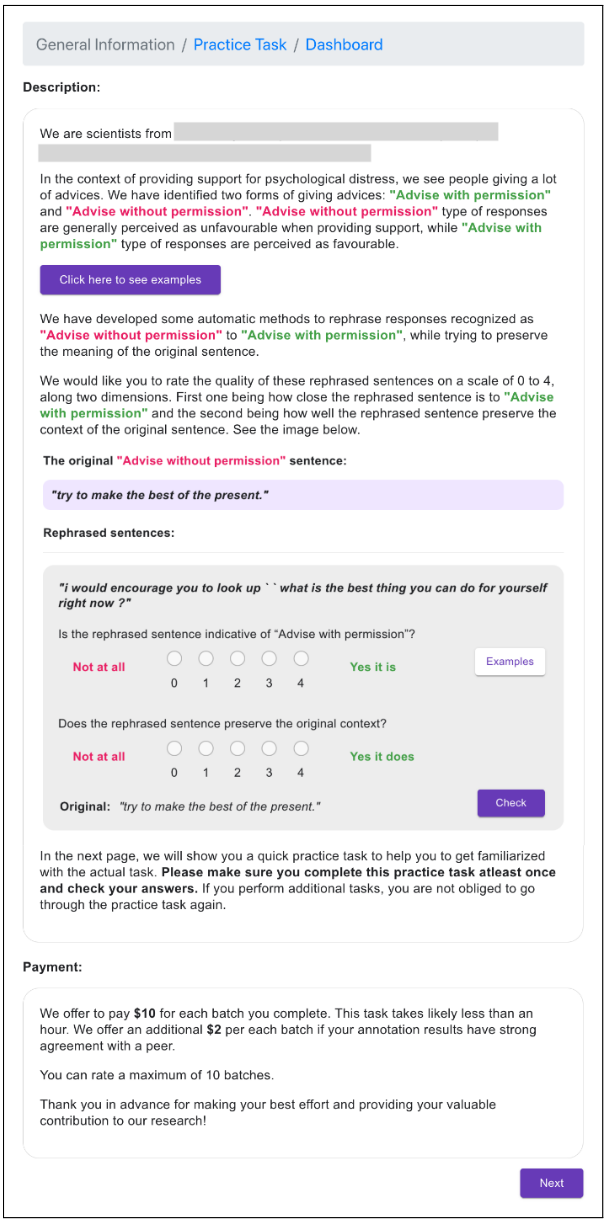
In human evaluation results, we observed in 97.5% of the cases, the average scores obtained for style transfer strength are better than the average scores obtained for semantic similarity. This observation is invariant of the type of backbone model used in training. This implies template-based and retrieval-based 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

[7] https://pypi.org/project/rouge/

[8] https://pypi.org/project/bert-score/

[9] https://github.com/YouzhiTian/Structured-Content-Preservation-for-Unsupervised-Text-Style-Transfer/blob/master/POS_distance.py

[10] https://radimrehurek.com/gensim/auto_examples/tutorials/run_wmd.html



General Information / Practice Task / Dashboard

Description:

We are scientists from [redacted]

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

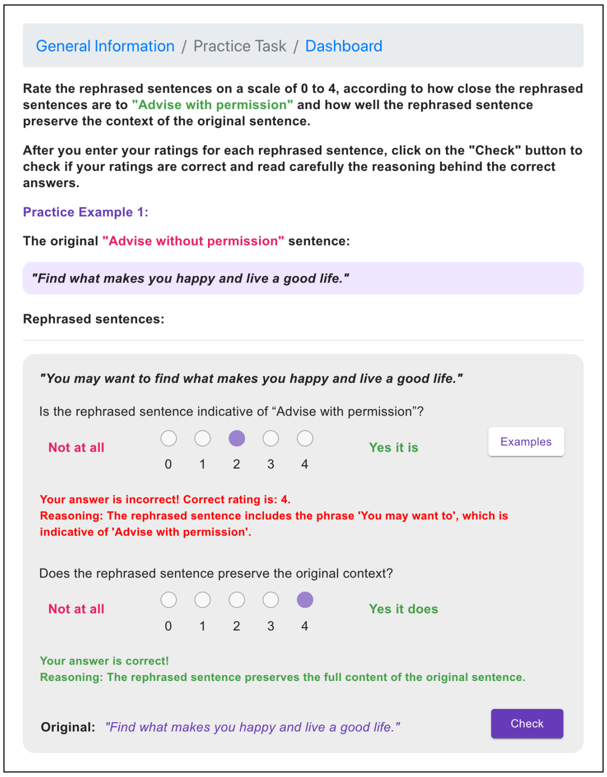
Figure 5: Human evaluation task description.

original sentence, it might still be able to generate rephrasings relevant to the preceding context.

It should 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.

G Distribution and Use of Artifacts

The artifacts produced, including the datasets and the models, will be released under the CC BY-NC-SA 3.0 license `https://creativecommons.org/licenses/by-nc-sa/3.0`, providing only non-commercial access to the users. We use artifacts such as the CounselChat dataset, and pre-trained language architectures such as BERT (Devlin et al., 2019), RoBERTA (Liu et al., 2019), Blender (Roller et al., 2021), and GPT3 (Brown et al., 2020) for research purposes only, which does not violate their intended use.



General Information / Practice Task / Dashboard

Rate the rephrased sentences on a scale of 0 to 4, according to how close the rephrased sentences are to "Advise with permission" and how well the rephrased sentence preserve the context of the original sentence.

After you enter your ratings for each rephrased sentence, click on the "Check" button to check if your ratings are correct and read carefully the reasoning behind the correct answers.

Practice Example 1:

The original "Advise without permission" sentence:

"Find what makes you happy and live a good life."

Rephrased sentences:

"You may want to find what makes you happy and live a good life."

Is the rephrased sentence indicative of "Advise with permission"?

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

Your answer is incorrect! Correct rating is: 4.
Reasoning: The rephrased sentence includes the phrase 'You may want to', which is indicative of 'Advise with permission'.

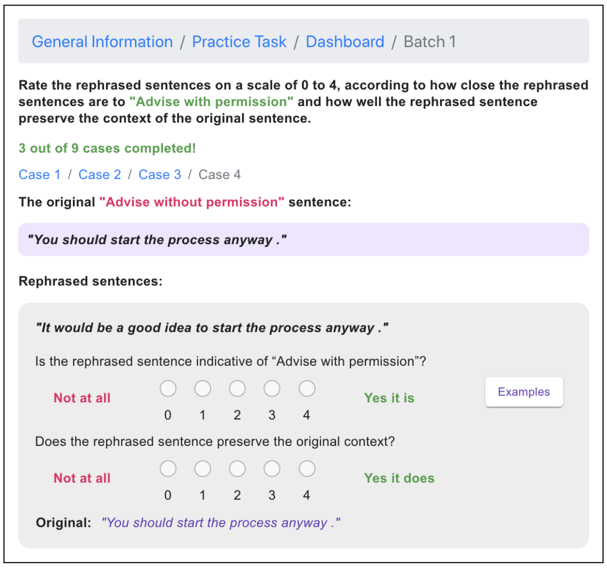
Does the rephrased sentence preserve the original context?

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

Your answer is correct!
Reasoning: The rephrased sentence preserves the full content of the original sentence.

Original: *"Find what makes you happy and live a good life."* Check

Figure 6: Self-evaluating practice task offered to the counselors to get familiarized with the rating task.



General Information / Practice Task / Dashboard / Batch 1

Rate the rephrased sentences on a scale of 0 to 4, according to how close the rephrased sentences are to "Advise with permission" and how well the rephrased sentence preserve the context of the original sentence.

3 out of 9 cases completed!

Case 1 / Case 2 / Case 3 / Case 4

The original "Advise without permission" sentence:

"You should start the process anyway ."

Rephrased sentences:

"It would be a good idea to start the process anyway ."

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: *"You should start the process anyway ."*

Figure 7: The human evaluation task interface.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 9: "Limitations"

☑ A2. Did you discuss any potential risks of your work?
Section 10: "Ethics Statement" under "Chatbots for Distress-Consolation"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
"Abstract" and Section 1: "Introduction"

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 "Datasets", Section 4 "MI Classifier", Section 5.1 "Pseudo-Parallel Corpora", and Section 5.2 "Models".

☑ B1. Did you cite the creators of artifacts you used?
Section 3 "Datasets", Section 4 "MI Classifier", and Section 5.2 "Models".

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix G: "Distribution and Use of Artifacts" and Section 10: Ethics Statement under "Chatbots for Distress-Consolation"

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Appendix G: "Distribution and Use of Artifacts" and Section 10: Ethics Statement under "Chatbots for Distress-Consolation"

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 10 "Ethics Statement" under "Data Curation".

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3 "Datasets", and Appendix A: "Datasets"

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 "Datasets", Section 4 "MI Classifier", Section 5.1 "Pseudo-Parallel Corpora", and Appendix A "Datasets".

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

Section 4: "MI Classifier" and Section 5.2: "Rephrasing Models"

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix B: "MI Classifier", and Appendix C.2: "Rephrasing Models"

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4: "MI Classifier", Section 5.2: "Rephrasing Models", Appendix B: "MI Classifier", and Appendix C.2: "Rephrasing Models"

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4: "MI Classifier", Section 6: "Automatic Evaluation", Appendix B: "MI Classifier", and Appendix D: "Automatic Evaluation"

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix D: "Automatic Evaluation"

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 7: "Human Evaluation"

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix D: "Human Evaluation"

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 7: "Human Evaluation" and Section 10: "Ethics Statement" under "Human Evaluation"

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. We recruited human workers only to rate rephrased responses generated by the rephrasing models that we developed. No personal information was collected during this experiment. But we discussed the details of our experiment and informed why we are conducting the experiment for the crowdworkers recruited. These details are denoted under Appendix E: "Human Evaluation".

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. We recruited human workers only to rate rephrased responses generated by the rephrasing models that we developed. No personal information was collected during this experiment.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. We recruited human workers only to rate rephrased responses generated by the rephrasing models that we developed. No personal information was collected during this experiment. But we did include the information that all workers recruited were professionally trained in the practice of counseling and all had professional competency in English.