Improving Open-Domain Dialogue Evaluation with a Causal Inference Model

Cat P. Le, Luke Dai, Michael Johnston, Yang Liu, Marilyn Walker,

and Reza Ghanadan

The International Workshop on Spoken Dialogue Systems Technology 2023

amazon alexa

About me

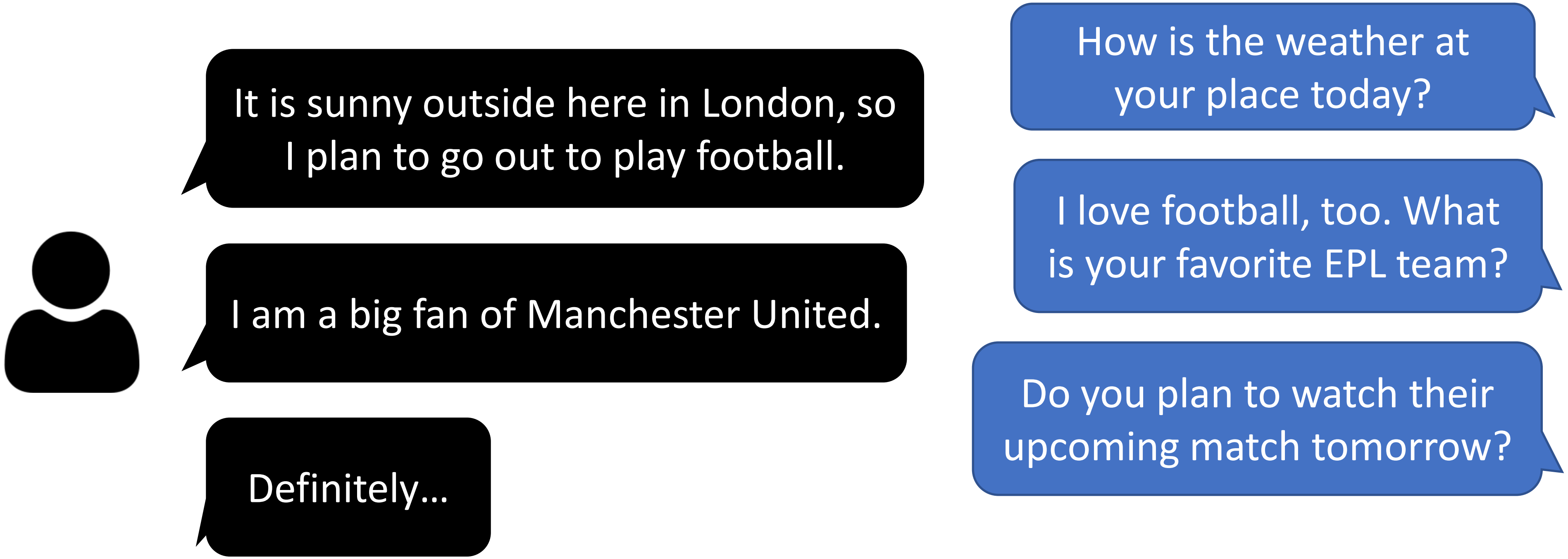
Duke UNIVERSITY

- Cat P. Le is a Ph.D. Candidate in machine learning at Duke University.
- His research interests: transfer learning, few-shot learning, multi-task learning, meta-learning.
- This work was done while he was a research scientist intern at Amazon Alexa AI.

amazon alexa

Open-Domain Dialogue System

- No specific goal but trying to be **engaging** and **interactive** with users.
- Generate coherent and meaningful responses.



Alexa Prize SocialBot

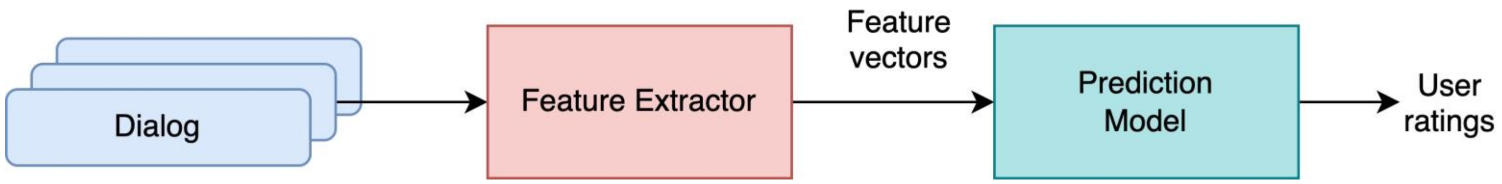
- The **SocialBot** is an Alexa skill that can chat coherently and engagingly with users on popular topics and genres with widely varying lengths.
- Dialogues are de-identified conversations from thousands of users.
- User are asked for the ***satisfaction ratings*** *(from 1-5 stars)* on the quality of SocialBot.
- However, these ratings are often biased or subjective.
- As a result, evaluation system for open-domain dialogue remains a challenging problem.

Evaluation Factors

- The ***end-to-end*** (E2E) evaluation approaches often perform well on expert-rated but not on user-rated data.
- Numerous factors can affect user *satisfaction* ratings.
  - Sentiment score is shown to be strongly correlated with user ratings.
  - Response relevance and specificity are also useful representations.
  - Triggered word (e.g., insults, complains) can be used as a quick indicator for bad dialogs.

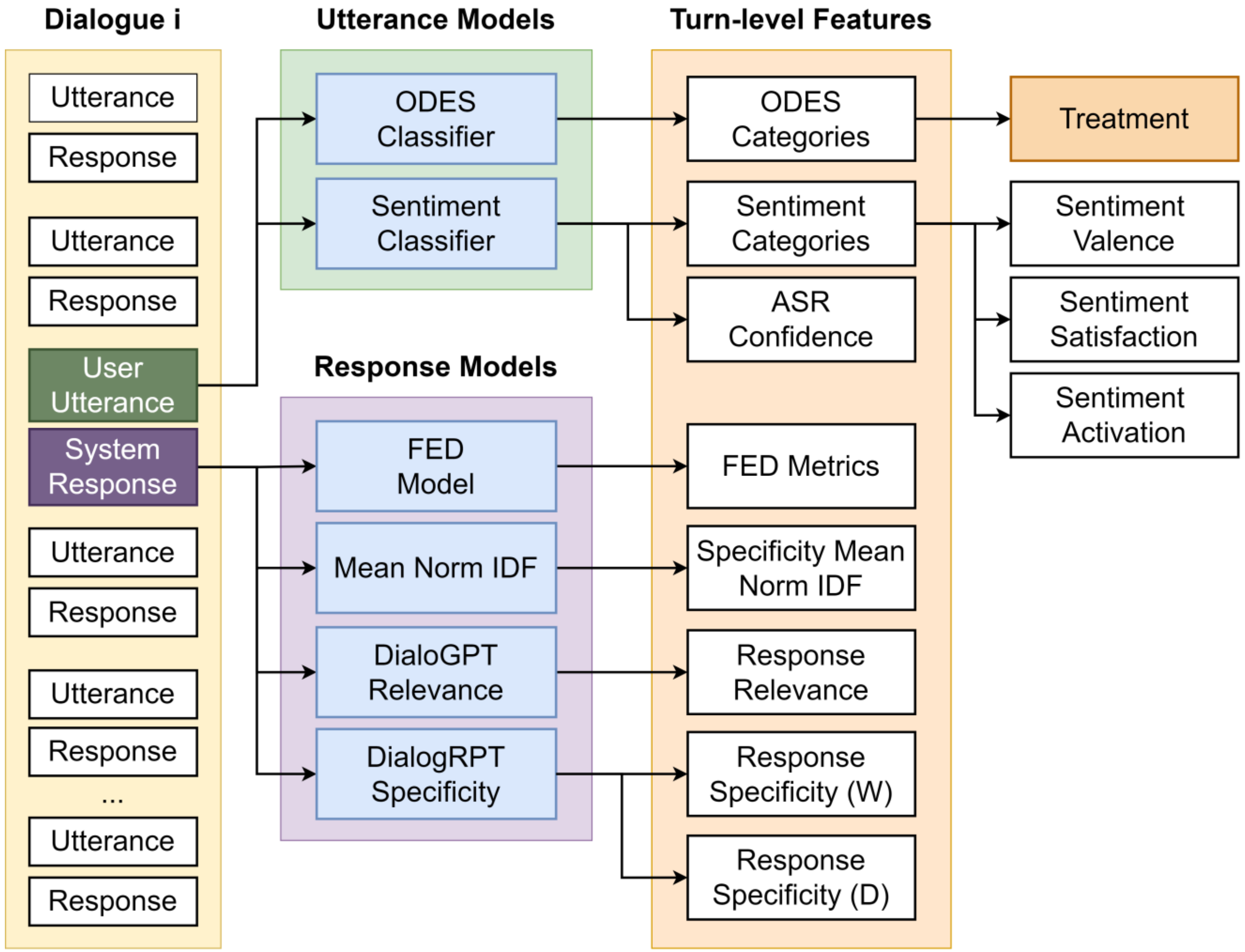
Proposed Method

- Our proposed method aims to predict the user ratings based on the meaningful features from the dialogue.
- Our framework consists of two components:
  - **Dialogue Turn-level Feature Extractors** convert each turn pair in the dialogue into a feature vector representations (e.g., sentiment, response relevance, specificity, text categories).
  - **Counterfactual LSTM (CF-LSTM)** utilizes the extracted dialogue features for learning to predict the user ratings.



amazon alexa

Dialogue Turn-level Feature Extractors



- Six models is utilized to extract features for each utterance-response pair.
- Two models for the utterances and four models for the responses.

amazon alexa

ODES Turn-level Feature

Class	ODES Name	Counts	Example
i	User disinterest	10,938	**SocialBot**: Did you know that otters sleep holding hands? **User**: I really couldn't care less.
ii	User critique	12,290	**SocialBot**: Who do you think will win Superbowl 2019? **User**: You are really a very stupid bot.
iii	User not understand	12,399	**SocialBot**: What is your favorite genre of video game? **User**: I don't know what genre means.
iv	User requests topic switch	21,278	**SocialBot**: Who is your favorite Batman actor? **User**: Can we talk about something else?
v	User obscenity	66,532	**SocialBot**: Do you have a favorite movie? **User**: B*** me.
vi	User rejects topic switch	4,389	**SocialBot**: I love country music too. But I was wondering, do you have a favorite sport? **User**: No, I want to keep talking about music.
vii	User requests to repeat	26,779	**SocialBot**: Would you rather climb the Eiffel Tower or the Empire State Building? **User**: Could you say that again?
viii	User requests to stop	78,504	**SocialBot**: I just love talking about music. What is your favorite kind of music? **User**: Please stop I need to go to bed.
ix	User insult	12,173	**SocialBot**: Wow, Taylor Swift has 97 albums. That's a lot! **User**: You are so full of sh**.
x	User compliment	57,539	**SocialBot**: Did you know LeBron James Jr has a college football scholarship, and he's only 12 years old? **User**: That's really interesting.
xi	User calls out repetition	15,052	**SocialBot**: What's your favourite football team? **User**: You already asked me that question twice.
xii	User calls out contradiction	3,147	**SocialBot**: I don't have any pets. **User**: You just said you had a cat.
xiii	System not understand	12,534	**User**: Can we talk about Elle King? **SocialBot**: I like BB King too. **User**: That's not what I said.
xiv	Others	2,320,515	**SocialBot**: Do you like K-Pop music? **User**: Yes, I often listen to Blackpink and BTS.

- Open-Dialogue Evaluation Signals Classifier (ODES) is proposed to categorize dialogues into the 14 classes.
- Subsequently, it is used to assign the treatment for causal analysis with CF-LSTM.

amazon alexa

Other Turn-level Features

- Sentiment analysis includes valence, satisfaction, and activation.
- Response relevance is based on DialoGPT.
- Response specificities (width and depth) are based on DialogRPT.
- Other features includes FED, ASR conf. score, and specificity mean norm IDF.

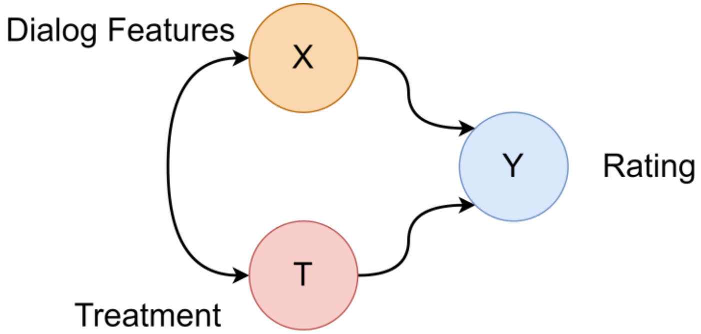
Turn-level to Dialogue-level Features

- The turn-level features are **stacked** into a dialogue-level feature vector for each dialogue.
- Since each dialogue has a different number of turn pairs, the dialogue-level features are varied-length vectors.
- Additionally, the treatment assignment is added for each dialogue-level feature.
- Particularly, treatment T = 1 is assigned for poorly-rated dialogues by the ODES classifier (e.g., dialogs where users complain, insult), and treatment T = 0 is assigned for the remaining dialogues.

amazon alexa

Causal Inference in Dialog Evaluation

- The ODES classifier is not perfect and can label dialogues incorrectly.
- Dialogue with the same extracted feature vectors might be totally different in terms of satisfaction ratings.
    - E.g., Sad sentiments might indicate the user's dissatisfaction or the heartbreaking content of the conversation.
- Causal analysis aims to study the **treatment effect** on open-domain dialogues.

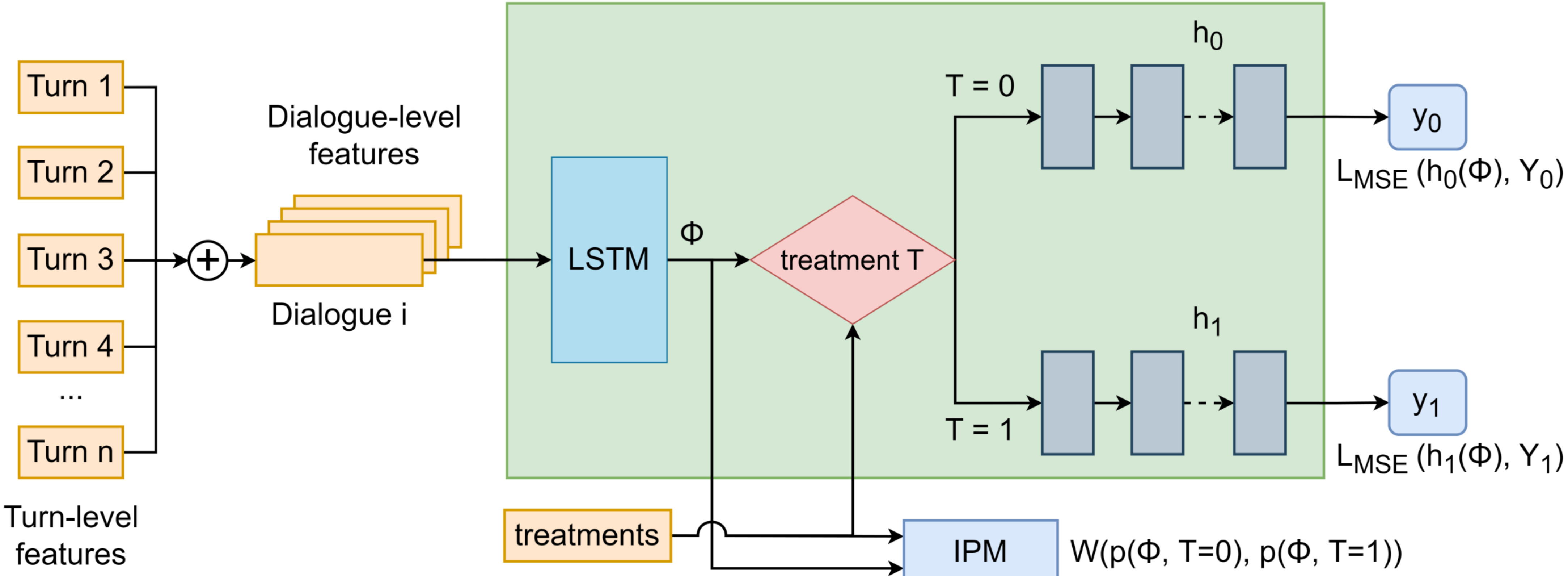


amazon alexa

Counterfactual-LSTM (CF-LSTM)

- CF-LSTM is a causal inference model that aims to investigate the **potential outcomes** (ratings) of the dialogs under different **hypotheses** (treatments).
- It maps the *dialogue-level features* to the *user ratings* based on a specific hypothesis based on the ODES classes.
- Its structure consists of the LSTM layers and 2 individual MLP regressors, each trained individually to predict ratings on a specific hypothesis.

Structure of CF-LSTM



amazon alexa

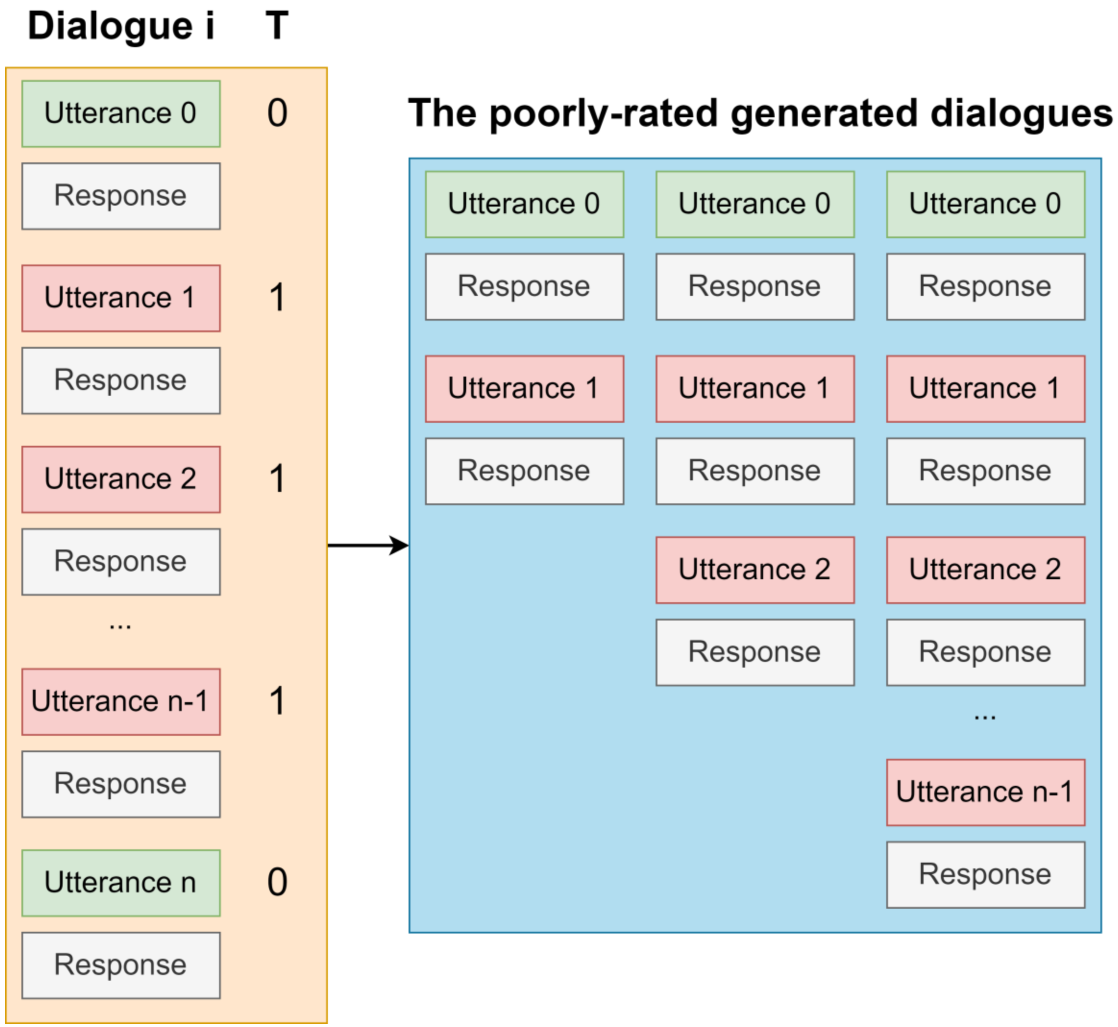
Loss Function

- The Integral probability metric (IPM) is used to avoid introducing variance and bias into the model.
- This metric measures the **Wasserstein distance** between two distributions $p(\phi, T = 0), p(\phi, T = 1)$.
- The loss function of the CF-LSTM is described as follows:

$$L = L_{MSE}(h_0(\phi), Y_0|T = 0) + L_{MSE}(h_1(\phi), Y_1|T = 1) \\ +\alpha W(p(\phi, T = 0), p(\phi, T = 1))$$

where α controls the *trade-off* between the similarity of the representations (ϕ) and the model's performance on the factual data.

Data Augmentation



- Data masking is used to generated poorly-rated dialogues.
- Expose the system with *"short"* poorly-rated dialogs.
- Help the system learn to quickly detect bad dialogues at *run-time*.

amazon alexa

Experiments

- Here, we compare our proposed method with the E2E Transformer, Dialogue-level MLP, and Dialog-level LSTM.
- The **regression** problems:
  - Individual rating prediction
  - Daily average rating prediction
  - 7-day average rating prediction
- The **classification** problems:
  - Binary classification: the label 0 is assigned for ratings with less than 3 stars (poorly-rated) and the label 1 if ratings are greater than or equal to 3 stars (highly-rated).
  - 5-class classification: the ratings are rounded half-up into 5 groups.

Prediction Performances

Table 2 The comparisons between open-dialogue evaluation methods for the regression problem in terms of Pearson correlation (e.g., individual, L1d, L7d predictions) and the classification problems (e.g., binary, 5-class) in terms of prediction accuracy, on SocialBot conversations

Methods	Individual Prediction	L1d* Prediction	L7d† Prediction	Binary Class.	5-class Class.
E2E Transformer	0.22	0.30	0.47	54.1%	32.8%
Dialogue-level LSTM	0.30	0.41	0.59	64.6%	43.5%
Dialogue-level MLP	0.31	0.40	0.66	62.5%	46.1%
CF-LSTM	**0.34**	**0.46**	**0.68**	**67.8%**	**48.2%**

* daily average prediction, † 7-day rolling average prediction

amazon alexa

Causal Analysis

- Consider the scenario in which the ODES classifier ***accidentally flips*** the treatment assignments.
- The model is robust where the predicted ratings show high correlations with the ground truth, even when treatment labels are incorrectly reversed.

Table 3 The correlations of CF-LSTM when the treatment assignments for all dialogues are inverted.

Pearson Correlation	Original Treatments	Inverted Treatments
Individual Prediction	0.34	0.26
L1D Prediction	0.46	0.35
L7D Prediction	0.68	0.52

* daily average, † 7-day rolling average

amazon alexa

Average Treatment Effect

- In Rubin causality, the common interest is to study the causal effect of the treatment assignment.
- The **average treatment effect (ATE)** is shown as follows:

$$ATE = E[Y_{1i} - Y_{0i}] = -0.7809$$

- The ATE indicates that, on average, dialogues with T = 1 (i.e., presumably bad dialogs) will have lower user ratings than dialogues with T = 0 by 0.7809.

amazon alexa

Conclusions

- CF-LSTM is robust in learning complex representations and can predict the ratings for dialogues under different hypotheses.
- It can be applied at run-time to identify low-quality dialogs and propose different topics to improve the user experience.

➢ In future work, increasing the number of treatments can help further improve the model's performance and flexibility.