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

Sena Necla Çetin
Supervisor: Anuradha Welivita

Outline

- Motivation
- Goal
- Previous work
- Dataset
- Methods
  - Data cleaning
  - Measuring speaker engagement
  - Measuring speaker satisfaction
  - Hyper-parameter tuning
  - Manually analyzing listener intents
- Results
- Discussion
- Conclusion
- Future work

Motivation

- ~1 billion people worldwide suffer from a mental disorder (Ritchie and Roser, 2018).
- Only a low proportion can get help due to social stigma and lack of financial resources (Olfson et al., 2009).

→ Use AI-driven chatbots to help people deal with emotional distress.

- Conversational data between therapists and patients are unavailable due to privacy and ethical reasons.
- Most feasible option is to train chatbots on data from online forums such as Reddit.
  - They may include inappropriate conversation patterns which may lead to user dissatisfaction and disengagement with the chatbot.

→ Identify conversation patterns that lead to disengagement and dissatisfaction.

Example of a highly satisfying and less engaging dialogue

Speaker: My step mom makes me freak out (...) 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.

Example of a highly engaging and less satisfying dialogue

Speaker: My boyfriend's mother asked my boyfriend to ask me for $250 for a couple times now (...) I am tired of her asking me for money (..)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.

Goal

- Develop a novel scoring function that measures the level of speaker satisfaction and engagement in distress oriented conversations.
- 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.

→ These techniques can serve as a set of rules to design and develop automatic chatbots from online mental health community (OMHC) data so that inappropriate responses can be avoided and speaker satisfaction and engagement can be increased.

Previous 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.
  - Zhang and Danescu-Niculescu-Mizil (2020) quantify the forwards- and backwards-orientations of listener utterances.
  - Sharma et al. (2018) find that certain online mental health communities direct more emotional support while others provide more informational support.
  - Pfeil and Zaphiris (2007) investigate the patterns of empathy in online communication and compare them to patterns of offline communication.
  - Welivita and Pu (2020) define a taxonomy of listener specific empathetic response intents capable of supporting automatic empathetic communication in conversations.

Previous Work

- A majority of the prior work studying engagement between users in OMHCs 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) who propose four indicators of user engagement based on attention and interaction in two popular OMHCs, TalkLife and Reddit.
  - Attention-based indicators: number of dialogue turns + number of listeners in the conversation
  - Interaction-based indicators: time between responses + degree of interaction

Previous Work

- Previous work has measured speaker satisfaction in social chatbots that aim to improve the emotional state of its users using self-reported measures.
  - Vaidyam et al. (2019) investigate 10 social chatbots using self-reports of the participants.
  - Gennaro et al. (2020) study the impact of an empathetic chatbot on participants who experience social exclusion.

  → Lack the computational methods to apply on a large-scale dialogue dataset for measuring satisfaction.

  → Not focused well enough in negative response strategies that can make the user feel dissatisfied and disengaged with the conversation.

.

Dataset

- 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.
    - Audio:
        - TED-LIUM (Rousseau et al., 2014), IEMOCAP (Busso et al., 2008), SEMAINE (McKeown et al., 2011), MELD (Poria et al., 2018)
    - TV/Movie transcripts:
        - EmotionLines (Chen et al., 2018), OpenSubtitles (Lison et al., 2019)
    - Telephone recordings:
        - Switchboard (Stolcke et al., 2000)

→ Text data in these datasets may not fully represent contextual intents due to the existence of other channels of information.

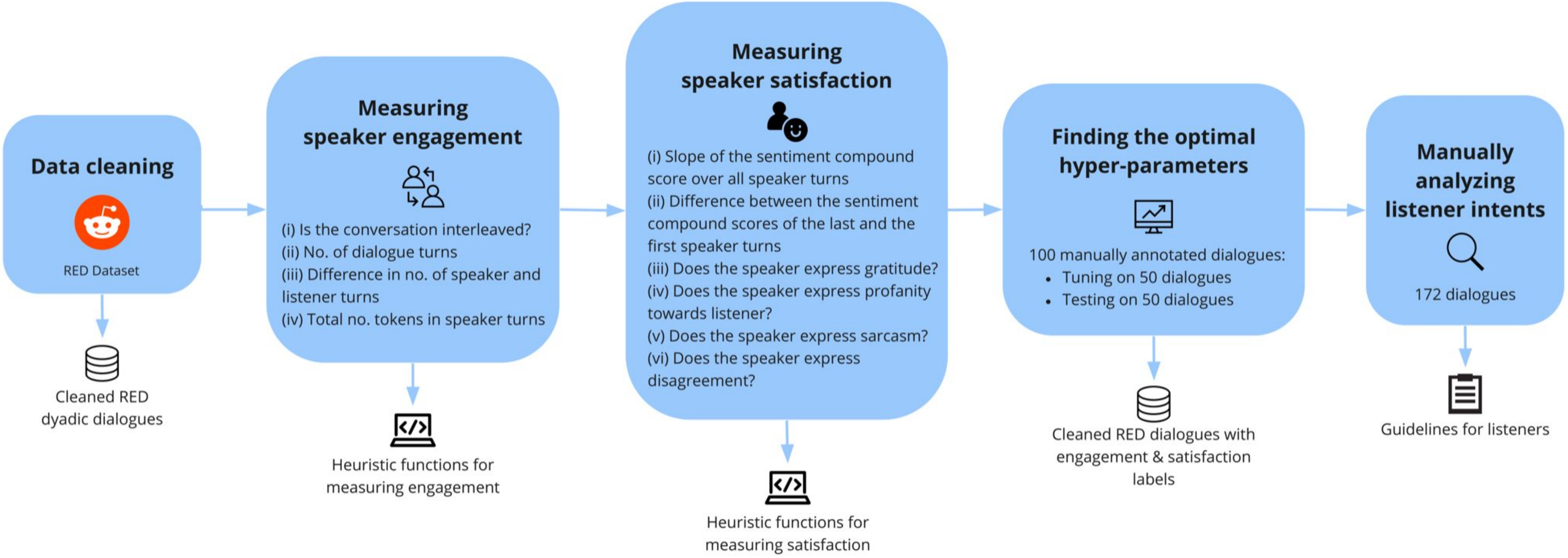
Dataset

- Purely text-based:
    - DailyDialog (Li et al., 2017)
      → Not guaranteed to contain empathetic responses.
    - EmpatheticDialogues (Rashkin et al., 2019)
      → Its limited size does not allow the training of a robust chatbot.
    - **Reddit Emotional Distress (RED)** (Yeh et al., 2020)
      → **Large-scale, preprocessed to contain almost no listener profanity, contains sentiment analysis and emotion prediction.**

Dataset

- RED
    - ~2 million conversations
        - 1.3 million dyadic and 0.6 million multiparty
    - 8 different emotional support subreddits
    - Most conversations end in 2 turns
    - Average number of dialogue turns is 4
    - Profanity removed from listener responses by profanity-check
    - Sentiment analysis by VADER
    - Emotion and intent analysis by the EmoBERT classifier trained on the EmpatheticDialogues dataset

Methods



Data cleaning

RED Dataset

Cleaned RED dyadic dialogues

Measuring speaker engagement

(i) Is the conversation interleaved?
(ii) No. of dialogue turns
(iii) Difference in no. of speaker and listener turns
(iv) Total no. tokens in speaker turns

Heuristic functions for measuring engagement

Measuring speaker satisfaction

(i) Slope of the sentiment compound score over all speaker turns
(ii) Difference between the sentiment compound scores of the last and the first speaker turns
(iii) Does the speaker express gratitude?
(iv) Does the speaker express profanity towards listener?
(v) Does the speaker express sarcasm?
(vi) Does the speaker express disagreement?

Heuristic functions for measuring satisfaction

Finding the optimal hyper-parameters

100 manually annotated dialogues:

- Tuning on 50 dialogues
- Testing on 50 dialogues

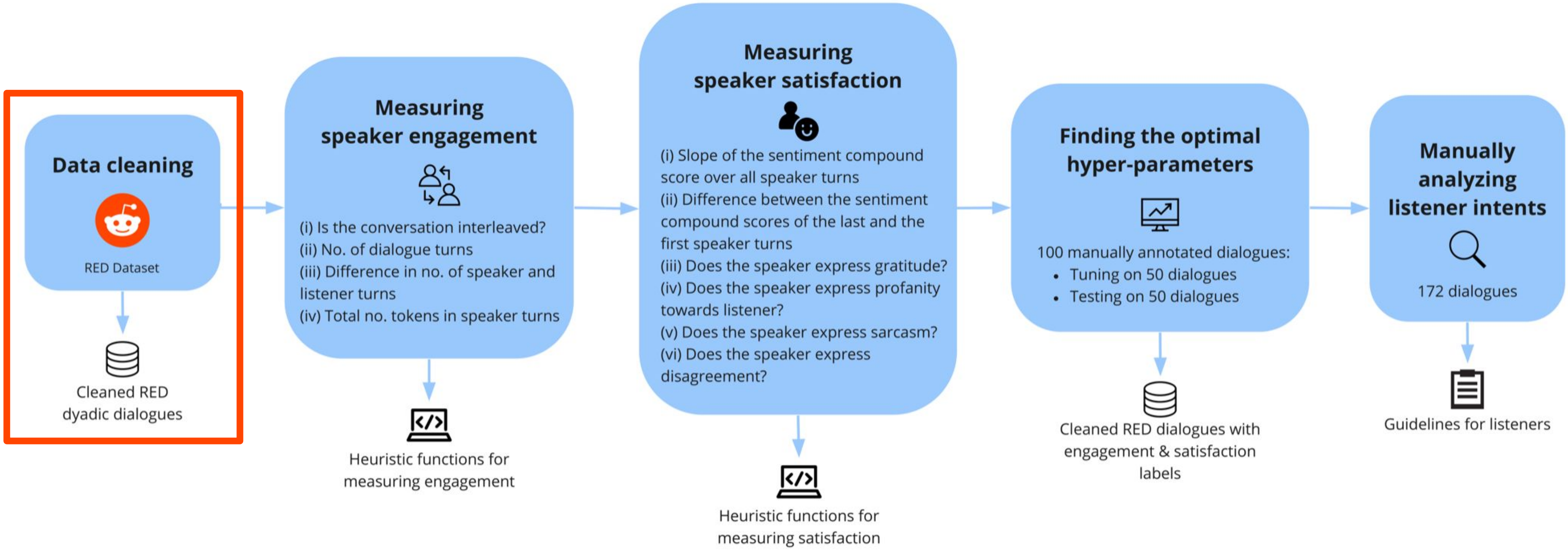
Cleaned RED dialogues with engagement & satisfaction labels

Manually analyzing listener intents

172 dialogues

Guidelines for listeners

Methods



Data cleaning

RED Dataset

Cleaned RED dyadic dialogues

Measuring speaker engagement

(i) Is the conversation interleaved?
(ii) No. of dialogue turns
(iii) Difference in no. of speaker and listener turns
(iv) Total no. tokens in speaker turns

Heuristic functions for measuring engagement

Measuring speaker satisfaction

(i) Slope of the sentiment compound score over all speaker turns
(ii) Difference between the sentiment compound scores of the last and the first speaker turns
(iii) Does the speaker express gratitude?
(iv) Does the speaker express profanity towards listener?
(v) Does the speaker express sarcasm?
(vi) Does the speaker express disagreement?

Heuristic functions for measuring satisfaction

Finding the optimal hyper-parameters

100 manually annotated dialogues:

- Tuning on 50 dialogues
- Testing on 50 dialogues

Cleaned RED dialogues with engagement & satisfaction labels

Manually analyzing listener intents

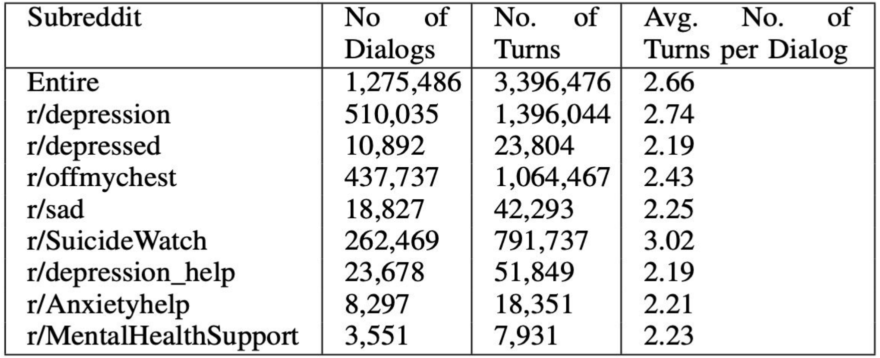
172 dialogues

Guidelines for listeners

Data Cleaning

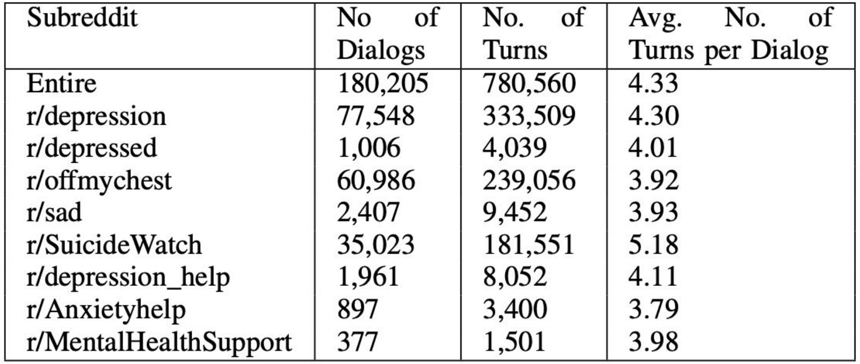
- Selected only the dyadic conversations.
- Removed duplicate turns.
- Removed conversations with less than 3 dialogue turns.
    - To be able to infer speaker satisfaction
- Removed faulty multiparty conversations inside the dyadic dataset.

Data Cleaning



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

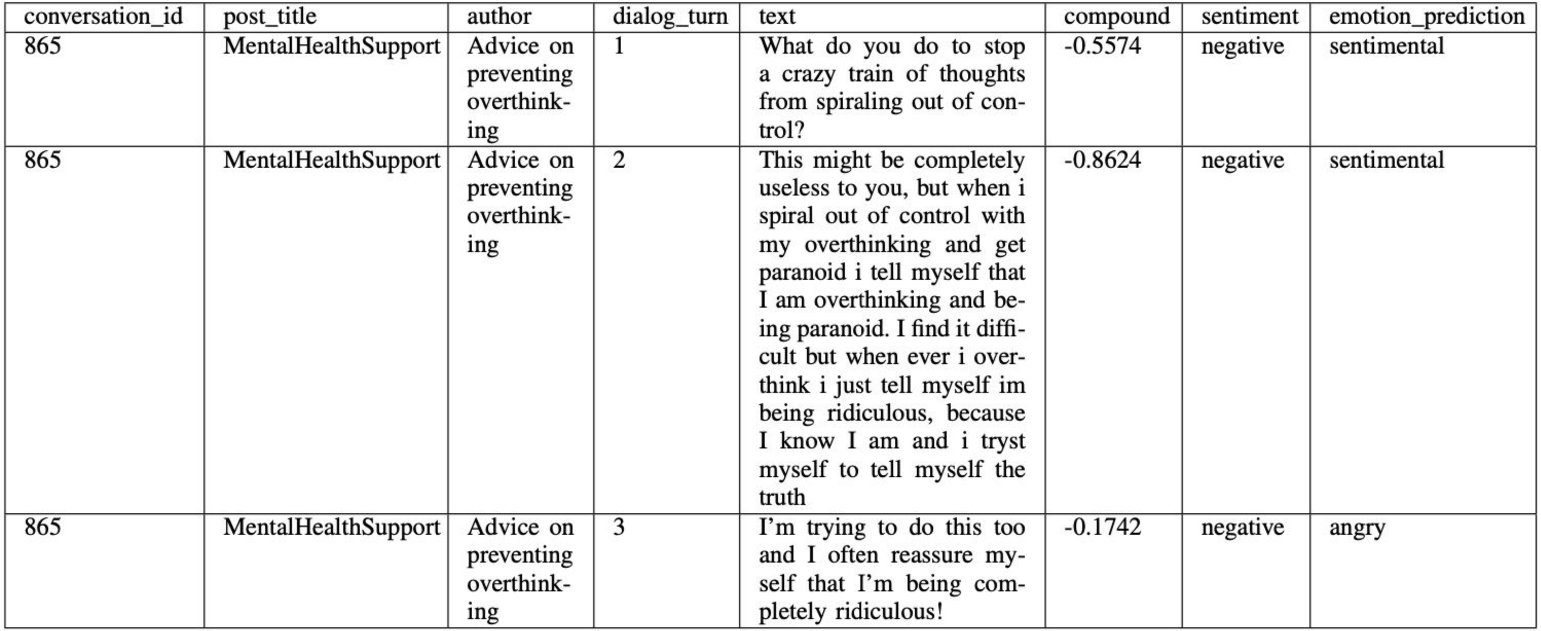
Table IV. Descriptive statistics of dyadic conversations in the entire dataset as well as in each subreddit (before cleaning)



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)

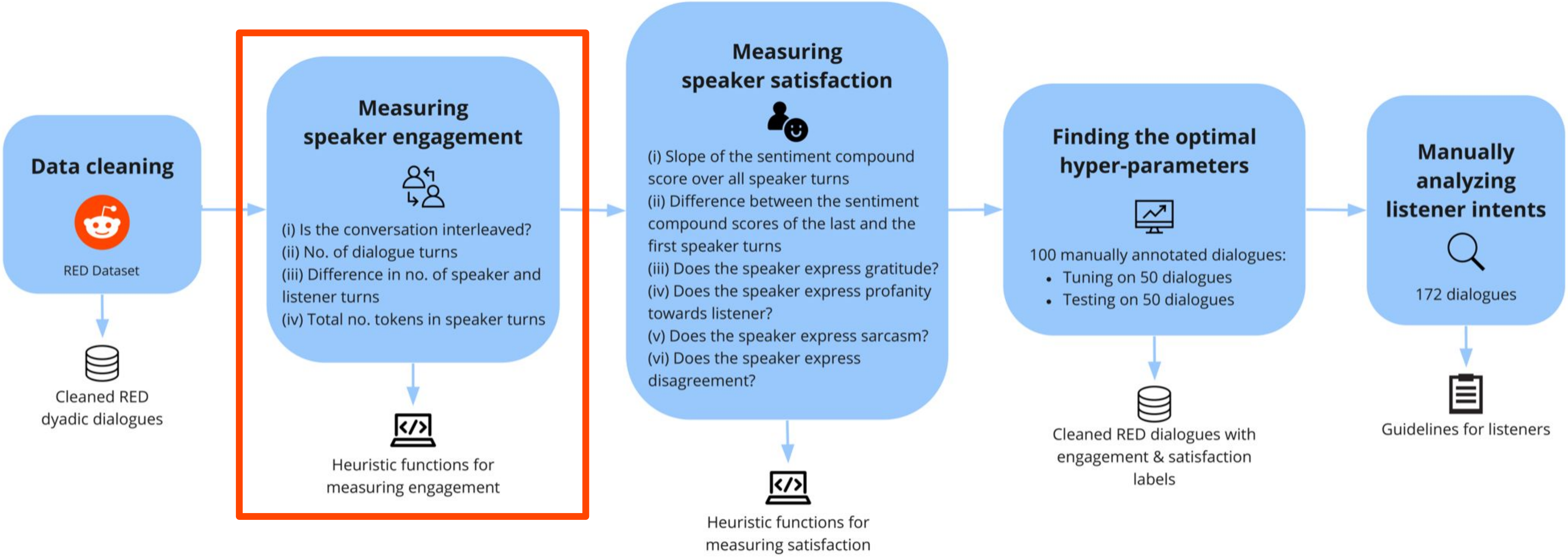
Data Cleaning



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

Methods



Data cleaning

RED Dataset

Cleaned RED dyadic dialogues

Measuring speaker engagement

(i) Is the conversation interleaved?
(ii) No. of dialogue turns
(iii) Difference in no. of speaker and listener turns
(iv) Total no. tokens in speaker turns

Heuristic functions for measuring engagement

Measuring speaker satisfaction

(i) Slope of the sentiment compound score over all speaker turns
(ii) Difference between the sentiment compound scores of the last and the first speaker turns
(iii) Does the speaker express gratitude?
(iv) Does the speaker express profanity towards listener?
(v) Does the speaker express sarcasm?
(vi) Does the speaker express disagreement?

Heuristic functions for measuring satisfaction

Finding the optimal hyper-parameters

100 manually annotated dialogues:

- Tuning on 50 dialogues
- Testing on 50 dialogues

Cleaned RED dialogues with engagement & satisfaction labels

Manually analyzing listener intents

172 dialogues

Guidelines for listeners

Measuring Speaker Engagement

- To measure speaker engagement, we applied various heuristic methods:

    i. Merged the consecutive speaker responses into a single speaker turn and checked if the conversation is interleaved.

    - ■ Based on models in communication theory (Bretz and Schmidbauer, 1983), (Williams et al., 1988).

    ii. Selected the number of dialogue turns.

    - ■ Similarly to Sharma et al. (2020).

    iii. Selected the absolute value of the difference in the number of speaker and listener turns.

    - ■ *Mutual Discourse* is the most desirable condition for OMHCs (Sharma et al., 2020).

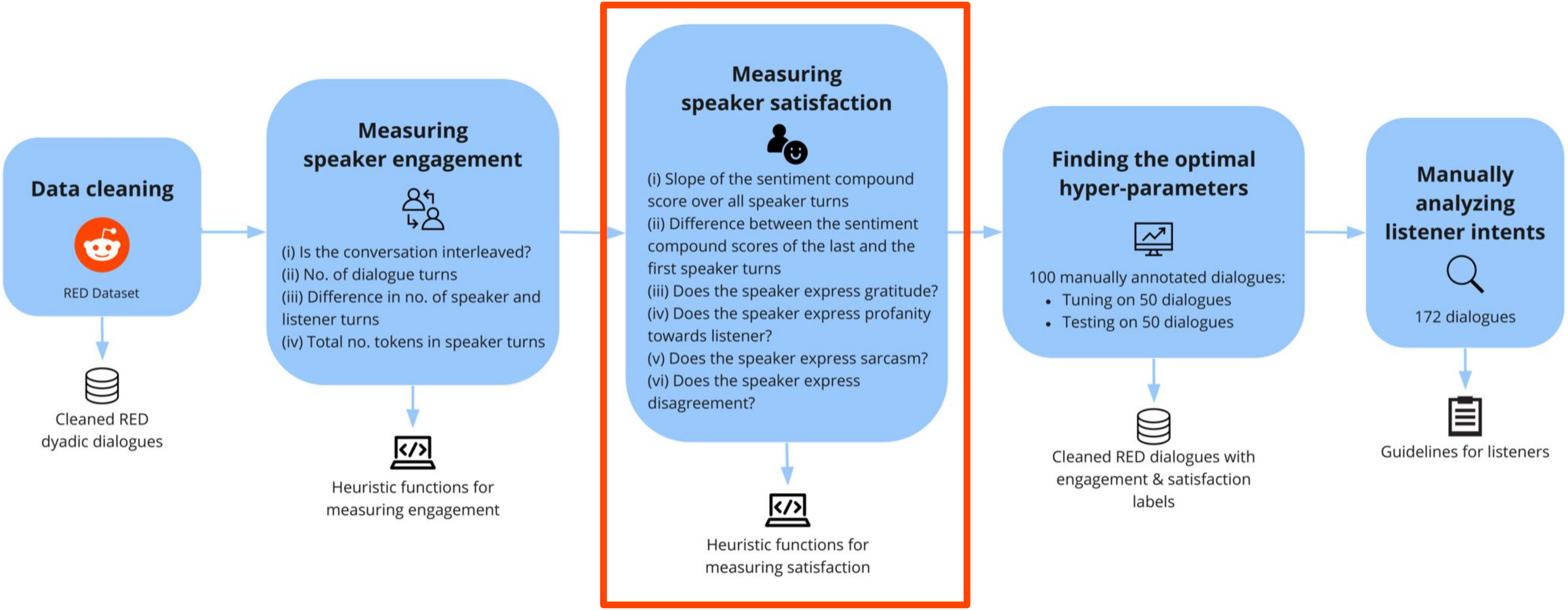
    iv. Selected the total number of tokens in the speaker turns.

    - ■ An upper limit of 30 tokens to limit the impact of very long responses.

Measuring Speaker Engagement

- Overall, we used four different predictors for predicting speaker engagement:
    i. Whether the conversation is interleaved (+)
    ii. The number of dialogue turns (+)
    iii. The absolute value of the difference in number of speaker and listener turns (-)
    iv. The total number of tokens in speaker turns (+)
- Assigned weights to each of the predictors.
- Defined a numerical threshold for the engagement score.

Methods



Data cleaning

RED Dataset

Cleaned RED dyadic dialogues

Measuring speaker engagement

(i) Is the conversation interleaved?
(ii) No. of dialogue turns
(iii) Difference in no. of speaker and listener turns
(iv) Total no. tokens in speaker turns

Heuristic functions for measuring engagement

Measuring speaker satisfaction

(i) Slope of the sentiment compound score over all speaker turns
(ii) Difference between the sentiment compound scores of the last and the first speaker turns
(iii) Does the speaker express gratitude?
(iv) Does the speaker express profanity towards listener?
(v) Does the speaker express sarcasm?
(vi) Does the speaker express disagreement?

Heuristic functions for measuring satisfaction

Finding the optimal hyper-parameters

100 manually annotated dialogues:

- Tuning on 50 dialogues
- Testing on 50 dialogues

Cleaned RED dialogues with engagement & satisfaction labels

Manually analyzing listener intents

172 dialogues

Guidelines for listeners

Measuring Speaker Satisfaction

- To measure speaker satisfaction, we applied various heuristic methods:

    i. Extended Yeh et al. (2020)'s work and applied sentiment analysis on the RED dataset on the sentence turn level using the VADER tool.
        - Assigned the sentiment with the strongest magnitude as the dialogue turn sentiment.
        - Calculated the slope of the sentiment throughout the conversation.
            - Captures the overall direction of the speaker's change in mood.

    ii. Calculated the change in sentiment from the first to the last speaker turn.
        - Captures a more fine-grained change in sentiment.

Measuring Speaker Satisfaction

iii. Two complementary methods of detecting expressions of gratitude:

- Checked if the last speaker turn was tagged with *grateful* emotion and *positive* sentiment.
  - Since EmoBERT is applied 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.
- Used the *matcher* module of the SpaCy library 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.

Measuring Speaker Satisfaction

iv. Checking if the speaker expresses profanity toward the listener:

- ■ We used the profanity-check library on all the speaker turns except the first one.
- ■ To differentiate between profanity toward else and toward listener, we utilized the *matcher* module of SpaCy.
    - Checked if the speaker response that contains profanity also contains the tokens “you” and/or “your”.

Measuring Speaker Satisfaction

v. Checking if the speaker expresses sarcasm:

- We used a Keras text classification model trained on the News Headlines for Dataset Sarcasm Detection (Misra and Arora, 2019), (Misra and Grover, 2021) and set the threshold for sarcasm prediction as 0.6 upon careful inspection.

vi. Checking if the speaker expresses disagreement:

- We used SpaCy again, and detected the existence of certain tokens and phrases that convey disagreement, e.g., “i don’t think so”, “diagree”, “no way”, which we obtained from an online English language teaching source, KSE Academy.

Measuring Speaker Satisfaction

- Overall, we used six 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 turns (+)

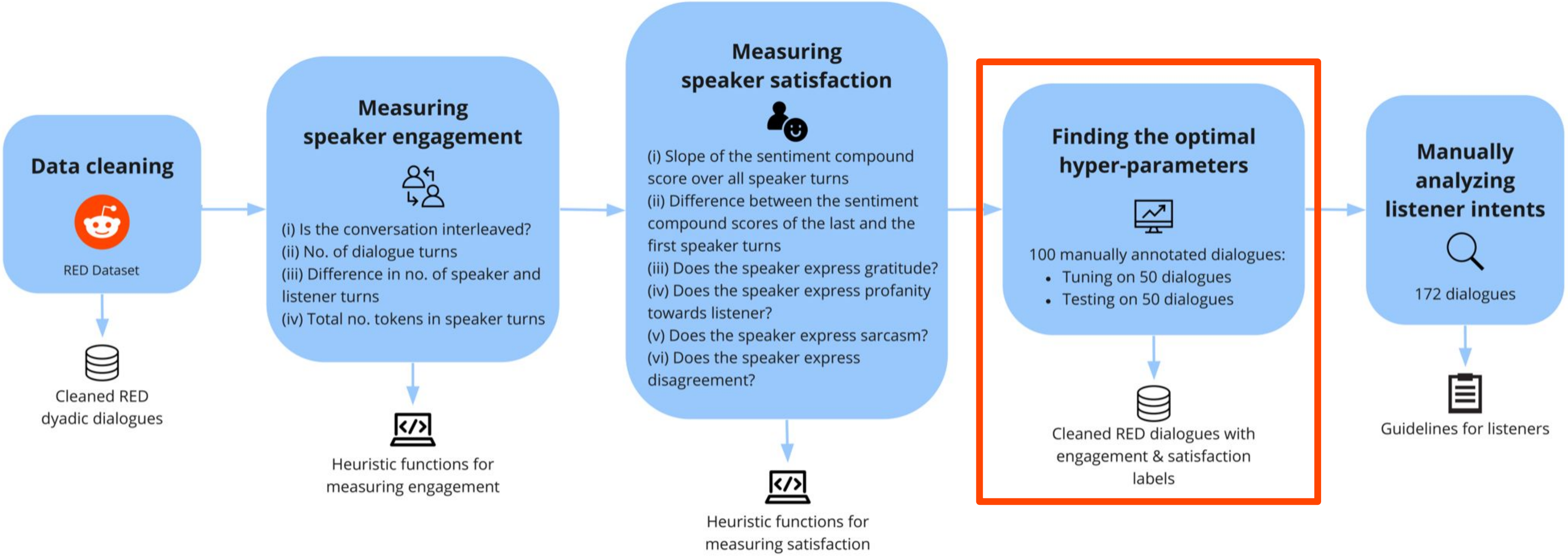
    iii. Whether the speaker expresses gratitude (+)

    iv. Whether the speaker uses profanity towards the listener (-)

    v. Whether the speaker uses sarcasm (-)

    vi. Whether the speaker expresses disagreement (-)
- Assigned weights to each of the predictors.
- Defined a numerical threshold for the satisfaction score.

Methods



Data cleaning

RED Dataset

Cleaned RED dyadic dialogues

Measuring speaker engagement

(i) Is the conversation interleaved?
(ii) No. of dialogue turns
(iii) Difference in no. of speaker and listener turns
(iv) Total no. tokens in speaker turns

Heuristic functions for measuring engagement

Measuring speaker satisfaction

(i) Slope of the sentiment compound score over all speaker turns
(ii) Difference between the sentiment compound scores of the last and the first speaker turns
(iii) Does the speaker express gratitude?
(iv) Does the speaker express profanity towards listener?
(v) Does the speaker express sarcasm?
(vi) Does the speaker express disagreement?

Heuristic functions for measuring satisfaction

Finding the optimal hyper-parameters

100 manually annotated dialogues:

- Tuning on 50 dialogues
- Testing on 50 dialogues

Cleaned RED dialogues with engagement & satisfaction labels

Manually analyzing listener intents

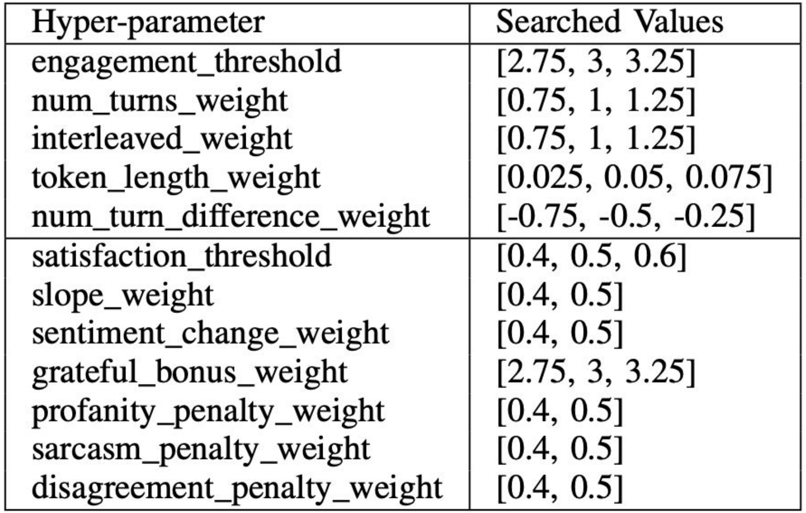
172 dialogues

Guidelines for listeners

Finding the Optimal Hyper-Parameters

- Annotated 100 dialogues with 12-13 samples from each of the 8 subreddits.
  - Ground truth labels for engagement and satisfaction
  - Created 50:50 validation and test sets
- Applied grid search on validation.
  - Selected optimal hyper-parameters wrt the best f1-score on validation.
- Applied best model onto the test set as well as the entire dataset.

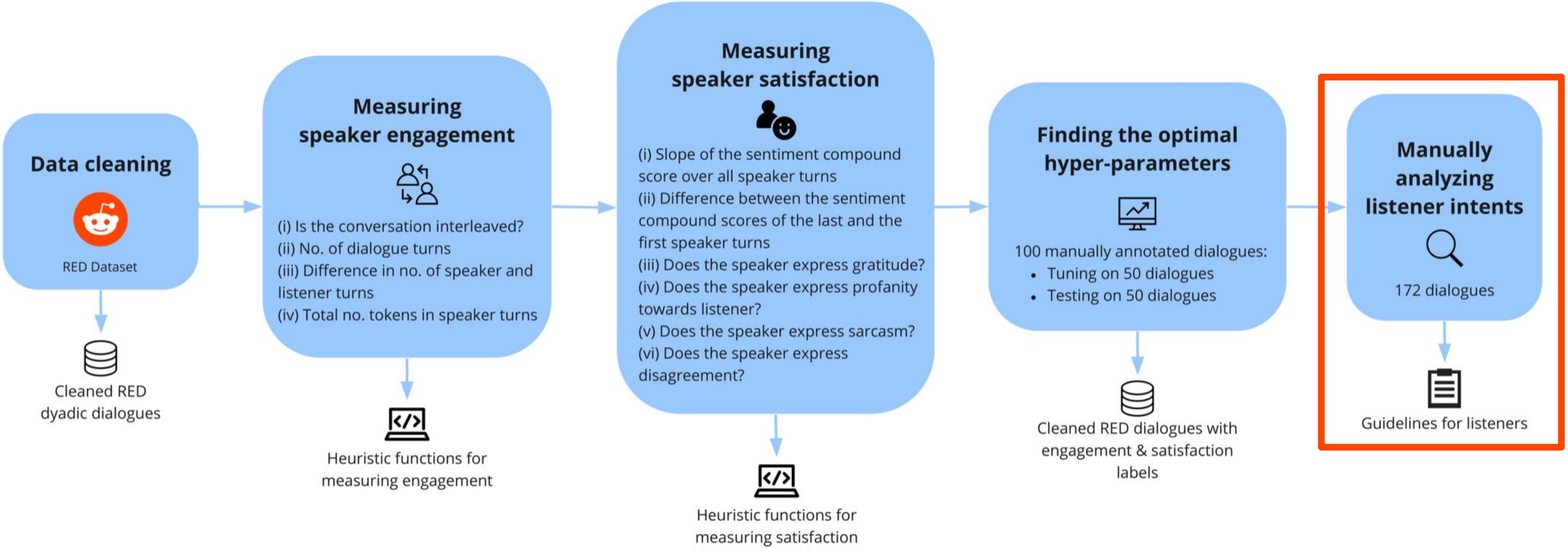
Finding the Optimal Hyper-Parameters



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.

Methods



Data cleaning

RED Dataset

Cleaned RED dyadic dialogues

Measuring speaker engagement

(i) Is the conversation interleaved?
(ii) No. of dialogue turns
(iii) Difference in no. of speaker and listener turns
(iv) Total no. tokens in speaker turns

Heuristic functions for measuring engagement

Measuring speaker satisfaction

(i) Slope of the sentiment compound score over all speaker turns
(ii) Difference between the sentiment compound scores of the last and the first speaker turns
(iii) Does the speaker express gratitude?
(iv) Does the speaker express profanity towards listener?
(v) Does the speaker express sarcasm?
(vi) Does the speaker express disagreement?

Heuristic functions for measuring satisfaction

Finding the optimal hyper-parameters

100 manually annotated dialogues:

- Tuning on 50 dialogues
- Testing on 50 dialogues

Cleaned RED dialogues with engagement & satisfaction labels

Manually analyzing listener intents

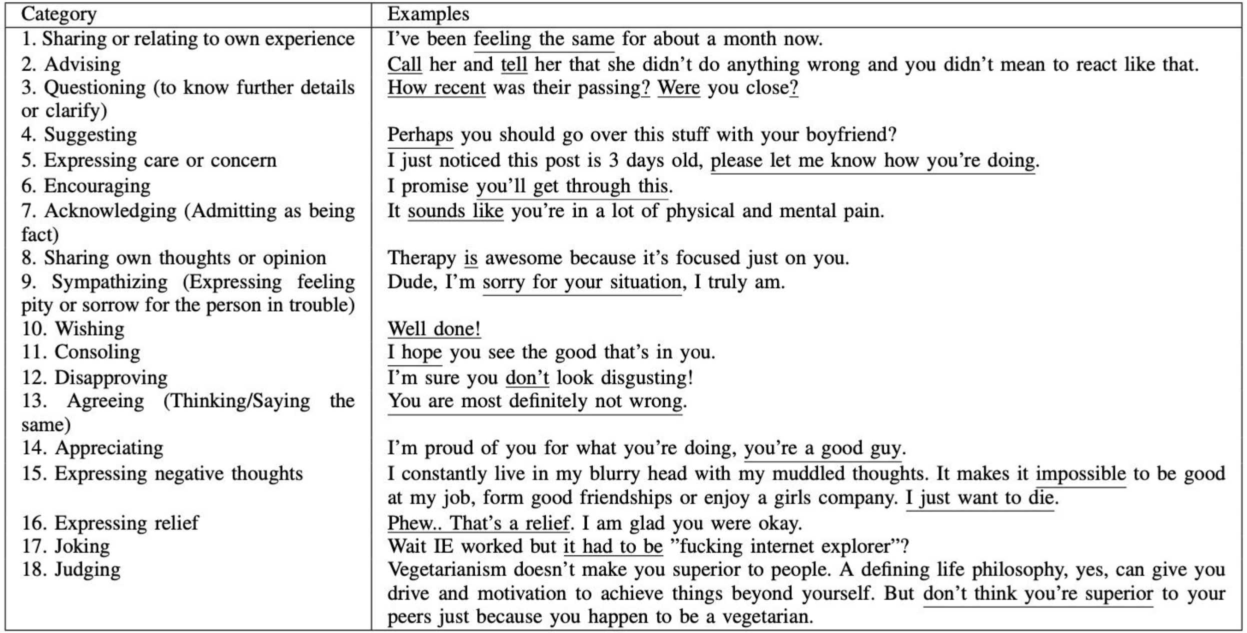
172 dialogues

Guidelines for listeners

Manually Analyzing Listener Intents

- Annotated 172 dialogues with their listener intents.
  - Using the taxonomy of empathetic response intents in Welivita and Pu (2020).
  - Added *judging*, *joking*, and *expressing negative thoughts* to account for the unempathetic listener intents.
- Manually analyzed the listener intents that lead to high/low satisfaction and high/low engagement.

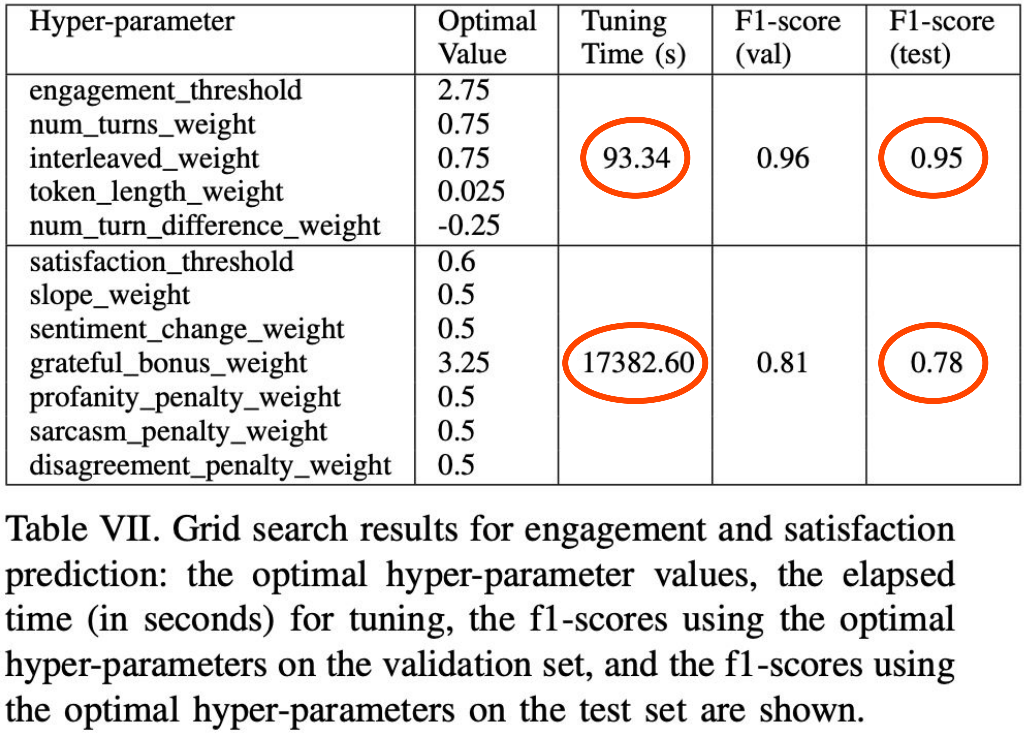
Manually Analyzing Listener Intents



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.

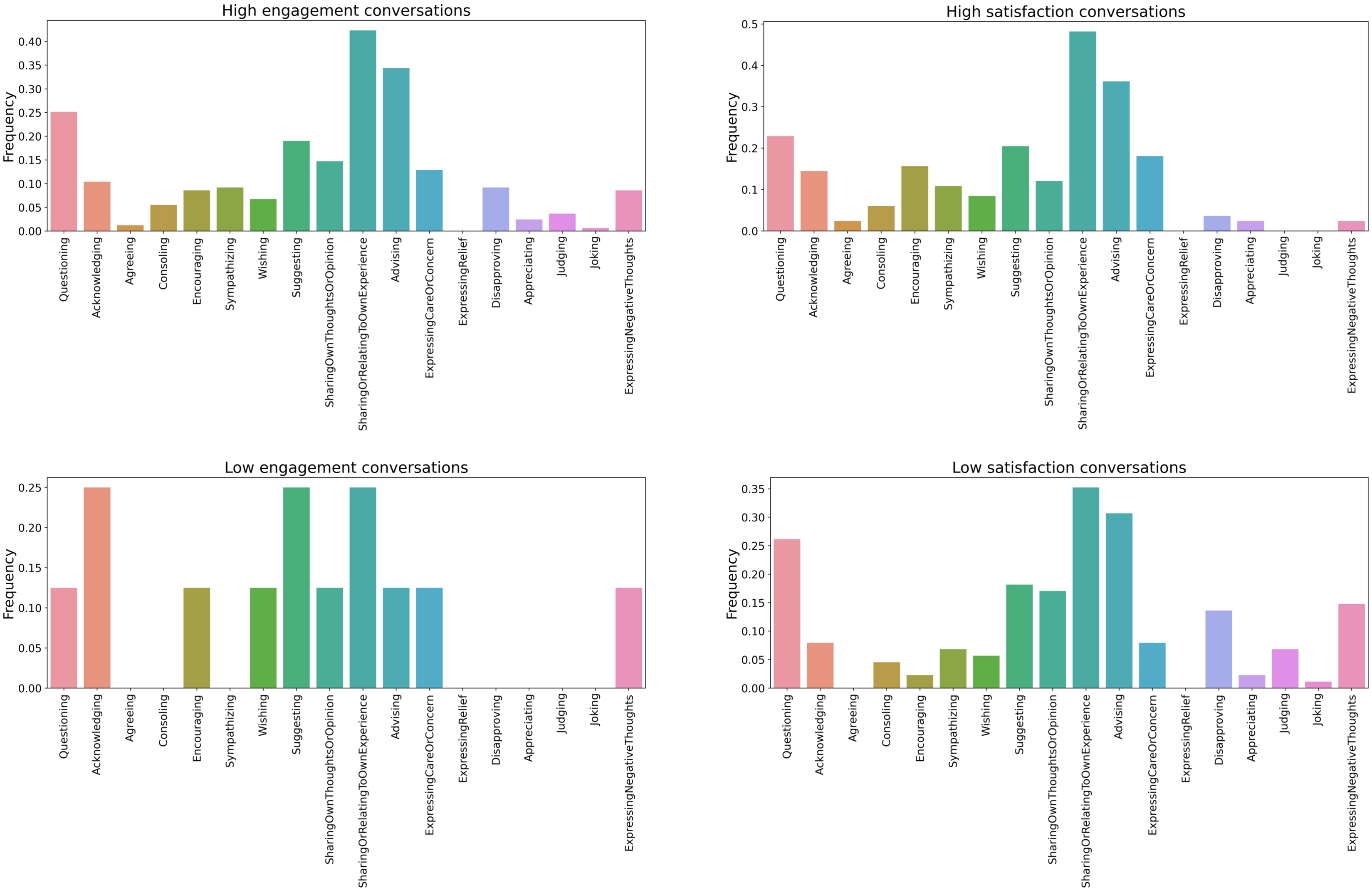
Table VIII. 18 listener intents and their examples. Note that 15 of these intents are taken from Welivita and Pu (2020) and 3 are newly introduced.

Results



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

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.


High engagement conversations
Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Questioning
Acknowledging
Agreeing
Consoling
Encouraging
Sympathizing
Wishing
Suggesting
SharingOwnThoughtsOrOpinion
SharingOrRelatingToOwnExperience
Advising
ExpressingCareOrConcern
ExpressingRelief
Disapproving
Appreciating
Judging
Joking
ExpressingNegativeThoughts
High satisfaction conversations
Frequency
0.0
0.1
0.2
0.3
0.4
0.5
Questioning
Acknowledging
Agreeing
Consoling
Encouraging
Sympathizing
Wishing
Suggesting
SharingOwnThoughtsOrOpinion
SharingOrRelatingToOwnExperience
Advising
ExpressingCareOrConcern
ExpressingRelief
Disapproving
Appreciating
Judging
Joking
ExpressingNegativeThoughts
Low engagement conversations
Frequency
0.00
0.05
0.10
0.15
0.20
0.25
Questioning
Acknowledging
Agreeing
Consoling
Encouraging
Sympathizing
Wishing
Suggesting
SharingOwnThoughtsOrOpinion
SharingOrRelatingToOwnExperience
Advising
ExpressingCareOrConcern
ExpressingRelief
Disapproving
Appreciating
Judging
Joking
ExpressingNegativeThoughts
Low satisfaction conversations
Frequency
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Questioning
Acknowledging
Agreeing
Consoling
Encouraging
Sympathizing
Wishing
Suggesting
SharingOwnThoughtsOrOpinion
SharingOrRelatingToOwnExperience
Advising
ExpressingCareOrConcern
ExpressingRelief
Disapproving
Appreciating
Judging
Joking
ExpressingNegativeThoughts

Discussion

- 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.
    - F1-scores on test set: 0.95 (engagement) and 0.78 (satisfaction)
- Listener intent analysis:
    - Some intuitive results, e.g., the correlation between expressing care or concern and higher satisfaction.
    - Some unintuitive results, e.g., the correlation between disapproving and higher user engagement.

Discussion

- Guidelines to be used in the development of chatbots from OMHC data:
  - Higher speaker satisfaction:
    - Sharing or relating to one’s own experience
    - Expressing care or concern
    - Acknowledging
    - Encouraging
  - Lower speaker satisfaction:
    - Questioning
    - Sharing own thoughts or opinion
    - Disapproving
    - Joking
    - Judging
    - Expressing negative thoughts

Discussion

- Higher speaker engagement:
  - Sharing or relating to one’s own experience
  - Advising
  - Questioning
  - Sharing own thoughts or opinion
  - Sympathizing
  - Consoling
  - Disapproving
  - Agreeing
  - Appreciating
  - Joking
  - Judging
- Lower speaker engagement:
  - Expressing care or concern
  - Encouraging
  - Acknowledging
  - Wishing
  - Expressing negative thoughts

Conclusion

- We introduced a new measure combining both speaker engagement and satisfaction to identify the optimal conversation patterns that can be used to train a chatbot to respond effectively to emotional distress.
    - F1-scores on test set: 0.95 (engagement) and 0.78 (satisfaction)

- We manually analyzed the frequency of listener intents that lead to high/low satisfaction and engagement.
    - Guidelines for the development of a chatbot:
        - For high speaker satisfaction:
            - Should be acknowledging, encouraging, share or relate to one's own experience, and express care or concern.
            - Should not express negative thoughts, or be disapproving or judging.
        - For high speaker engagement:
            - Should be asking more questions and offer advice.

Future Work

- Use a larger grid search to possibly find better hyper-parameters resulting in more accurate classifications on the RED dataset.
- Apply EmoBERT on the sentence level to improve the emotion and speaker satisfaction predictions.
- Use a separate dataset which contains conversations of length two for identifying conversational strategies that lead to lower speaker engagement.
- Incorporate multiparty conversations.