What Makes a Conversation Satisfying and Engaging? — An Analysis on Reddit Distress Dialogues

Sena Necla Çetin
Supervisors: Anuradha Welivita, Dr. Pearl Pu Faltings
School of Computer and Communication Sciences
École polytechnique fédérale de Lausanne
Switzerland
{sena.cetin, kalpani.welivita, pearl.pu}@epfl.ch

***Abstract*—Recently, AI-driven chatbots have gained interest to help people deal with emotional distress and help them regulate emotion. However, since conversational data between patients who are in emotional distress and therapists who are actively offering emotional support is hardly available publicly due to privacy and ethical reasons, the most feasible option is to train chatbots on data from online forums such as Reddit. One challenge is ensuring that the data collected from these platforms contain responses that lead to high engagement and satisfaction and avoid those that lead to dissatisfaction and disengagement. We have developed a novel scoring function that can measure the level of satisfaction and engagement in distress oriented conversations. Using this scoring function, we classified dialogues in the Reddit Emotional Distress (RED) dataset as highly satisfying, less satisfying, highly engaging, and less engaging. By analysing these separate dialogues, we finally came up with a set of guidelines that describes which conversational strategies lead to highly satisfying and highly engaging conversations and which conversational strategies lead to less satisfying and less engaging conversations. Our guidelines can serve as a set of rules when developing therapeutic chatbots from online mental health community data so that inappropriate responses could be avoided and speaker satisfaction and engagement with these chatbots could be increased.**

I. Introduction

Nearly a billion people worldwide suffer from a mental or substance use disorder [1]. Only a low proportion of these people can get help due to the stigma around mental health as well as the unavailability of financial resources [2]. On the other hand, in countries with free mental healthcare, such as England, the waiting times to receive psychological therapy can reach up to 18 weeks [3]. Nevertheless, social support is critical for individuals with mental and/or emotional disorders to help them deal with their difficult situations. One recent solution is the social chatbot, a system capable of conversing and interacting with human users using natural language.

Recently, online mental health communities (OMHCs) have emerged to provide emotional support for people in distress [11]. The empathetic responses expressed in these OMHCs bring a positive shift to the poster's feelings [12]. Even though these OMHCs provide significant help in alleviating emotional distress, successful social support requires users to engage with each other and failures may lead to serious consequences [13]. Sharma et al. (2020) highlight the importance of early replies and mutual discourse for seeker retention in these OMHCs [13]. Sharma et al. (2018) reveal that the amount of support a post receives in these online mental health communities is positively correlated with the amount of linguistic accommodation the poster exhibits [8]. This signifies the importance of building a therapeutic chatbot, which could be of immediate help to seekers, regardless of their linguistic accommodation. Gennaro et al. (2020) examine the effectiveness of an empathetic chatbot in combatting the adverse effects of social exclusion on mood. They find that participants who interacted with an empathetic chatbot reported a higher mood compared to the control group.

Recent work shows that AI-driven chatbots can effectively generate appropriate syntactic and contextual responses. Since therapeutic data between patients who are in emotional distress and therapists who are actively offering emotional support is hardly available publicly due to privacy and ethical reasons, the most feasible option is to train chatbots on data from online forums such as Reddit and TalkLife. However, the responses in these platforms are not given by professionals. Thus, they may include conversation patterns that are not ideal to be used when addressing someone's emotional distress. One example is Microsoft's Tay bot that learned from racist comments in Twitter and responded inappropriately to users [16]. Especially, when it comes to mental health, which is way more sensitive, these inappropriate responses can lead to user's dissatisfaction and disengagement with the chatbot.

Both engagement and satisfaction are integral components of therapeutic conversations to provide successful emotional support to speakers' needs. However, they are two disconnected dimensions, such that the level of one measure may not necessarily imply the level of the other. Specifically, a highly satisfying dialogue might not necessarily be highly engaging and a highly engaging dialogue might not necessarily be highly satisfying. Examples of each case are shown in Tables I and II, respectively.

To the best of our knowledge, no one else has attempted to identify conversation strategies leading to satisfaction and engagement in distress oriented conversations and those that lead to dissatisfaction and disengagement. By identifying these conversation patterns, we can control chatbots from making

Speaker:	My step mom makes me freak out. For example, recently she found out I was gay and that one of my friends were transgender. Not only did she say that I was gay because I had gay friends and wanted to fit in, but she said my friend isn't transgender (...) Her and my dad also constantly make fun of my mom because she doesn't make as much money as them (...) Does anyone know any tips to help me stand up for my mom?
Listener:	Have you ever tried mindfulness? I know your struggles, and how hard it can be, but mindfulness can really help put you in a better frame of mind, and calm you down when you're worked up. Regarding when you're with your step-mum, there's always the obvious breathing exercises, but you could also try getting one of those fidget toys that can keep your mind distracted enough to not leave you too emotionally distressed. Best of luck to you!
Speaker:	Thanks for the help. I will try.
Listener:	Keep me updated if you can, I hope I've helped.

Table I. Example of a highly satisfying and less engaging dialogue

Speaker:	My boyfriend's mother asked my boyfriend to ask me for $250 for a couple times now (...) I have told my boyfriend I am tired of her asking me for money, but I feel bad for feeling this way. I am also worried she might not be able to pay me back as she isn't very good with money.
Listener:	Sounds like she doesn't want them to know, from her point of view she should be in a much better position than she is, and she feels weak and vulnerable. It's annoying but at least she will pay you back.
Speaker:	Yeah I can see your point. Problem is, she could be in a better position but she wastes her money on her useless daughter, who is a topic for a whole other post!
Listener:	Gotcha, what do you think you will do? You could always tell her that you don't have the money at the moment.
Speaker:	I have already given it to her. I feel bad saying no :/
Listener:	You sound like me :/ I am a bit of a pushover.

Table II. Example of a highly engaging and less satisfying dialogue

inappropriate responses and drive them to achieve specific conversational goals such as speaker satisfaction and engagement. To fill this gap, following Yeh et al. (2020)'s work [10], we analyze the Reddit Emotional Distress (RED) dataset, which consists of dialogues between speakers and listeners, in which the speakers convey their ongoing issues and the listeners from the community offer emotional support to them in various ways. The dataset contains the the predicted emotion and sentiment of each dialogue turn. We identify the techniques that the listeners use when providing support to those in distress. These techniques can then be utilized in designing and developing automatic chatbots that can coherently and consistently provide therapeutic support for people in distress.

In this work, we develop a novel scoring function that can measure the level of speaker satisfaction and engagement in distress oriented conversations. Using this scoring function, we classify the dialogues in the RED dataset as highly satisfying, less satisfying, highly engaging, and less engaging. By manually analysing the occurrence frequencies of listener intents on a subset of these dialogues, we finally come up with a set of guidelines that describes, which conversational strategies lead to highly satisfying and highly engaging conversations and which conversational strategies lead to less satisfying and less engaging conversations. Our guidelines can serve as a set of rules when developing therapeutic chatbots from OMHC data so that inappropriate responses could be avoided and speaker satisfaction and engagement with these chatbots could be increased.

In the following, we first describe the previously conducted research to identify empathy generation in therapeutic conversations. We follow this by introducing the dataset and explaining the preprocessing steps (see Figure 1). Then, we explain the methods we introduced to measure engagement and satisfaction in emotional support dialogues. Thereafter, we explain how we measured the performance of these methods. Finally, we suggest a set of guidelines for counselors, in light of the findings from our application of these methods to classify the conversations in the curated RED dataset.

II. Related work

Previous research has introduced various computational methods to identify and control empathy generation in therapeutic conversations.

Sharma et al. (2020) develop the EPITOME framework for characterizing empathy in text-based conversations [7]. This framework consists of three communication mechanisms, i.e., *Emotional Reactions, Interpretations,* and *Explorations*, along with their level of strength from 0 (none) to 2 (strong). In addition, they create a new corpus of 10k conversations and a computational method for identifying empathy in text-based conversations. The framework reveals that speakers have a better impression of high empathy interactions. Namely, strong *Emotional Reactions, Interpretations,* and *Explorations*.

Zhang and Danescu-Niculescu-Mizil (2020) quantify the forwards- and backwards- orientations of listener utterances. A forwards-oriented utterance prompts the speaker towards problem-solving, whereas a backwards-oriented utterance reflects or affirms what the speaker has said to check understanding and show empathy. They find that speakers favor relatively backwards-oriented interactions, where the listener is inclined towards addressing the situation, and that keeping a balance between these orientations is of utmost importance for speaker satisfaction and engagement.

Sharma et al. (2018) find that certain online mental health communities direct more emotional support while others provide more informational support [8]. Their results validate Cutrona and Russell's (1990) *Optimal Matching Theory* such that people with a certain type of disorder (e.g., Mood Disorders) may need more emotional support than those with another type of disorder (e.g., Compulsive Disorders) [9].

Pfeil and Zaphiris (2007) investigate the patterns of empathy in online communication and compare them to the patterns of offline communication [15]. Compared to offline studies, the participants seem to have fewer inhibitions to give away their feelings, which leads to less need for prompting to get more information. Moreover, understanding is not as prevalent in online conversations. Deep support can be seen as an implicit proof of understanding.

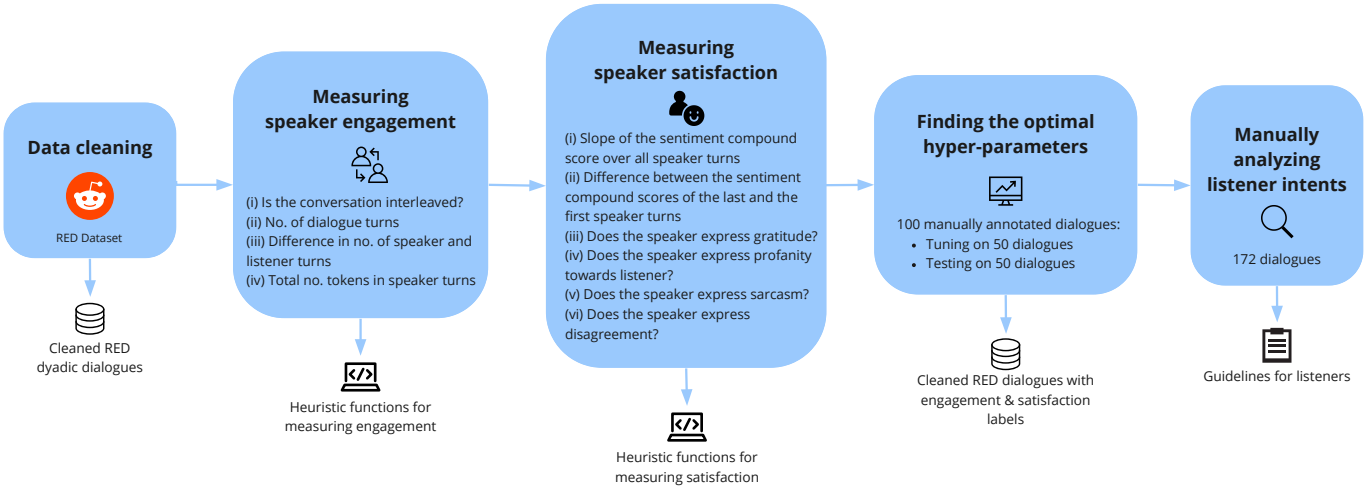

Figure 1: Development steps for classifying speaker engagement and satisfaction on the RED dataset, and for constructing the listener guidelines

Welivita and Pu (2020) define a taxonomy of listener specific empathetic response intents capable of supporting automatic empathetic communication in conversations [5]. They reveal that certain listener intents are associated with positive speaker emotions while others are associated with negative speaker emotions. More specifically, *encouraging* and *wishing* intents are associated with positive speaker emotions while *sympathizing* and *consoling* are associated with negative speaker emotions. The taxonomy is suggested to be incorporated into the design of a social chatbot to gain more controllability and interpretability of the generated responses.

A majority of the prior work studying engagement between users in OMHCs have derived their findings using correlations between user and platform characteristics [30], [31], [32]. However, they have not considered the conversational aspects of engagement, but merely examined the dimensions such as the number of posts and likes. One exception is Sharma et al. (2020), which propose four indicators of user engagement based on attention and interaction in two popular OMHCs, TalkLife and Reddit. As attention-based indicators, they use the number of dialogue turns and number of listeners in the conversation. As interaction-based indicators, they use time between responses and degree of interaction, i.e., whether the speaker responds back to the listener and whether the listener responds back to the speaker.

Previous work has measured speaker satisfaction in social chatbots, which aim at improving the emotional state of its users, using self-reported measures. For example, Vaidyam et al. (2019) investigated 10 social chatbots, which are primarily built towards the care of those with high risk of mental disorders. To measure user satisfaction, they used self-reports of the participants on ease of use, desire to continue using the system, liking, and trust. The participants rated the chatbots on all scales as highly satisfying (>4.2 out of 5) and reported their helpfulness, ease to use, and informativeness. The study shows the potential for conversational agents in psychiatric use, but lacks the computational methods to be applied on a large-scale dialogue dataset for measuring satisfaction. Gennaro et al. (2020) studied the impact of an empathetic chatbot on participants who experience social exclusion. Their results reveal that interacting with an empathetic chatbot (e.g., "I'm sorry that this happened to you") increased the mood of the participants compared to those in the control condition who interacted with a chatbot that merely acknowledged their responses (e.g., "Thank you for your feedback"). However, they measured speaker satisfaction using self-report, which again is inapplicable in the labeling of a large-scale dialogue dataset. In addition, existing studies have not focused well enough in negative response strategies that can make the user feel dissatisfied and disengaged with the conversation.

To address the above limitations stated with respect to the existing studies, we develop a scoring function that measures the level of speaker satisfaction and engagement in distress oriented conversations and discover conversational strategies that can make a conversation highly satisfying and engaging and also those that lead to dissatisfaction and disengagement by applying this function on a large-scale Reddit distress dialogues dataset.

III. Dataset

Over the years, researchers have constructed numerous dialogue datasets to be utilized in the design and development of automatic chatbots that can appropriately provide therapeutic support to its users. A number of these datasets utilize audio, such as the TED-LIUM corpus [18], to recognize speech and sentiment in real-time. In addition to audio, some datasets also contain visual signals for facial expression detection, such as the IEMOCAP [19], SEMAINE [20], and MELD [21] datasets. However, the text data in these datasets may not fully represent contextual intents due to the existence of other channels of information. Similarly, datasets that utilize TV or movie transcripts (e.g., EmotionLines) [17], (e.g., OpenSubtitles)

conversation_id	post_title	author	dialog_turn	text	compound	sentiment	emotion_prediction
865	MentalHealthSupport	Advice on preventing overthinking	1	What do you do to stop a crazy train of thoughts from spiraling out of control?	-0.5574	negative	sentimental
865	MentalHealthSupport	Advice on preventing overthinking	2	This might be completely useless to you, but when i spiral out of control with my overthinking and get paranoid i tell myself that I am overthinking and being paranoid. I find it difficult but when ever i overthink i just tell myself im being ridiculous, because I know I am and i tryst myself to tell myself the truth	-0.8624	negative	sentimental
865	MentalHealthSupport	Advice on preventing overthinking	3	I'm trying to do this too and I often reassure myself that I'm being completely ridiculous!	-0.1742	negative	angry

Table III. Example conversation taken from the RED dataset after data cleaning

[22], and telephone recordings (e.g., Switchboard) [23] fail to fully model the interactions occurring via only text. Even purely text-based dialogue datasets (e.g., DailyDialog) [24] do not guarantee to contain empathetic responses.

Rashkin et al. (2019) develop the EmpatheticDialogues dataset, which contains 24,856 human-human conversations with an average of almost four dialogue turns per conversation. Almost all conversations in this dataset are empathetic, purely-text based, and contain no toxicity. Extending their work, Welivita and Pu (2020) develop a taxonomy of empathetic listener intents using the EmpatheticDialogues dataset. They automatically annotate the dataset based on the most frequent intents in their taxonomy and the 32 types of emotion categories in EmpatheticDialogues. Their results can be used to gain more controllability in the generated responses of an empathetic chatbot. However, the limited size of the dataset does not allow the training of a robust neural chatbot. Moreover, there does not exist a dataset consisting of conversations between speakers in emotional distress and listeners who offer emotional support to them in the literature. Therefore, Yeh et al. (2020) curate the RED dialogues dataset and analyzes it at the lexical, sentiment, and emotional level, to be potentially used in the training of a neural chatbot [10]. Since this dataset is large-scale, preprocessed to contain almost no listener profanity, contains sentiment analysis and emotion prediction, we chose it to apply our computational methods to identify user engagement and satisfaction.

The RED dataset consists of two million conversations, curated from 8 different emotional support subreddits. It contains 1.3 million dyadic dialogues (Table IV) and 0.6 million multiparty dialogues. The number of dialogue turns inside conversations follow a power law distribution, where most conversations end in two turns and the average number of dialogue turns is four. The raw text is preprocessed to remove the HTML tags, URLs, and to replace the numerals with a <NUM> tag. Profanity from listener responses is removed using PROFANITY-CHECK [26], a fast and robust library to detect offensive language. It uses the Support Vector Machine classifier [27] trained on a human-labeled 200k samples of clean and profane text. It returns the probability of predicting profanity in a given text along with its prediction. Yeh et al. (2020) set the threshold to 0.95 upon thorough manual inspection to allow the responses of the listeners who tend to aggressively express themselves with no mean intention. For sentiment analysis, they use VADER [28], a lexicon and rule-based tool specifically attuned for social media data. For emotion and intent analysis, they use the EmoBERT classifier [5] trained on the EmpatheticDialogues dataset. EmoBERT is a BERT [29] based emotion classifier that predicts the emotion or intent of a particular dialogue turn with an accuracy of 65.88%.

Subreddit	No of Dialogs	No. of Turns	Avg. No. of Turns per Dialog
Entire	1,275,486	3,396,476	2.66
r/depression	510,035	1,396,044	2.74
r/depressed	10,892	23,804	2.19
r/offmychest	437,737	1,064,467	2.43
r/sad	18,827	42,293	2.25
r/SuicideWatch	262,469	791,737	3.02
r/depression_help	23,678	51,849	2.19
r/Anxietyhelp	8,297	18,351	2.21
r/MentalHealthSupport	3,551	7,931	2.23

Table IV. Descriptive statistics of dyadic conversations in the entire dataset as well as in each subreddit (before cleaning)

IV. METHODOLOGY

As shown in Figure 1, our classification and guideline formation pipeline consists of 5 main stages: 1) data cleaning; 2) measuring speaker engagement; 3) measuring speaker satisfaction; 4) finding the optimal hyper-parameters; 5) manually analyzing listener intents. The methodology is described in detail in the following.

A. Data cleaning

Due to the dyadic nature of a chatbot, we selected only the dyadic dialogues from the RED dataset to apply our methods. To prepare our data for analysis, we first applied cleaning. Noticing the existence of some duplicate turns within conversations, some of which occurred hundreds of times either due to a bug or as a spam, we kept only the first of the duplicate texts within the same conversation. To be able to infer speaker satisfaction from the later turns, we removed conversations with less than 3 dialogue turns. Moreover, we noticed the existence of some multiparty conversations inside the dyadic conversations and removed them. Table III shows an example conversation taken from the RED dataset after data cleaning. Table V shows the descriptive statistics of the dataset after the cleaning process.

Subreddit	No of Dialogs	No. of Turns	Avg. No. of Turns per Dialog
Entire	180,205	780,560	4.33
r/depression	77,548	333,509	4.30
r/depressed	1,006	4,039	4.01
r/offmychest	60,986	239,056	3.92
r/sad	2,407	9,452	3.93
r/SuicideWatch	35,023	181,551	5.18
r/depression_help	1,961	8,052	4.11
r/Anxietyhelp	897	3,400	3.79
r/MentalHealthSupport	377	1,501	3.98

Table V. Descriptive statistics of dyadic conversations in the entire dataset as well as in each subreddit (after cleaning)

B. Measuring speaker engagement

To be able to measure speaker engagement, we applied various heuristic methods. First, we merged the consecutive speaker turns into a single speaker turn and checked if the conversation is interleaved, such that a speaker turn is always followed by a listener turn and a listener turn is always followed by a speaker turn. We based this method on models in communication theory [**?**], [**?**], where interactions are separated into two categories. In the first, the speaker responds back to the listener, whereas in the second, the speaker never responds back to the listener. As the second predictor of engagement, similarly to Sharma et al. (2020), we selected the number of dialogue turns. Since equal number of speaker and listener turns is the most desirable condition, referred to as *Mutual Discourse* in Sharma et al. (2020), we selected the absolute value of the difference in number of speaker and listener turns as the third predictor. Finally, we selected the total number of tokens in speaker turns as our last predictor. However, to limit the impact of very long texts on speaker engagement, we placed an upper limit of 30 tokens, such that any dialogue turn of 30 or more tokens are counted as having 30.

Overall, we used 4 different predictors for predicting speaker engagement: (i) whether the conversation is interleaved (+), (ii) the number of dialogue turns (+), (iii) the difference in number of speaker and listener turns (-), and (iv) the total number of tokens in speaker turns (+). Note that the positive or negative impact of each predictor to speaker engagement is indicated with a "(+)" or "(-)", respectively. To come up with a numerical engagement score, we assigned weights to each of the predictors. Moreover, we defined a numerical threshold for the engagement score.

C. Measuring speaker satisfaction

To be able to measure speaker satisfaction, we applied various heuristic methods. Yeh et al. (2020) apply sentiment analysis on the RED dataset on the dialogue turn level using the VADER [28] tool, as explained in Section III. Since a single dialogue turn can contain numerous sentences, we extended their work, and applied sentiment analysis using the VADER tool on the sentence level. Within the sentence-level sentiments, we assigned the sentiment with the strongest magnitude as the dialogue turn sentiment. As the first predictor of satisfaction, we used the slope of sentiment throughout the conversation. This allowed us to capture the overall direction of the speaker's change in mood. Next, we calculated the change in sentiment from the last to the first speaker turn and used it as the second predictor of satisfaction. This allowed us to capture a more fine-grained change in sentiment.

As the third predictor of satisfaction, we aimed to predict satisfaction by detecting expressions of gratitude by two complementary methods. As the first method, we checked if the last speaker turn was tagged with the *grateful* emotion and *positive* sentiment. However, since the EmoBERT classifier is applied to the RED dataset on the dialogue turn level and has a classification accuracy of 66%, it may not always return the *gratitude* tag expressed in one of the sentences in a dialogue turn. Thus, as the second method, we converted all speaker responses to lowercase and used the *matcher* module of the SpaCy library [33] to match any tokens (e.g., "thank") and phrases (e.g., "your help") that convey gratitude in all the speaker responses except the first one.

Next, we checked if the speaker expresses profanity toward the listener. For this, we used the PROFANITY-CHECK library on all the speaker turns except the first one, and checked if the prediction is equal to 1. To differentiate between profanity toward else and toward listener, we utilized the *matcher* module of SpaCy again, and checked if the speaker response that contains profanity also contains the tokens "you" and/or "your".

Then, we checked if the speaker expresses sarcasm in any of the dialogue turns (except the first one), using a Tensorflow [34] text classification model [35] trained on the News Headlines Dataset for Sarcasm Detection [36], [37], which returns the probability of sarcasm in the input text. Upon careful inspection and trials on speaker response samples from the RED dataset, we selected the threshold for sarcasm prediction as 0.6.

Finally, we checked if the speaker expresses disagreement in any of their responses except the first one. Similarly to the other phrase detection steps we have used, we used SpaCy, and detected the existence of certain tokens and phrases that convey disagreement, e.g., "i don't think so", "disagree", "no

way", which we obtained from an online English language teaching source [38].

Overall, we used 6 different predictors for predicting speaker satisfaction: (i) the slope of the sentiment from the first speaker turn to the last speaker turn (+), (ii) the change in the sentiment compound score between the last and the first speaker turn (+), (iii) whether the speaker expresses gratitude (+), (iv) whether the speaker uses profanity towards the listener (-), (v) whether the speaker uses sarcasm (-), and (vi) whether the speaker expresses disagreement (-). Note that the positive or negative impact of each predictor to speaker satisfaction is indicated with a "(+)" or "(-)", respectively. To come up with a numerical satisfaction score, we assigned weights to each of the predictors. Moreover, we defined a numerical threshold for the satisfaction score.

D. Finding the optimal hyper-parameters

To tune the weights of each of the engagement and satisfaction predictors as well as the engagement and satisfaction thresholds, we first annotated 100 dialogues with 12-13 samples from each of the 8 subreddits with their ground truth labels of engagement (less engaging: 0, highly engaging: 1) and satisfaction (less satisfying: 0, highly satisfying: 1). Then, we separated these 100 samples into validation and test sets with a 1:1 ratio and applied grid search on the validation set. However, due to the large number of hyper-parameters and the limited computational resources, we could only select a small range of hyper-parameters (Table VI), with 2-3 values for each of the hyper-parameters. We selected the optimal set of hyper-parameters with respect to the best f1-score on the validation set and applied them onto the test set to test their performance on unseen samples. Finally, we applied the optimal hyper-parameters onto the entire dataset to predict the level of engagement and satisfaction of all dialogues.

Hyper-parameter	Searched Values
engagement_threshold	[2.75, 3, 3.25]
num_turns_weight	[0.75, 1, 1.25]
interleaved_weight	[0.75, 1, 1.25]
token_length_weight	[0.025, 0.05, 0.075]
num_turn_difference_weight	[-0.75, -0.5, -0.25]
satisfaction_threshold	[0.4, 0.5, 0.6]
slope_weight	[0.4, 0.5]
sentiment_change_weight	[0.4, 0.5]
grateful_bonus_weight	[2.75, 3, 3.25]
profanity_penalty_weight	[0.4, 0.5]
sarcasm_penalty_weight	[0.4, 0.5]
disagreement_penalty_weight	[0.4, 0.5]

Table VI. Grid search hyper-parameters: the prediction thresholds of engagement and satisfaction scores, and the weights of predictors, along with their searched ranges.

E. Manually analyzing listener intents

Following the classification on the entire dataset, we selected a subset of 172 dialogues to manually analyze the listener intents that lead to high satisfaction or less satisfaction, and high engagement or less engagement. We labeled the dialogues with the taxonomy of empathetic response intents in Welivita and Pu (2020). In order to account for the unempathetic listener intents as well, we added *judging, joking,* and *expressing negative thoughts* to the set of listener intents. All listener intents used in the analysis can be found in Table VIII along with an example for each.

V. Results

The results and performances of the engagement and satisfaction scoring functions are given in Table VII. Engagement prediction has better performance compared to satisfaction prediction on both the validation and test sets. Also, engagement hyper-parameter tuning requires significantly less time than satisfaction hyper-parameter tuning even though the number of tested hyper-parameters are in the same order of magnitude, i.e., 243 and 288, respectively. This difference is due to the use of more sophisticated and computationally-intensive methods (e.g., phrase matching, slope of sentiment scores, profanity-checking) required in measuring satisfaction as compared to measuring engagement.

Hyper-parameter	Optimal Value	Tuning Time (s)	F1-score (val)	F1-score (test)
engagement_threshold	2.75			
num_turns_weight	0.75			
interleaved_weight	0.75	93.34	0.96	0.95
token_length_weight	0.025			
num_turn_difference_weight	-0.25			
satisfaction_threshold	0.6			
slope_weight	0.5			
sentiment_change_weight	0.5			
grateful_bonus_weight	3.25	17382.60	0.81	0.78
profanity_penalty_weight	0.5			
sarcasm_penalty_weight	0.5			
disagreement_penalty_weight	0.5			

Table VII. Grid search results for engagement and satisfaction prediction: the optimal hyper-parameter values, the elapsed time (in seconds) for tuning, the f1-scores using the optimal hyper-parameters on the validation set, and the f1-scores using the optimal hyper-parameters on the test set are shown.

Figure 2 shows the listener intent frequencies in highly satisfying, less satisfying, highly engaging, and less engaging conversations. Sharing or relating to own experience (or *self-disclosure*) is the most frequent listener intent regardless of the class. However, it occurs more in highly satisfying conversations by a significant margin of 12.97% compared to less satisfying conversations. Moreover, listeners express more disapproving (+10.03%), less acknowledging (-6.51%), less encouraging (-13.39%), and more judging (+6.82%) intents in less satisfying conversations. They also express less care or concern (-10.12%) and more negative thoughts (+12.36%). On the other hand, listeners express less questioning (-12.65%), less disapproving (-9.20%), less advising (-21.86%), and more acknowledging (+14.58%) intents in less engaging conversations compared to highly engaging conversations.

VI. Discussion

To measure speaker engagement and satisfaction in OMHC datasets, we applied data cleaning, several heuristic methods,

Category	Examples
1. Sharing or relating to own experience	I've been feeling the same for about a month now.
2. Advising	Call her and tell her that she didn't do anything wrong and you didn't mean to react like that.
3. Questioning (to know further details or clarify)	How recent was their passing? Were you close?
4. Suggesting	Perhaps you should go over this stuff with your boyfriend?
5. Expressing care or concern	I just noticed this post is 3 days old, please let me know how you're doing.
6. Encouraging	I promise you'll get through this.
7. Acknowledging (Admitting as being fact)	It sounds like you're in a lot of physical and mental pain.
8. Sharing own thoughts or opinion	Therapy is awesome because it's focused just on you.
9. Sympathizing (Expressing feeling pity or sorrow for the person in trouble)	Dude, I'm sorry for your situation, I truly am.
10. Wishing	Well done!
11. Consoling	I hope you see the good that's in you.
12. Disapproving	I'm sure you don't look disgusting!
13. Agreeing (Thinking/Saying the same)	You are most definitely not wrong.
14. Appreciating	I'm proud of you for what you're doing, you're a good guy.
15. Expressing negative thoughts	I constantly live in my blurry head with my muddled thoughts. It makes it impossible to be good at my job, form good friendships or enjoy a girls company. I just want to die.
16. Expressing relief	Phew.. That's a relief. I am glad you were okay.
17. Joking	Wait IE worked but it had to be "fucking internet explorer"?
18. Judging	Vegetarianism doesn't make you superior to people. A defining life philosophy, yes, can give you drive and motivation to achieve things beyond yourself. But don't think you're superior to your peers just because you happen to be a vegetarian.

Table VIII. 18 listener intents and their examples. Note that 15 of these intents are taken from Welivita and Pu (2020) and 3 are newly introduced.

and hyper-parameter tuning. Applying our novel engagement and satisfaction scoring functions on the RED dataset, we classified each conversation as less engaging, highly engaging, less satisfying, and highly satisfying.

Looking at the test set performances, our novel speaker engagement and satisfaction scoring functions are promising to classify conversations in OMHCs for selecting highly engaging and highly satisfying conversations to be used in the training of an empathetic chatbot. The close performances of the engagement and satisfaction measures on the validation and test sets indicate that the models do not overfit to the validation set.

Analyzing the listener intents, we notice the difference in the occurence frequencies between each of the classes (Figure 2). Among these communication strategies, some are more intuitive such as the correlation between expressing care or concern and higher satisfaction. However, some findings are less intuitive such as the correlation between disapproving and higher user engagement. Drawing from our results, we come up with the following set of guidelines to be used in the development of therapeutic chatbots from OMHC data:

- Sharing or relating to one's own experience, expressing care or concern, acknowledging, and encouraging intents can lead to higher speaker satisfaction.
- Questioning, sharing own thoughts or opinion, disapproving, joking, judging, and expressing negative thoughts intents can lead to lower speaker satisfaction.
- Sharing or relating to one's own experience, advising, questioning, sharing own thoughts or opinion, sympathizing, consoling, disapproving, agreeing, appreciating, joking, and judging intents can lead to higher speaker engagement.
- Expressing care or concern, encouraging, acknowledging, wishing, and expressing negative thoughts intents can lead to lower speaker engagement.

VII. Conclusion

In this work, we introduced novel speaker engagement and satisfaction scoring functions to be used in the development of therapeutic chatbots from OMHC data so that speaker satisfaction and engagement can be increased and inappropriate responses can be avoided. For measuring engagement, we used predictors such as, whether the conversation is interleaved, the number of dialogue turns, the difference in number of speaker and listener turns, and the total number of tokens in speaker turns. For measuring satisfaction, we used predictors such as, the slope of the sentiment score, the difference in sentiment from the last to the first speaker turn, and used various tools to detect if the speaker expresses gratitude, profanity toward the listener, sarcasm, or disagreement. Using a weighted sum of these predictors, we came up with the engagement and satisfaction scores of each conversation and classified them according to their respective thresholds. After hyper-parameter tuning, our engagement and satisfaction scoring functions performed with f1-scores of 0.95 and 0.78, on the test sets, respectively.

After classifying conversations on the RED dataset using a subset of 172 conversations, we manually analyzed the frequency of listener intents that lead to high satisfaction, high engagement, low satisfaction, and low engagement. Using our analysis, we came up with a set of guidelines that can serve as a set of rules for developing therapeutic chatbots from OMHC data. Based on our analysis, we observed that to achieve high

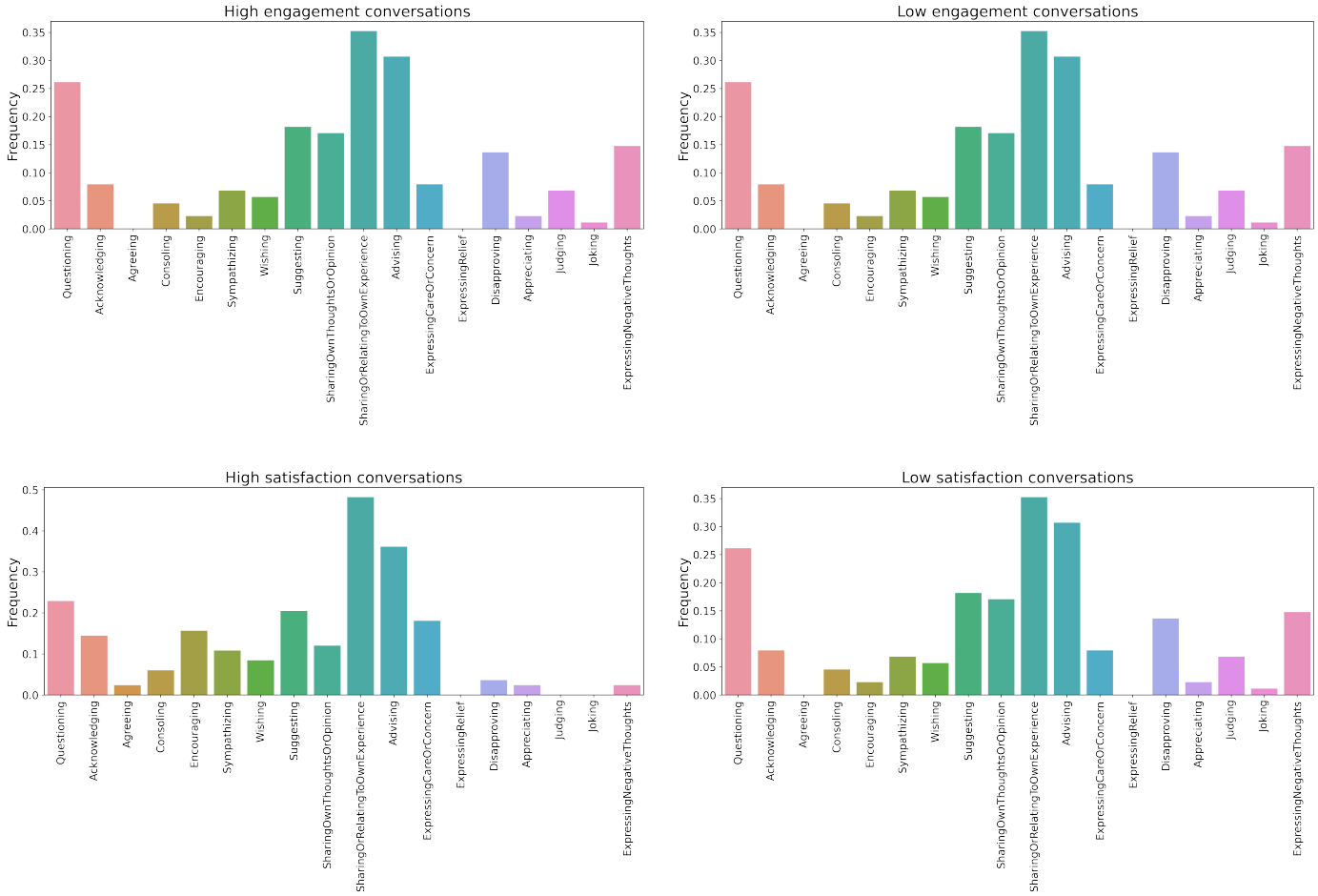

Figure 2: Occurrence frequencies of listener intents on the manually annotated 172 listener utterances in the RED dataset. Note that one of the 15 empathetic intents taken from Welivita and Pu (2020), i.e., *expressing relief*, does not occur in any of the 172 utterances.

speaker satisfaction, the chatbot should generate responses which are acknowledging and encouraging, share or relate to one's own experience, and express care or concern. It should not express negative thoughts, or be disapproving or judging. On the other hand, for high speaker engagement, the chatbot should be asking more questions and offer advice.

VIII. Future Work

There are some limitations to this work. Due to the limited computational resources and the time constraints, only a small grid search with 2-3 values per hyper-parameter have been applied. A larger grid search to find the optimal engagement and satisfaction hyper-parameters could yield more accurate classifications on the entire RED dataset.

Another limitation is the application of the EmoBERT classifier onto the dataset on the dialogue turn level. Since a dialogue turn can include multiple sentences with several different emotions, applying the EmoBERT classifier could improve the emotion prediction of each dialogue. This would result in a better estimation of speaker satisfaction.

Another limitation is the limited number of low engagement conversations that we used to draw the listener intent occurrence frequencies from. During data cleaning, we inherently removed conversations that are less engaging by removing all conversations with less than three turns to be able to calculate satisfaction in conversations. A separate dataset which contains conversations of length two could be created for identifying conversational strategies that lead to less speaker engagement in a more accurate manner.

As future work, in order to make better use of the large size of the RED dataset, multiparty conversations could also be analyzed and adapted to be used in the training of a more robust therapeutic chatbot.

IX. Acknowledgement

I would like to thank Anuradha Welivita for guiding me with her expertise and offering me help throughout the semester, and Dr. Pearl Pu for enabling me to work on this project, in the intersection between data science, natural language processing, and psychology.

REFERENCES

[1] Ritchie, H., & Roser, M. (2018) - "Mental Health". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/mental-health' [Online Resource]

[2] Olfson, M., Mojtabai, R., Sampson, N. A., Hwang, I., Druss, B., Wang, P. S., Wells, K. B., Pincus, H. A., & Kessler, R. C. (2009). Dropout from outpatient mental health care in the United States. Psychiatric services (Washington, D.C.), 60(7), 898–907. https://doi.org/10.1176/appi.ps.60.7.898

[3] Community and Mental Health Team (2020, July 30). Psychological Therapies - Annual Report on the Use of IAPT Services, England 2019-20 (United Kingdom, NHS Digital). Retrieved June 7, 2021, from https://files.digital.nhs.uk/B8/F973E1/psych-ther-2019-20-ann-rep.pdf

[4] Andrade, L. H., Alonso, J., Mneimneh, Z., Wells, J. E., Al-Hamzawi, A., Borges, G., Bromet, E., Bruffaerts, R., de Girolamo, G., de Graaf, R., Florescu, S., Gureje, O., Hinkov, H. R., Hu, C., Huang, Y., Hwang, I., Jin, R., Karam, E. G., Kovess-Masfety, V., Levinson, & D., Kessler (2014). Barriers to mental health treatment: results from the WHO World Mental Health surveys. Psychological medicine, 44(6), 1303–1317. https://doi.org/10.1017/S0033291713001943

[5] Welivita, A., & Pu, P. (2020). A Taxonomy of Empathetic Response Intents in Human Social Conversations. Proceedings of the 28th International Conference on Computational Linguistics. doi:10.18653/v1/2020.coling-main.429

[6] Zhang, J., & Danescu-Niculescu-Mizil, C. (2020). Balancing Objectives in Counseling Conversations: Advancing Forwards or Looking Backwards. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.470

[7] Sharma, A., Miner, A., Atkins, D., & Althoff, T. (2020). A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). doi:10.18653/v1/2020.emnlp-main.425

[8] Sharma, E., & Choudhury, M. D. (2018). Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. doi:10.1145/3173574.3174215

[9] Cutrona, C. E., & Russell, D. W. (1990). Type of social support and specific stress: Toward a theory of optimal matching. Social support: An interactional view (1990), 319–366.

[10] Chun-Hung Yeh, Anuradha Welivita, Pearl Pu Faltings. 2020. A Dialogue Dataset Containing Emotional Support for People in Distress.

[11] Whittaker, S., Isaacs, E., & O'Day, V. (1997). Widening the net: workshop report on the theory and practice of physical and network communities. SIGCHI BULLETIN 29 (1997), 27–30.

[12] Khanpour, H., Caragea, C., & Biyani, P. (2017). Identifying Empathetic Messages in Online Health Communities. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers) (pp. 246–251). Asian Federation of Natural Language Processing.

[13] Sharma, A., Choudhury, M., Althoff, T., & Sharma, A. (2020). Engagement Patterns of Peer-to-Peer Interactions on Mental Health Platforms.

[14] Gennaro, M., Krumhuber, E., & Lucas, G. (2020). Effectiveness of an Empathic Chatbot in Combating Adverse Effects of Social Exclusion on Mood. Frontiers in Psychology, 10, 3061.

[15] Pfeil, U., & Zaphiris, P. (2007). Patterns of Empathy in Online Communication. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 919–928). Association for Computing Machinery.

[16] Hunt, E. (2016, March 24). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. The Guardian. https://www.theguardian.com

[17] Chen, S. Y., Hsu, C. C., Kuo, C. C., & Ku L. W. (2018). Emotionlines: An emotion corpus of multi-party conversations. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation, Miyazaki, Japan.

[18] A. Rousseau, P. Del´eglise, and Y. Est'eve, "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks", in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 2014.

[19] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4):335.

[20] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE transactions on affective computing, 3(1):5–17.

[21] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 527–536, Florence, Italy.

[22] Pierre Lison, J¨org Tiedemann, and Milen Kouylekov. 2019. Open subtitles 2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In Eleventh International Conference on Language Resources and Evaluation (LREC), European Language Resources Association (ELRA).

[23] Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational linguistics 26(3):339–373.

[24] Yanran Li, Hui Su, Xiaoyu Shen,Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing, pages 986–995, Taipei, Taiwan.

[25] Hannah Rashkin, Eric Michael Smith, Margaret Li and Y-Lan Boureau. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5370–5381, Florence, Italy.

[26] Victor Zhou, Domitrios Mistriotis and Vadim Shestopalov. Profanitycheck, 2018, Github repository, https://github.com/vzhou842/profanitycheck.git

[27] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.

[28] Hutto, C.J. & Gilbert, E.E. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

[29] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pretraining of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[30] Andalibi, N.; Haimson, O. L.; Choudhury, M. D.; and Forte, A. 2018. Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media. ACM TOCHI 25(5):28.

[31] Andalibi, N., and Forte, A. 2018. Responding to sensitive disclosures on social media: A decision-making framework. ACM TOCHI 25(6):31.

[32] Ernala, S. K.; Labetoulle, T.; Bane, F.; Birnbaum, M. L.; Rizvi, A. F.; Kane, J. M.; and De Choudhury, M. 2018. Characterizing audience engagement and assessing its impact on social media disclosures of mental illnesses. In ICWSM'18.

[33] Hannibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). SpaCy [Program documentation]. Retrieved from https://doi.org/10.5281/zenodo.1212303

[34] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[35] Tensorflow - Course 3 - Week 2 - Lesson 2.ipynb. (n.d.). Retrieved from https://colab.research.google.com/github/lmoroney/dlaicourse/blob/master/TensorFlowInPractice/Course3-NLP/Course3-Week2-Lesson2.ipynb#scrollTo=0eJSTTYnkJQd

[36] Misra, R., Arora, P. (2019). Sarcasm Detection using Hybrid Neural Network. arXiv preprint arXiv:1908.07414.

[37] Misra, R., Grover, J. (2021). Sculpting Data for ML: The first act of Machine Learning.

[38] KSE Academy. (2019, November 25). How to Express Agreement and Disagreement. Retrieved from https://kseacademy.com/how-express-agreement-disagreement/