A Dialogue Dataset Containing Emotional Support for People in Distress

Chun-Hung Yeh
Supervisor: Kalpani Anuradha Welivita, Dr. Pearl Pu Faltings
Human Computer Interaction Group, IC, EPFL

***Abstract*—Many people suffer from emotional distress due to many reasons, such as significant life change, financial crisis, or various physical and mental health conditions. Inability to regulate emotion can potentially lead to self-destructive behavior such as substance abuse, self-harm, or suicide. Even though many helplines and therapeutic consultations are available to assist such people in distress, most people do not reach for help due to the public and personal stigma associated with mental health. Even therapeutic consultations are limited and are not available 24/7 to support people going through a traumatic episode. Therefore, it is important to assess the ability of AI-driven chatbots to help people deal with emotional distress. One of the significant limitations in developing such a chatbot is the lack of a large, curated dialogue dataset containing emotional support. This work curates a dataset containing two million emotional dialogues from Reddit and analyzes it at the lexical, sentiment, and emotional level. In this paper, we outline the data selection and curation process and the important observations made related to sentiment and emotion in speaker and listener turns of the resulting dataset. This dataset is expected to help build future chatbots that could offer emotional support for people in distress.**

***Index Terms*—Emotional support, Web crawling, Natural language processing**

I. Introduction

According to World Health Organization research, it is estimated that mental distress affect 29% of people in their lifetime [1]. Globally, mental health issues have been considered one of the most common causes of disability [2]. Despite the availability of mental health services, people hesitate to reach them because of the public stigma associated with mental health. Even worse, there is a severe shortage of mental health workers [3]. Due to the lack of resources, it is challenging to offer interventions using one-to-one traditional therapeutic methods. Accordingly, insufficient services have facilitated the utilization of technology to meet the needs of people suffering from mental distress. One of the technological solutions is the chatbot, a system capable of conversing and interacting with human users using spoken languages.

Through recent years, chatbots have become popular in the natural language processing community. Recent advances show that deep neural networks can be effectively applied to the development of task-oriented and open-domain conversational systems [4], [5], [6]. Nowadays, most existing systems can generate appropriate responses from the syntactic and contextual points of view. However, one of the challenges is identifying emotions from the conversational human counterpart and making a suitable response correspondingly. One reason behind it is the inadequacy of a curated conversational datasets consisting empathetic responses.

Within the last few years, building emotion-labelled datasets for the purpose of developing dialogue systems that generate emotion-aware and empathetic responses has gained much research interest. For instance, Chen et al., (2018) introduced EmotionLines [7], a dataset containing 2K dialogues, in which each utterance is manually annotated with an emotion label. One of seven emotions, six Ekman's basic emotions [8] plus the neutral emotion, is used to annotate each utterance. Another realization comes from EmoContext [9], published by A. Chatterjee et al.. This dataset consists of 30K conversations annotated with one of four emotion labels: Happy; Sad; Angry; and Others. Though these datasets are well designed using the interaction between human users and chatbots, their shortcomings still pose several problems to build a useful chatbot with empathy. First, none of these can be used for emotional reasoning as they lack the necessary annotation details required for the reasoning task. Specifically, in the EmoContext dataset, an emotion label is assigned to only the last utterance of each conversation. Second, the datasets are annotated with coarse-grained emotions and often contain a label 'Neutral' or 'Others', which introduces vagueness. Third, the datasets are often limited in size, which makes them difficult to be used in training neural chatbots.

Rashkin et al., (2019) proposed EmpatheticDialogues [10], a novel dataset containing 25K dialogues grounded in emotional situations. Expressly, each conversation is comprised of communication between a speaker and a listener, where the speaker initiates the dialogue by giving a setting according to a given emotion label and the listener responds based on the underlying situation. The authors consider 32 emotion labels, with a single label given in each dialogue for making a situation strongly related to one particular emotional experience. Welivita and Pu (2020) extend the above dataset by annotating each utterance with 32 emotions and additional 9 empathetic response intents [11]. Nevertheless, the limited size of the EmpatheticDialogues dataset is not enough to train a robust chatbot that can deal with different emotional situations. Thus to fill the above gap, in this project, we build a larger curated dialogue dataset, named RED (Reddit Emotional Distress) containing emotional support for people in acute distress. In addition, we present our detailed analysis on our proposed dataset to evaluate the empathetic dialogue characteristics between speakers and listeners. The ultimate goal of this dataset is to be utilized for training a robust conversational agent that can recognize human feelings and offer emotional support consistently.

Our data curation pipeline contains 4 main stages (see Figure 1):

1) Web Scraping: To develop a conversational dataset for training a chatbot, first and foremost we need to collect the relevant data. In this stage, we select and scrape the text data from several empathy-related subreddits in Reddit.

2) Extract Conversations: The conversations among different Reddit users can be either dyadic or multiparty. If a conversation is dyadic, it is communication between a speaker and a listener. On the other hand, a multiparty dialogue is inclusive of a speaker and multiple listeners. In this stage, we build both dyadic and multiparty conversations out of the scraped Reddit conversational threads.

3) Preprocessing: In this stage, we clean the extracted conversations by removing HTML tags, URLs and numbers. Since our goal is to develop a dialogue dataset containing empathetic responses, we also detect and remove the listener utterances containing profane words from the dataset.

4) Exploratory Data Analysis (EDA): After distilling the raw data, in this state, we conduct an analysis to find out the descriptive statistics about the curated data. In addition, sentiment analysis and emotion prediction are utilized to observe sentiment and emotion distribution among speakers and listeners.

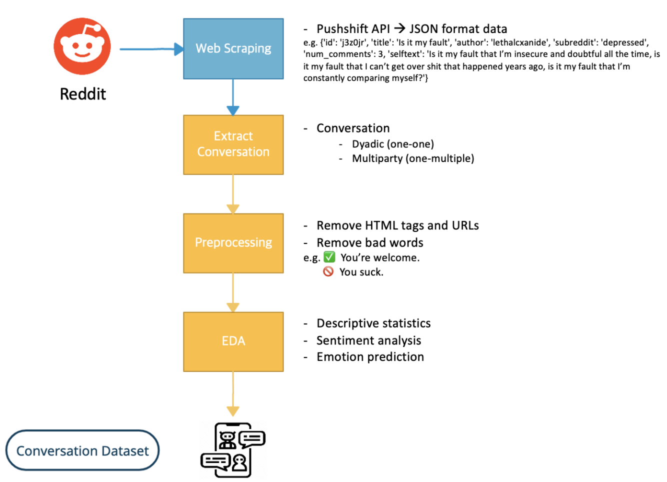


Fig. 1: Data Curation Pipeline

In the following sections, we describe the above steps in detail and present our observations related to exploratory data analysis. The main contributions of this paper are two-fold: 1) we curate a dialogue dataset containing two million dialogues out of Reddit conversational threads. These dialogues express emotional distress and how listeners offer emotional support for those in distress; 2) we thoroughly analyze the sentiment and emotional characteristics of the speaker and listener turns in this dataset and present our observations so that they can be useful in the design and development of empathetic chatbots that can offer emotional support to people in distress.

II. Literature Review

A number of dialogue datasets were previously constructed to make chatbots understand users' emotions and responses suitably. For example, Bertero et al. (2016) [12] built a dataset from TED-LIUM corpus [13] for real-time speech emotion and sentiment recognition in interactive dialogue systems. The data are in the form of audio and text, annotated with six emotion categories: criticism, anxiety, anger, loneliness, happiness, and sadness. Overall, the dataset contains 32,464 dialogue segments. Busso et al. (2008) [14], McKeown et al. (2011) [15], and Poria et al. (2018) [16] proposed IEMOCAP, SEMAINE, and MELD datasets, respectively. These datasets contain visual, acoustic, and textual signals. In general, they consist of various back-channel communication via facial expressions and speech tones, but the text may not fully represent contextual intents.

More recent works such as EmotionLines (Chen et al., 2018) [7] and OpenSubtitles (Lison et al., 2019) [17] are conversation datasets equipped with TV or movie transcripts translated from voice to text. Specifically, each dialogue turn in the EmotionLines corpus is labeled with an emotion based on its textual content. The dialogues are collected from Friends TV scripts and Facebook Messenger dialogues. One of seven emotions (six Ekman's basic emotions plus the neutral emotion) is labeled on each utterance by Amazon MTurkers. Overall, a total of 29,245 utterances from 2,000 dialogues are labeled in EmotionLines. Besides, OpenSubtitles is composed of 3.7 million subtitles in over 60 languages. The authors have made use of a quality score determined using a neural network trained on a sample of aligned sentence pairs, to filter out low-quality utterances. Though these works intend to build the dialogue datasets by improving the sentence quality and adding emotion labels, they are still unable to fully model the interactions occurring only via text. Meanwhile, A. Chatterjee et al. (2019) introduced a purely text-based dataset, EmoContext [9]. Given a textual dialogue along with two previous turns of context, the goal is to infer the underlying emotion in the utterance by choosing from four emotion classes: Happy; Sad; Angry; and Others. To facilitate the participation in this task, textual dialogues from user interaction with a chatbot are taken and annotated with emotion classes.

To guarantee empathetic responses, Rashkin et al. (2019) presented EmpatheticDialogues dataset [10], inclusive of 24,856 human-human conversations to train and evaluate dialogue systems to enable them to converse in an empathetic manner. Each conversation is based on a scenario related to one of 32 given emotions. In the stage of data collection, the conversations were generated by 810 Amazon MTurkers using the ParlAI platform [18]. Since a conversation model trained for empathetic responding needs to deal with the less frequently chosen labels, the workers were forced to select an emotion label of the three least-chosen labels to ensure balanced emotion coverage. Overall, almost all the conversations are empathetic, purely textual, and without any toxic response. Extending the above, Welivita et al. (2020) [11] developed a taxonomy of empathetic listener intents by analyzing the EmpatheticDialogues dataset. They automatically labeled all speaker and listener utterances of the EmpatheticDialogues dataset. In total, 32 types of emotion categories and 9 response intents were used to annotate the conversation messages. Lastly, the results validate that the taxonomy can be utilized to develop empathetic chatbots to achieve more interpretability in the generated responses. However, due to the limited size of this dataset, it is difficult to use it to train neural chatbots that could generate robust empathetic responses. Also, to the best of our knowledge, a dataset that specifically portrays conversations between speakers who are emotionally distressed and listeners who actively offer emotional support to them is lacking in the literature. This type of conversations could be available as recorded therapy sessions between mentally distressed patients and therapists. However, they are not available to the public due to privacy reasons. Our aim here is to curate such a dataset, which is both large-scale and contains emotional support that could potentially be used to train a therapeutic neural chatbot.

III. Methodology

As shown in Figure 1, our data curation pipeline consists of four main stages: 1) web scraping; 2) conversation formation; 3) preprocessing; and 4) exploratory data analysis. The sections below describe these steps in detail.

A. Web Scraping

Nowadays, social media platforms incorporate countless textual dialogues. For example, in 2020 Facebook has approximately 1.73 billion daily active users who visit the networking site for communication [19]. At the same time, Twitter has roughly 180 million active users interacting daily [20]. These social media platforms contain a massive number of conversations generated by users. Still, they have changed permissive data access provisions due to major scandals around data privacy ethics that occurred in recent years [21]. Consequently, the ability to collect timely data and reproduce findings has been curtailed.

Although the popular social media restrict people from accessing their data completely, some online communities such as Wikipedia, GitHub, and Reddit still continue to offer open application programming interfaces (APIs) and data dumps, which are valuable for researchers. Specially, Reddit consists of millions of subreddits, hundreds of millions of users, and hundreds of billions of comments, which are all accessible to people. Because of its abundance and variety, we chose Reddit to specifically gather conversations that provide support for people in emotional distress.

To curate dialogues containing emotional support for people in distress, we need conversations between speakers showing mental distress and listeners showing empathy and emotional support towards the speakers. In this regard, we choose 8 subreddits where such conversations were present: *depression*; *depressed*; *Off My Chest*; *SuicideWatch*; *Depression Help*; *sad*; *Anxiety Help*; and *Mental Health Support*. All of the subreddits provide abundant text data unveiling the conversations between authors of Reddit posts (speakers) undergoing personal distress and authors of comments (listeners) providing supportive remarks. Besides, the comments integrate real, personal understanding of a speaker's feeling as encouragements.

Conversation ID	Sub-reddit	Post title	Author	Dialog turn	Text
0000001	sad	I was depressed	abc	1	I was depressed due to work.
0000001	sad	I was depressed	u_1	2	It's so sad to hear
0000002	sad	I was depressed	abc	1	I was depressed due to work.
0000002	sad	I was depressed	u_2	2	I have also had experience with that.
0000002	sad	I was depressed	abc	3	I don't care
0000002	sad	I was depressed	u_2	4	I do some stress release meditation.
0000003	sad	I was depressed	abc	1	I was depressed due to work.
0000003	sad	I was depressed	u_2	2	Take some time to do Yoga.
0000003	sad	I was depressed	abc	3	Sure, I will. Thanks

Fig. 2: Example of a data frame after transforming the raw data extracted from Reddit.

Of many open APIs, we utilized Pushshift's APIs [22] to collect and process the text data from various subreddits containing conversations related to emotional distress. Practically, Pushshift is easier to query and retrieve a large quantity of historical data without strict limitations. For instance, it has a size limit five times greater than the Reddit official APIs, limiting only 100 objects per usage. Additionally, Pushshift makes all the submissions and comments from 2005 to 2019 available. If we parsed all its data, the dataset would be composed of 651,778,198 submissions and 5,601,331,385 comments posted on 2,888,885 subreddits.

B. Extracting Conversations

The data scraped using the Pushshift APIs comes in the form of JSON data. To facilitate further analysis, we transformed the data into data frames. An example of a data frame is depicted in Figure 2.

Following the transformation, the conversations could be classified into dyadic and multiparty conversations separately. Dyadic conversations were built by extracting a post and its first comment thread, and the conversations were restricted to the post authors and the authors of the first comment. Multiparty dialogues were built by extracting a post and its longest comment thread, including all the authors involved in the conversation. Figures 3 and 4 show an example of how dyadic and multiparty conversations were extracted. Note that in these figures the speaker turns are highlighted in red while the listeners turns are highlighted in blue.

C. Preprocessing

We transformed the raw text data into a more cleaned format by detecting and removing the HTML tags, and URLs, and replacing the numerals with a special tag <NUM>. However, various punctuation marks, emoticons, and emojis were preserved since they can be used as indicators to identify users' feelings.

Profane words also have a direct impact on human emotions. They are a spontaneous reflection of intense emotional states such as anger, fear, or passion. They are unequaled in their

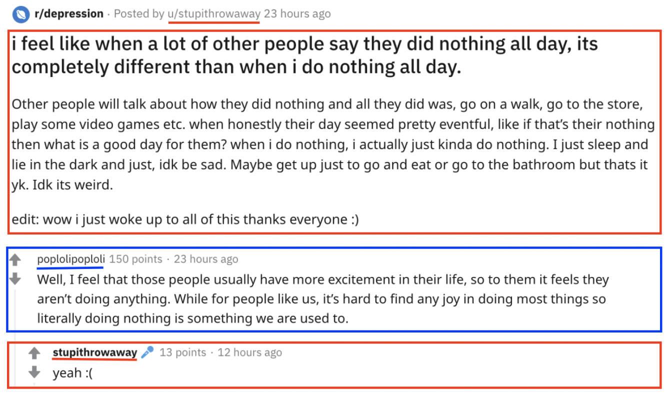


Fig. 3: Example of a Reddit dyadic dialogue containing three dialogue turns. Only two users are enagaged in the conversation.

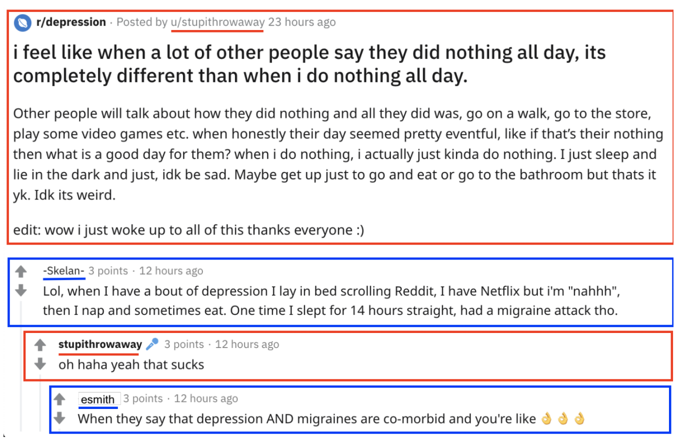


Fig. 4: Example of a Reddit multiparty dialogues containing four dialogue turns. Multiple users are engaged in the conversation.

capacity to inflict emotional pain and incite violent disagreement. In short, bad words are negatively powerful, specially imposing a risk of severely harming those who are mentally fragile.

To mitigate the aggressiveness, it is important to remove profanity from the comment threads. Moreover, since our goal is to curate a dataset to build an empathy-oriented chatbot, getting rid of profanity becomes apparent. To fulfill this need, we applied `profanity-check` [23], a fast and robust library to detect offensive language. Instead of using hard-coded lists of profane words, it makes use of linear Support Vector Machine [24] trained on 200k human-labeled samples of clean and profane text. It is simple but surprisingly effective to generalize profanity checking. Also, when exploiting `profanity-check` on a text message, it returns the probability of predicting profanity. Thus, we could set up a threshold to classify the message as profane or not. In our case, we manually set the threshold to be as high as 0.95 because the users sometimes express their feeling aggressively but with no mean intention. This threshold was determined after a thorough inspection of the profane text returned at different thresholds. Above this threshold, the message in a dialogue turn is classified as cussing, and the entire dialogue turn, along with the turns that follow, were removed from the conversation to maintain consistency.

D. Exploratory Data Analysis

Exploratory data analysis helps to discover patterns from data via summary statistics and graphical representations. Here, we focus on analysing the RED conversations. In particular, we mainly analysed the distribution of dialogues and their turns with respect to each subreddit. Then, we constructed tables to showcase the descriptive statistics.

Next, we conducted sentiment analysis separately on speaker and listener turns in our dialogue dataset and visualized their distribution. We hypothesized that the listeners should express more positiveness than the speakers since they are offering support to the speakers to uplift the mood. To confirm our hypothesis, we applied *Vader* [25], a lexicon and rule-based sentiment analysis tool specifically attuned to sentiments expressed in social media. Given a textual message, *Vader* is capable of classifying it to one of three classes: positiveness; negativeness; or neutrality. Particularly, *Vader* provides the compound score computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive). Then, by tuning the threshold, this useful metric can be exploited to obtain a single uni-dimensional measure of sentiment for a given sentence.

On top of that, we conducted emotion and intent analysis on speaker and listener turns in the RED dataset, separately. We used the EmoBERT classifier [11] trained on the EmpatheticDialogues dataset for this purpose. EmoBERT is a BERT [26] based emotion classifier, which predicts the emotion or intent of a particular dialogue turn. Its structure is composed of a representation network based on BERT and a classification network. During training, the weights inside the representation network are initialized from the pre-trained language model, RoBERTA [27]. Then, the model is fine-tuned on the situation descriptions from Empathetic Dialogues Dataset [10] labeled with 32 emotions and listener utterances tagged with 9 empathetic response intents. The model is trained on a total of 25,023 sentences, validated using a validation set containing 3,544 sentences, and tested using a testing set containing 3,225 sentences. It has a significant accuracy of 65.88% on the testing dataset.

IV. Results

A. RED dataset

In general, the resultant RED dataset contains approximately 1.9 million conversations, with 1.3 million dyadic dialogues and 0.6 million multiparty dialogues. On average, there are roughly 4 turns inside a conversation. The speakers tend to convey negative attitudes while the listeners are inclined to express more positiveness. Nonetheless, as we could observe during emotion prediction, the speakers and listeners could possibly communicate in both pessimistic and optimistic tone. Figure 5 and 6 showcase examples of dyadic and multiparty conversations present in the RED dataset.

Situation: Looking for emergent help…
Conversation:
Format: Turn # (sentiment, emotion)
Turn 1 (negative, sentimental)
Speaker: I can't do this anymore and I can't see any reason to stay alive. I am just scared. Where can I find the courage ?

Turn 2 (positive, encouraging)
Listener: I know this sounds cliche but you *need* to think of the impact that will be left on your family and friends. I recently found the song Joyner Lucas - I'm Sorry (508)-507-2209 and it really helped to show me the other side of the coin and showed the emotions others will feel. Music has helped me a lot so far and I am amping myself up to seek professional help and I advise you to do the same.

Fig. 5: Example of a dyadic dialogue in the RED dataset annotated with sentiment and emotion labels.

Situation: My boss thinks I'm a liar.
Conversation:
Format: Turn # (sentiment, emotion)
Turn 1 (negative, afraid)
Speaker: So I've suffered from some form of anxiety/depression all my life. It's worsened over the last few months as I left a comfortable job where I was relatively happy for a promotion within the same company. Basically I was pushed into the job, I had second thoughts at the beginning but was assured that my new manager would be supportive, my training would be excellent, the job would be amazing for me, etc etc. This couldn't be further from the truth. The people I work with are judgemental, they aren't supportive at all and I'm being bullied by an older woman who I think is resentful of me. It's got to the point that even walking into the building where I work affects my breathing and I feel like I'm dying. I finally found the courage to voice all of this to my manager today, who basically outright called me a liar and said she "just didn't know what was wrong with me". I'm absolutely devastated, I can't go back to my old job because my position was filled. I have bills to pay and no way of paying them unless I go back to this job but the thought of doing that makes me want to die. I feel like I could just end it but I wouldn't solely because of my mother who would be devastated. I don't even know what advice I'm looking for really, I just don't know what to do.

Turn 2 (positive, consoling)
Listener 1: You don't have to face it alone, get help from people who have been there [URL]

Turn 3 (positive, agreeing)
Speaker: Thank you, means a lot that you replied❤️

Turn 4 (positive, anticipating)
Speaker: Spoke to my doctor who is extremely supportive and understanding which helped a lot, and he has referred me for counselling. I took some time off work and am happy to say within that time I've found a new job with a different company. I hope this is a step in the right direction

Turn 5 (positive, caring)
Listener 2: Great! Keep up the good OP. I'll be waiting for your updates. Feel free to PM me if you want to chat :)

Turn 6 (positive, grateful)
Speaker: Thank you so much!

Fig. 6: Example of a multi-party dialogue in the RED dataset annotated with sentiment and emotion labels.

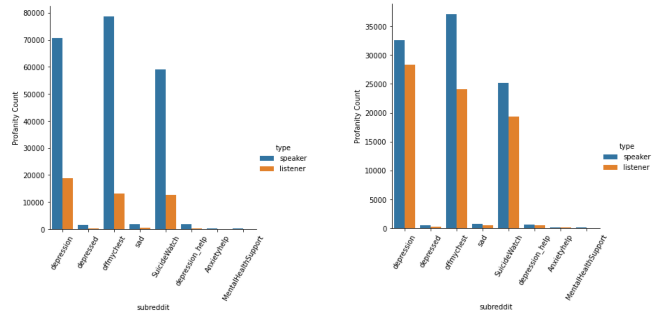
B. Profanity Detection

Figure 7 shows the distribution of speaker and listener turns that are detected as profane in each subreddit. We could see that profane words are mostly contained inside popular subreddits such as *depression*, *Off My Chest*, and *SuicideWatch*. In particular, the speaker turns contain far more profane words than the listeners turns. In the final dataset, we retained the profane speaker turns and only removed the profane listener turns since the chatbot could potentially use the profanity present in the speaker turns to understand the intensity of speaker's emotion.

C. Descriptive Statistics

Succeeding removing offensive language from the conversations, further analysis on the RED dataset revealed more characteristics of dialogues belonging to each subreddit. Statistical analysis on the dataset releaved the following.

1) Total number of dyadic and multiparty conversations per subreddit: Figure 8 shows the distribution of the number of dyadic and multiparty dialogs across subreddits. The subreddits *depression*, *Off My Chest*, and *SuicideWatch* contribute conversations mostly to the RED dataset. All the three subreddits consist of much more dyadic dialogues than multiparty ones.



(a) Dyadic (b) Multiparty

Fig. 7: Distribution of profane speaker and listener turns across subreddits.

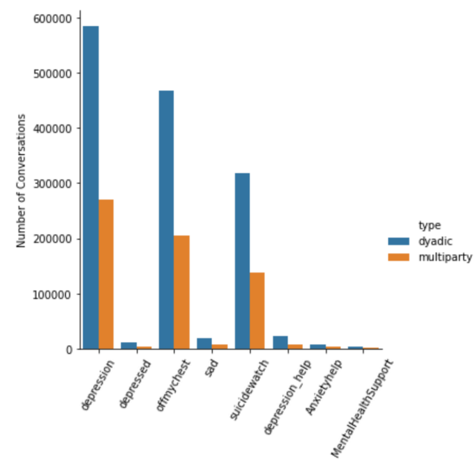


Fig. 8: Distribution of the number of dyadic and multiparty dialogs across subreddits.

2) Total number of dialogue turns per subreddit: Similar to the comparison of the number of dialogs, Figure 9 shows the distribution of dialog turns in dyadic and multiparty conversations in each subreddit. The three subreddits *depression*, *Off My Chest*, and *SuicideWatch* have considerably more turns than the rest. Despite the majority of dyadic conversations present in the dataset, the multiparty conversations contribute more turns than the dyadic ones.

3) Turn Distribution: The distribution of dialogue turns in the dataset reveals the level of engagement between speakers and listeners. Figure 10 shows the distribution of dialogue turns in the entire dataset. It implies that most conversations are dyadic and end up in two turns. However, taking the *sad* subreddit as an example, in Figure 11 we show that there are few lengthy conversations taking place in all subreddits. In some cases the dialogues span over 100 turns. We have included visualizations of the turn distributions in other subreddits as Appendix.

4) Summary: Tables I and Table II display the summary of descriptive statistics of both dyadic and multiparty conversations present in the entire dataset as well as in individual

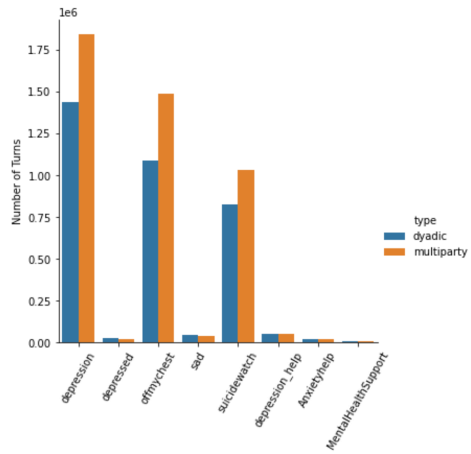


Fig. 9: Distribution of dialogues turns in dyadic and multiparty conversations in each subreddit.

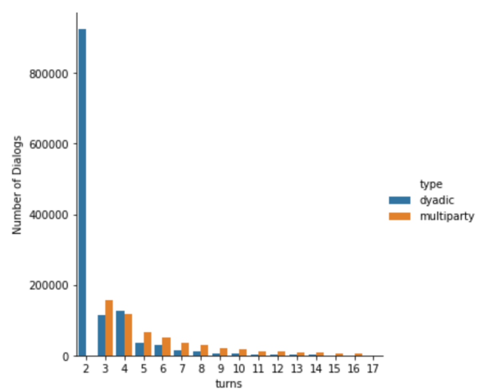
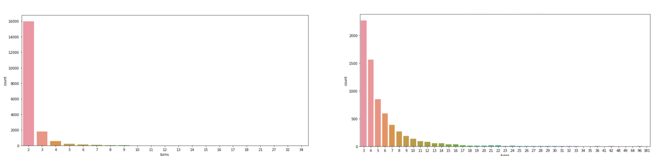


Fig. 10: Distribution of dialog turns in the RED dataset



(a) Dyadic

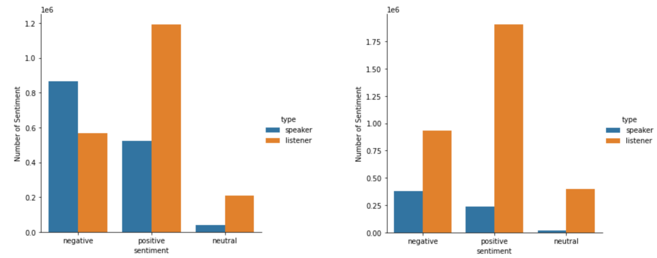
(b) Multiparty

Fig. 11: Distribution of dialog turns in the *sad* Subreddit

subreddits. Note that when computing the number of tokens, punctuation marks and emojis are also included, since they are important indicators of predicting emotion.

D. Sentiment Analysis

In Figure 12, we summarize the results of sentiment analysis conducted separately on speaker and listener turns in the RED dataset. As a whole, the plots reasonably reflect that the speaker sentiments are more negative while the listener sentiments are more positive. Thus, the dataset indeed corresponds to the fact that the positivity and encouragement offered by listeners can truly assist in uplifting the mood of people going through negativity and distress.



(a) Dyadic

(b) Multiparty

Fig. 12: Sentiment Distribution between Speakers and Listeners

E. Emotion Detection

Figures 13 and 14 present the predicted emotion distributions of dyadic and multiparty dialogues. As per the results obtained, in dyadic conversations, the speakers are likely to express positive emotion, such as *impressed*, *hopeful*, and *encouraging* as well as negative emotion, such as *furious* and *ashamed*. Correspondingly, the listeners tend to convey both positive and negative emotions the same way as the speakers. In multiparty conversations, we could observe more *afraid* and *sad* emotions being expressed compared to the dyadic ones. Turn-wise analysis is further needed to derive at specific conclusions.

V. Discussion

We followed web scraping, conversation extraction, preprocessing, and EDA steps to curate the RED dialogue dataset consisting of two million dyadic and multi-party conversations. We outlined the details of the above stages and the results obtained through EDA. The EDA gives us a basic overview of the entire dataset. We observed that most conversations come from *depression*, *Off My Chest* and *SuicideWatch* subreddits. Most conversations are dyadic with their length limited to two turns. The sentiment analysis results clearly indicates that the speakers mostly express negativity while the listeners mostly respond back positively indicating support and encouragement. The emotion prediction results showed that both speakers and listeners tend to express positive as well as negative emotions and also intents. However, turn-wise analysis is further needed to derive specific conclusions on how the emotions and intents evolve as the dialogues proceed.

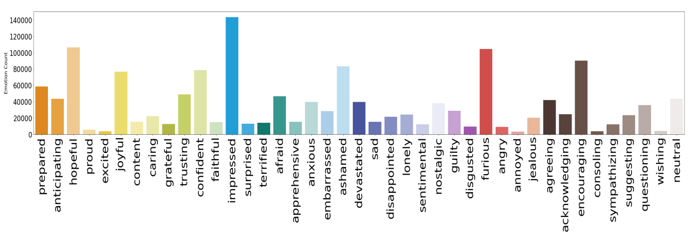
There are some limitations in this work. The classifier used to perform emotion analysis on the dataset is trained on short text conversations from EmpatheticDialogues and has a limited accuracy of 66%, which results in quite noisy emotion and intent analysis on the resultant dataset. Since, dialogue turns in Reddit are quite lengthy (on average 84.89 tokens per dyadic dialogue turn and on average 50.52 tokens per multi-party dialogue turn), different sections of the same turn may carry different emotions or intents, and this requires an advanced emotion and intent classifier to separately identify these parts and separately predict an emotion or intent label for these parts. Fine-tuning the emotion classfier on part of Reddit data with manual annotations could also help in improving performance.

Subreddit	No. of Dialogs	No. of Turns	No. of Tokens	Avg No. of Turns per Dialog	Avg No. of Tokens per Dialog	Avg No. of Tokens per Turn
Entire	1,275,486	3,396,476	288,336,762	2.66	226.06	84.89
r/depression	510,035	1,396,044	106,967,833	2.74	209.73	76.62
r/depressed	10,892	23,804	1,940,000	2.19	178.11	81.50
r/offmychest	437,737	1,064,467	109,459,738	2.43	250.06	102.83
r/sad	18,827	42,293	3,088,562	2.25	164.05	73.03
r/SuicideWatch	262,469	791,737	59,267,000	3.02	225.81	74.86
r/depression_help	23,678	51,849	5,412,390	2.19	228.58	104.39
r/Anxietyhelp	8,297	18,351	1,428,287	2.21	172.14	77.83
r/MentalHealthSupport	3,551	7,931	772,952	2.23	217.67	97.46

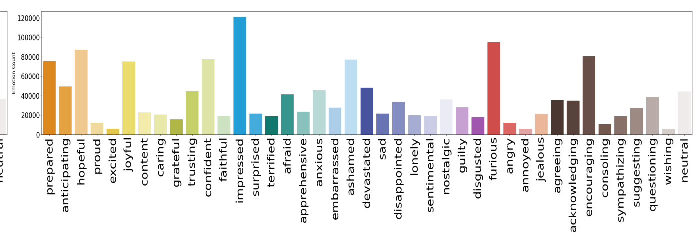
TABLE I: Descriptive statistics of dyadic conversations in the entire dataset as well as in each subreddit.

Subreddit	No. of Dialogs	No. of Turns	No. of Tokens	Avg No. of Turns per Dialog	Avg No. of Tokens per Dialog	Avg No. of Tokens per Turn
Entire	584,427	3,863,841	195,187,799	6.61	333.98	50.52
r/depression	246,268	1,609,795	76,789,493	6.54	311.81	47.70
r/depressed	3,434	18,658	923,429	5.43	268.91	49.49
r/offmychest	196,566	1,232,645	69,483,975	6.27	353.49	56.37
r/sad	6,756	35,085	1,577,803	5.19	233.54	44.97
r/SuicideWatch	119,577	899,460	42,468,629	7.52	355.16	47.22
r/depression_help	7,758	45,446	2,791,784	5.86	359.86	61.43
r/Anxietyhelp	2,990	16,959	825,710	5.67	276.16	48.69
r/MentalHealthSupport	1,078	5,793	326,976	5.37	303.32	56.44

TABLE II: Descriptive statistics of multi-party conversations in the entire dataset as well as in each subreddit.

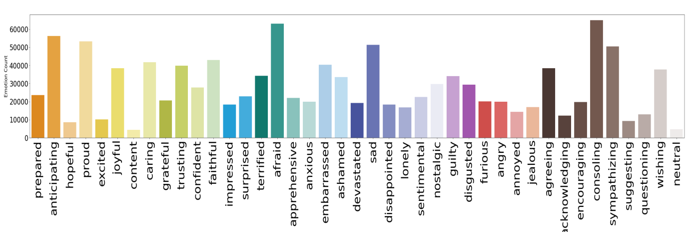


(a) Speakers

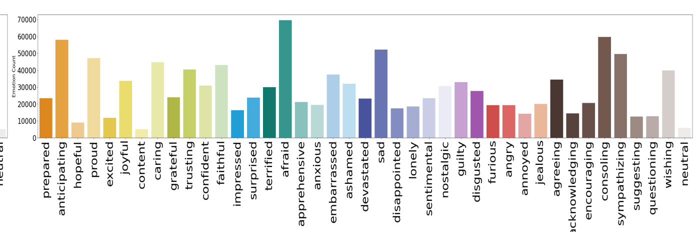


(b) Listeners

Fig. 13: Emotion Prediction in Dyadic Dialogues



(a) Speakers



(b) Listeners

Fig. 14: Emotion Prediction in Multiparty Dialogues

As future work, we can further extend the dataset by extracting more dialogues from the scraped Reddit data. Due to resource limitations, in each post submission, we merely picked the first comment thread to extract dyadic conversations and the longest comment thread to extract multiparty dialogues. Therefore, it has more room for extension. In addition, dialogues from other forms of social media such as Quora and Tumblr, which also provide open APIs for parsing their content, can be incorporated into the dataset. Thus, the dataset can be more diversified to enable the trained chatbot to handle a variety of scenarios. Finally, this dataset containing two million dyadic and multi-party dialogues could be readily utilized to train neural chatbots that could offer empathetic support to people in distress.

VI. Conclusion

In the paper, we proposed the RED (Reddit Emotional Distress) dataset containing two million conversations related to mental distress and emotional support extracted from several carefully chosen subreddits from the Reddit social media platform. We outlined in detail the steps in curating and analyzing the proposed dialogue dataset. The results of exploratory data analysis unveiled the characteristics of the RED dataset. The sentiment analysis results confirmed our hypothesis that the speakers tend of express more negativity and the listeners tend to reply back with positivity offering support and encouragement to the speakers.

Considering the future work, we put forth several suggestions to improve our dataset. We plan to diversify our dataset by accommodating more conversations from different social media platforms. Furthermore, we discussed plans to better predict and analyse the emotion-intent distribution in the dataset. Altogether, our proposed dataset has the potential to be used for training a robust chatbot, which is capable of providing emotional support for people in distress.

VII. Acknowledgement

I would like to warmly thank everyone helping me towards my goal. The biggest nod of appreciation goes to my mentor, Kalpani Anuradha, who faithfully monitored my progress every week. In addition, I express my gratitude to Dr. Pearl Pu for arranging the project, enabling me to deepen my knowledge in data science.

References

[1] Zachary Steel, Claire Marnane, Changiz Iranpour, Tien Chey, John W Jackson, Vikram Patel, Derrick Silove, The global prevalence of common mental disorders: a systematic review and meta-analysis 1980–2013, International Journal of Epidemiology, Volume 43, Issue 2, April 2014, Pages 476–493, https://doi.org/10.1093/ije/dyu038

[2] The Global Burden of Disease: 2004 Update, World Health Organization, WorldHealth Organization, Geneva, 2008.

[3] A.N. Vaidyam, H. Wisniewski, J.D. Halamka, M.S. Kashavan and J.B. Torous, Chatbots and conversational agents in mental health: a review of the psychiatric landscape, Can. J. Psychiatry (2019), Article 0706743719828977

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014, Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 2104-3112.

[5] Oriol Vinyals, and Quoc Le. 2015. A neural conversational model. In Proceedings of the 31st International Conference and Machine Learning, Deep Learning Workshop.

[6] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1711-1721, Lisbon, Portugal.

[7] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, L.-W. Ku et al., "EmotionLines: An emotion corpus of multi-party conversations," arXiv preprint arXiv:1802.08379, 2018.

[8] Ekman, Paul (1999). Basic emotions. In Tim Dalgleish & M. J. Powers (eds.), Handbook of Cognition and Emotion. Wiley. pp. 4–5.

[9] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, "Understanding emotions intext using deep learning and big data, "Computers in Human Behavior, vol. 93, pp. 309–317, 2019.

[10] Hannah Rashkin and Eric Michael Smith and Margaret Li and Y-Lan Boureau, Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset, ACL, 2019.

[11] Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4886–4899.

[12] Bertero, Dario and Siddique, Farhad Bin and Wu, Chien-Sheng and Wan, Yan and Chan, Ricky Ho Yin and Fung, Pascale. Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016.

[13] A. Rousseau, P. Deléglise, and Y. Estève, "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.

[14] 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.

[15] 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.

[16] 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.

[17] Pierre Lison, J. 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).

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

[19] Maryam Mohsin, 10 Facebooks Statistics Every Marketer Should Know. Available at https://www.oberlo.com/blog/facebook-statistics, 2020.

[20] J. Clement, Twitter: number of monetizable daily active users worldwide 2017-2020, Available at https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide, 2020.

[21] S. Walker, D. Mercea, and M. Bastos. The disinformation landscape and the lockdown of social platforms. Information, Communication & Society, 2019.

[22] Jason Baumgartner and Savvas Zannettou and Brian Keegan and Megan Squire and Jeremy Blackburn. The Pushshift Reddit Dataset. International AAAI Conference on Web and Social Media, 2020.

[23] Victor Zhou, Domitrios Mistriotis and Vadim Shestopalov. Profanity-check, 2018, Github repository, https://github.com/vzhou842/profanity-check.git

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

[25] 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.

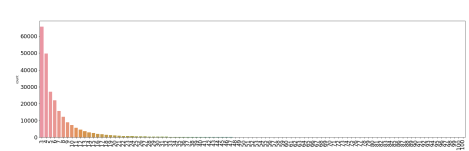
[26] Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

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

VIII. Appendix

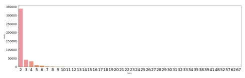
In *Turn Distribution* subsection, the distribution in the *Sad* subreddit is shown. Here, to get the whole picture from all the subreddits, their turn distributions are presented (Figure 11, 18, 19, 15, 16, 17, 20, 21).

(a) Dyadic

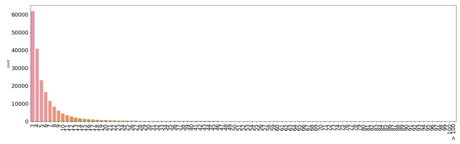


(b) Multiparty

Fig. 15: Turn Distribution in Depression Subreddit

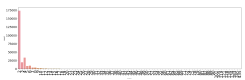


(a) Dyadic

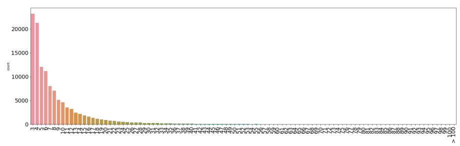


(b) Multiparty

Fig. 16: Turn Distribution in OffMyChest Subreddit

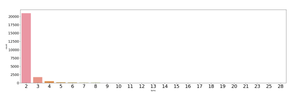


(a) Dyadic

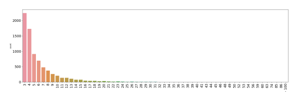


(b) Multiparty

Fig. 17: Turn Distribution in SuicideWatch Subreddit

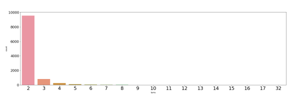


(a) Dyadic

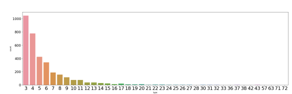


(b) Multiparty

Fig. 18: Turn Distribution in Depression_Help Subreddit

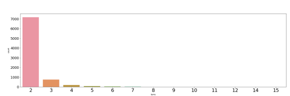


(a) Dyadic

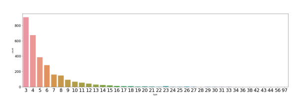


(b) Multiparty

Fig. 19: Turn Distribution in Depressed Subreddit

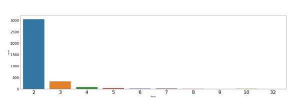


(a) Dyadic

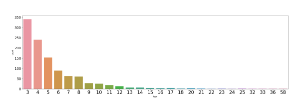


(b) Multiparty

Fig. 20: Turn Distribution in AnxietyHelp Subreddit



(a) Dyadic



(b) Multiparty

Fig. 21: Turn Distribution in MentalHealthSupport Subreddit