



香港浸會大學  
HONG KONG BAPTIST UNIVERSITY



DEPARTMENT OF  
COMPUTER SCIENCE  
HONG KONG BAPTIST UNIVERSITY  
香港浸會大學計算機科學系

# Understanding User Feedback on Recommendations in Conversational Systems

**Dr. Li Chen**

[lichen@comp.hkbu.edu.hk](mailto:lichen@comp.hkbu.edu.hk)

**September 25, 2020**

**Invited talk for the 2nd International Workshop on Context-Aware  
Recommender Systems (CARS 2020), in conjunction with RecSys'20**

# Traditional Conversational Recommender Systems (CRS)

UKRAINIAN VILLAGE. TWO bedroom rehab garden apartment. Lr, Eurokitchen, hwfl, excellent security, forced air, lots of closets, laundry in building. Garage space included. Dogs OK. Available immediately. \$600/ mo. 312-489-1554. / ;

Phone: 312-489-1554	2-bedrooms	\$600	60622 (West Town Rucktown)
---------------------	------------	-------	----------------------------------

This apartment is OK,

bigger cheaper ni

This neighborhood cov


convenient conservative

FindMe (Burke et al., 1997)

QWIKSHOP.COM

» Digital Cameras

Shop for: Digital Cameras Computers Holidays



**Adjust your preferences to find the right camera for you**

Manufacturer	X	Canon	X
Optical Zoom	↓	7x	↑
Memory (MB)	↓	512	↑
Weight (Grams)	↓	780	↑
Resolution	↓	6.2 M Pixels	↑
Size	X	Large	X
Case	X	Magnesium	X
Price	↓	995	↑

**Product Found: Canon EOS 30**

6.3 Megapixel CMOS sensor  
7-point wide-area AF  
High-performance DIGIC processor  
100-1600 ISO speed range  
Compatible with all Canon EF lenses and EX Speedlites  
PictBridge, Canon Direct Print and Bubble Jet Direct compatible - no PC required

I've found the Camera I want!

No lets start again

**Explain:**

**1. Less Memory and Lower Resolution and Cheaper**

This Critique covers 153 other Digital Cameras

**Less Memory**  
Current Value: 512 MB  
Critique: Less Than  
Remaining: (0 to 256 MB)

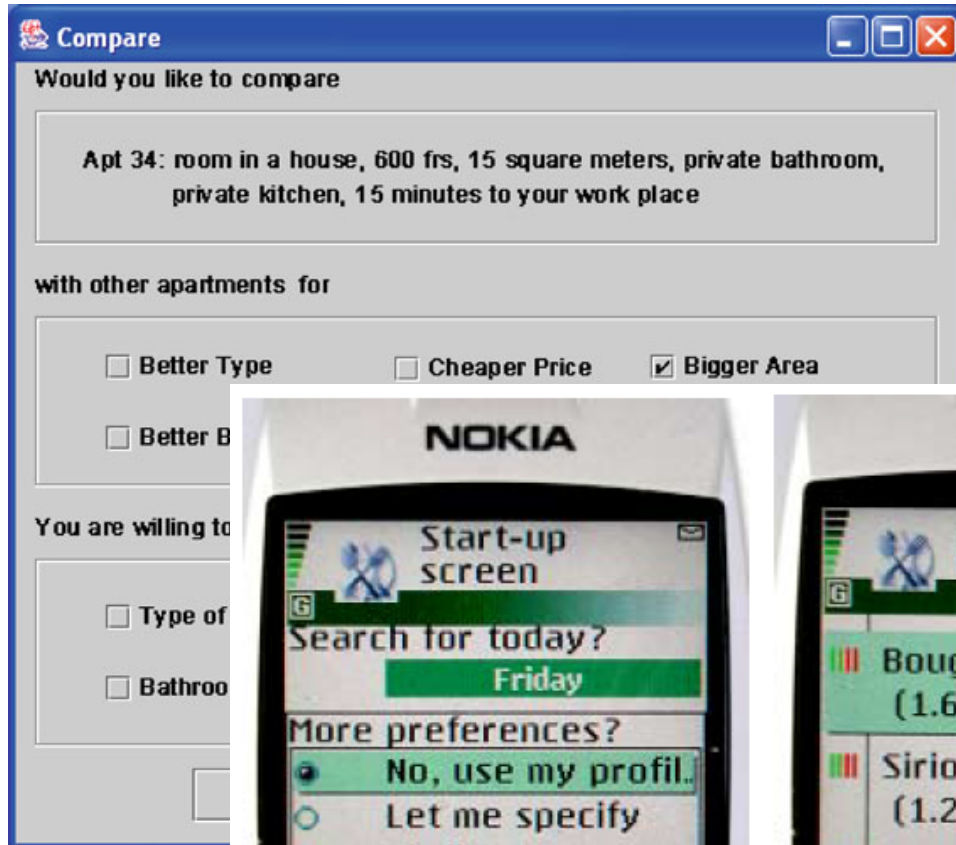
**Lower Resolution**  
Current Value: 6.2 M Pixels  
Critique: Less Than  
Remaining: (1.4 to 5.9 M Pixels)

**Cheaper**  
Current Value: 995 €  
Critique: Less Than  
Remaining: (75€ to 960€)

**We have more matching cameras with the following:**

1. Less Memory and Lower Resolution and Cheaper EXPLAIN PICK
2. Different Manufacturer and Less Zoom and Lighter EXPLAIN PICK
3. Lighter and Smaller and Different Case EXPLAIN PICK

Dynamic Critiquing (McCarthy et al., 2005)



Apartment



(a)



(b)

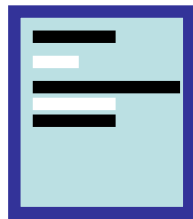


(c)

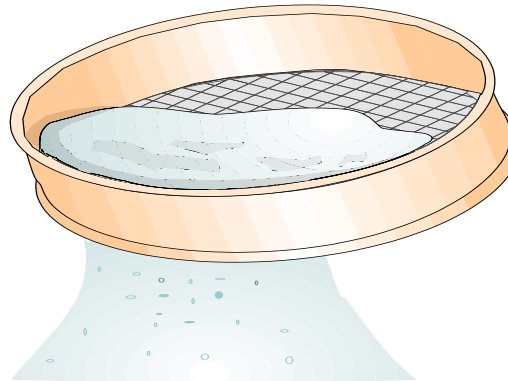
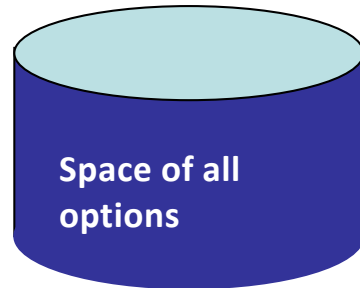
MobyRek with mobile critiquing (Ricci and Nguyen, 2005)

# Critiquing-based Recommender Systems

**Step 1: User states initial preferences**



Preference Model



$K$  items are displayed in the recommended set

**Conversational interaction**

- ✓ Feedback elicitation
- ✓ Preference refinement

**Step 2: System recommends multiple examples**

**Step 3: User revises preferences via critiquing**

**Step 4: User picks the final choice**




# Motivated by **Adaptive Decision Theory**

- Users are likely to construct their preferences in a **context-dependent and adaptive fashion** during the decision process (Payne et al., 1993; Carenini and Poole, 2000).
- Users **become aware of their latent preferences only when proposed solutions violate them** (Pu and Faltings, 2000 & 2002).
- Compensatory decision strategy (i.e., **tradeoff making**) normally leads to rational and high-quality decision (Frisch and Clemen, 1994)

Unfamiliar  
product  
domain



To find similar products with better values than this one



**Canon PowerShot S2 IS Digital Camera** [Add to saved list](#)  
**\$424.15**  
 Canon, 5.3 M pixels, 12x optical zoom, 16 MB memory, 1.8 in screen size, 2.97 in thickness, 404.7 g weight. [detail](#)

would you like to improve some values?

	Keep	Improve	Take any suggestion
Manufacturer	<input checked="" type="radio"/> Canon	<input type="radio"/> Sony <input type="button" value="v"/>	<input type="radio"/>
Price	<input type="radio"/> \$424.15	<input checked="" type="radio"/> less expensive <input type="button" value="v"/>	<input type="radio"/>
Resolution	<input checked="" type="radio"/> 5.3 M pixels	<input type="radio"/> less expensive <input type="button" value="v"/> <input type="radio"/> \$100 cheaper <input type="radio"/> \$200 cheaper <input type="radio"/> \$300 cheaper	<input type="radio"/>
Optical Zoom	<input checked="" type="radio"/> 12x	<input type="radio"/>	<input type="radio"/>
Removable Flash Memory	<input checked="" type="radio"/> 16 MB	<input type="radio"/> more memory <input type="button" value="v"/>	<input type="radio"/>
LCD Screen Size	<input checked="" type="radio"/> 1.8 in	<input type="radio"/> larger <input type="button" value="v"/>	<input type="radio"/>
Thickness	<input checked="" type="radio"/> 2.97 in	<input type="radio"/> thinner <input type="button" value="v"/>	<input type="radio"/>
Weight	<input checked="" type="radio"/> 404.7 g	<input type="radio"/> lighter <input type="button" value="v"/>	<input type="radio"/>

To find similar products with better values than this one



**Canon PowerShot S2 IS Digital Camera** [Add to saved list](#)  
**\$424.15**  
 Canon, 5.3 M pixels, 12x optical zoom, 16 MB memory, 1.8 in screen size, 2.97 in thickness, 404.7 g weight. [detail](#)

We have the following

1. Less Optical Zoom and Thinner and Lighter Weight
2. Different Manufacturer and Lower Resolution and Cheaper
3. Larger Screen Size and More Memory and Heavier

OR would you like to improve some value(s) by yourself?

	Keep	Improve	Take any suggestion
Manufacturer	<input checked="" type="radio"/> Canon	<input type="radio"/> Sony <input type="button" value="v"/>	<input type="radio"/>
Price	<input checked="" type="radio"/> \$424.15	<input type="radio"/> less expensive <input type="button" value="v"/>	<input type="radio"/>
Resolution	<input checked="" type="radio"/> 5.3 M pixels	<input type="radio"/> higher <input type="button" value="v"/>	<input type="radio"/>
Optical Zoom	<input checked="" type="radio"/> 12x	<input type="radio"/> more zoom <input type="button" value="v"/>	<input type="radio"/>
Removable Flash Memory	<input checked="" type="radio"/> 16 MB	<input type="radio"/> more memory <input type="button" value="v"/>	<input type="radio"/>
LCD Screen Size	<input checked="" type="radio"/> 1.8 in	<input type="radio"/> larger <input type="button" value="v"/>	<input type="radio"/>
Thickness	<input checked="" type="radio"/> 2.97 in	<input type="radio"/> thinner <input type="button" value="v"/>	<input type="radio"/>
Weight	<input checked="" type="radio"/> 404.7 g	<input type="radio"/> lighter <input type="button" value="v"/>	<input type="radio"/>

**User-initiated critiquing:** Unit or compound (Chen and Pu, AAAI'06)

**Hybrid critiquing:** User-initiated critiquing + system-suggested critiques (Chen and Pu, IUI'07)

- Critiquing-based system can significantly improve users' **decision accuracy** by **up to 57%**, against non-critiquing based
- Hybrid critiquing can achieve the **desired user control** and effectively save users' interaction effort

# Sentiment-based critiquing

The screenshot displays a product comparison interface with three tabs: 'Static View', 'Opinion View', and 'Mixture View'. The 'Mixture View' tab is selected and circled in red. Below the tabs, a 'Related Cameras' section is highlighted with a red dashed box. It contains two yellow critique boxes: 'They have better value at optical zoom and better opinions at effective pixels, weight, but worse value at price' and 'They have better value at weight and better opinions at video quality, optical zoom, but worse value at effective pixels'. Below these are four camera listings: Sony Cyber-shot DSC-HX200V, Nikon Coolpix S9100, Sony Cyber-shot DSC-HX100V, and Sony Cyber-shot DSC-HX5. Each listing includes a price, a star rating, and a list of features with their respective ratings. A red callout box points to the critique text, containing the text: 'Tradeoff-oriented category explanation based on both static specifications (e.g., "better value at optical zoom", "worse value at price") and sentiment features (e.g., "better opinions at effective pixels, weight")'. The interface also includes buttons for 'More Details' and 'Better Products' for each camera.

Static View    Opinion View    **Mixture View**

**Related Cameras**

They have better value at optical zoom and better opinions at effective pixels, weight, but worse value at price

They have better value at weight and better opinions at video quality, optical zoom, but worse value at effective pixels

**Sony Cyber-shot DSC-HX200V**  
Price: \$416.8    ★★★★★ 4.5 (397 reviews)  
Screen Size:    ★★★★★ 4.5 (397 reviews)  
Effective pixels:    ★★★★★ 4.5 (397 reviews)  
Optical zoom:    ★★★★★ 4.5 (397 reviews)  
Weight:    ★★★★★ 4.5 (397 reviews)  
Image quality:    ★★★★★ 3.9 (223 reviews)  
Video quality:    ★★★★★ 3.3 (124 reviews)  
Ease of use:    ★★★★★ 4.0 (175 reviews)  
[More Details](#)    [Better Products](#)

**Nikon Coolpix S9100**  
Price: \$379.0    ★★★★★ 3.3 (126 reviews)  
Image quality:    ★★★★★ 3.5 (228 reviews)  
Video quality:    ★★★★★ 3.3 (113 reviews)  
Ease of use:    ★★★★★ 3.6 (173 reviews)  
[More Details](#)    [Better Products](#)

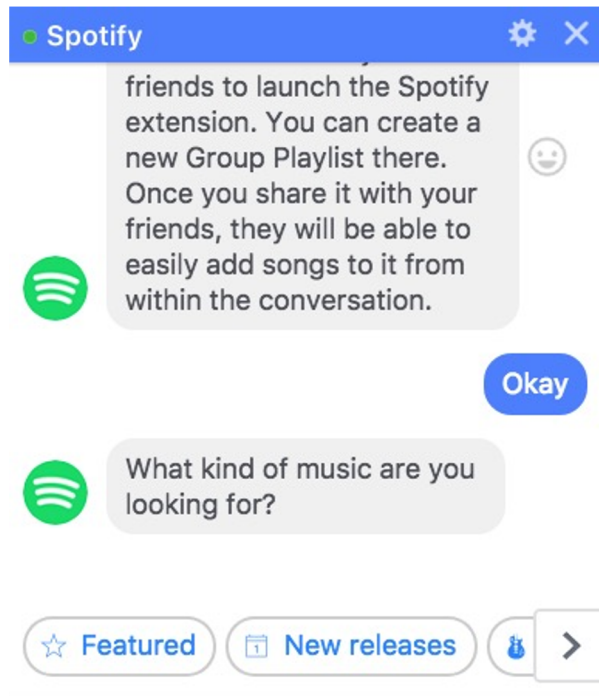
**Sony Cyber-shot DSC-HX100V**  
Price: \$420.0    ★★★★★ 3.9 (60 reviews)  
Screen Size: 3.0 inches    ★★★★★ 3.9 (31 reviews)  
Effective pixels: 16.0    ★★★★★ 3.6 (20 reviews)

**Sony Cyber-shot DSC-HX5**  
Price: \$400.0    ★★★★★ 3.4 (145 reviews)  
Screen Size: 3.0 inches    ★★★★★ 4.2 (45 reviews)  
Effective pixels: 10.0    ★★★★★ 3.3 (38 reviews)

Tradeoff-oriented category explanation based on both static specifications (e.g., "better value at optical zoom", "worse value at price") and sentiment features (e.g., "better opinions at effective pixels, weight")

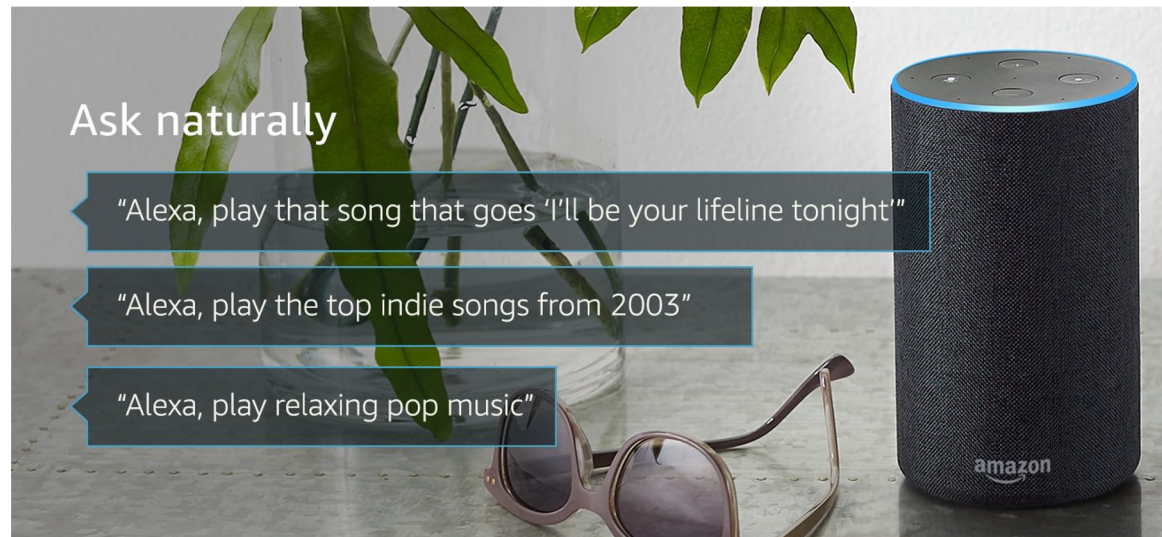
Incorporation of sentiment features into the critiquing interface can improve users' **product knowledge** and **preference certainty**

# Dialogue-based CRS (DCRS)



Composer is disabled for this thread.

<https://www.poptin.com/blog/how-to-use-chatbots-drive-sales-engagement/>



<https://www.amazon.co.uk/b?ie=UTF8&node=11368385031>



# Challenges

- **Dialogue-based CRS:** Users can freely express their preference in a way that they feel at ease
- *But,*
  - in such less controlled setting, how to elicit their feedback on recommendation?
  - can we accurately understand their intents behind utterances?
  - can we predict their satisfaction with recommendation?
- Little work has investigated these issues in a **multi-turn, mixed initiative** dialogue-based CRS

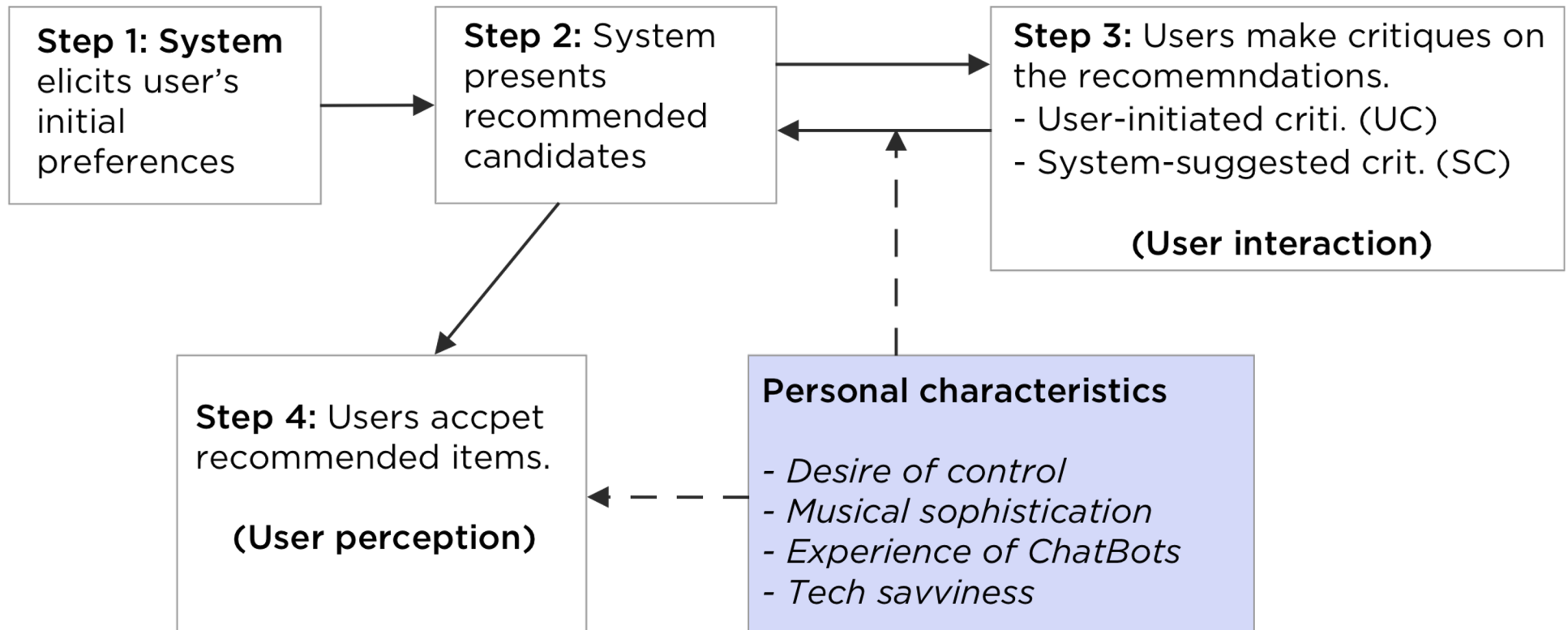
# Our Focuses

**Empirical study: User perception of and interaction with critiquing-integrated DCRS**

Classification of user intents for dialogue-based conversational recommendations

Prediction of user intents and satisfaction

# Critiquing-based interaction in dialogue system



# Interface design of our MusicBot

**(a)** Task: look for 5 good songs that fit the current scenario and your taste.

Crying Out For Me - Radio Edit  
 ★★★★★ Four Stars

**(b)** You would like to listen to on the trip.

I have found some songs for you based on your preference, but you can also search for other songs by using the tips shown on the right side.

We recommend this song because you like the songs of middle danceability, and the song Halo.

Crying Out For Me ...  
 Mario

I like this song.

Please rate your liked song.

Good, please try the next song.

Type of Way  
 Rich Homie Quan

Like Next Let bot suggest

Chat with me!

**(c)** Tips for tuning the recommendations by audio features

Currently the system supports searching by 5 audio features,

**Energy:** To tweak recommendation by energy, you can say "I need more/less energy", "I need higher/lower energy".

**Danceability:** To tweak recommendation by danceability, you can say "I need higher/lower danceability", "I need to dance", "Play a song for dancing".

**Speechiness:** To tweak recommendation by speechiness, you can say "I need more/less speech", "Play a song with less speech".

**Tempo:** To tweak recommendation by tempo, you can say "I like slow/fast songs", "Play some fast music".

**valence:** To tweak recommendation by tempo, you can say "I feel happy", "feel sad".

► Tips for tuning the recommendations by music categories

► Tips for tuning the recommendations by music languages

► Tips for tuning the recommendations by artists

**(d)** User-initiated critiquing

I like fast songs

Sure

OK, I recommend this song to you, because you like the fast songs

Crying Out For ...  
 Mario

Like Next Let bot suggest

System-suggested critiquing

I need some suggestions

Based on your music preference, we think you might like the pop English songs with lower danceability?

Yes No

# User Experiment



- Participants: 45 valid (19 female)
- User initiated (UC) critiquing vs. Hybrid critiquing (UC + SC)
- Experimental task

Find 5 ❤️ songs  
in two scenarios  
and give ratings



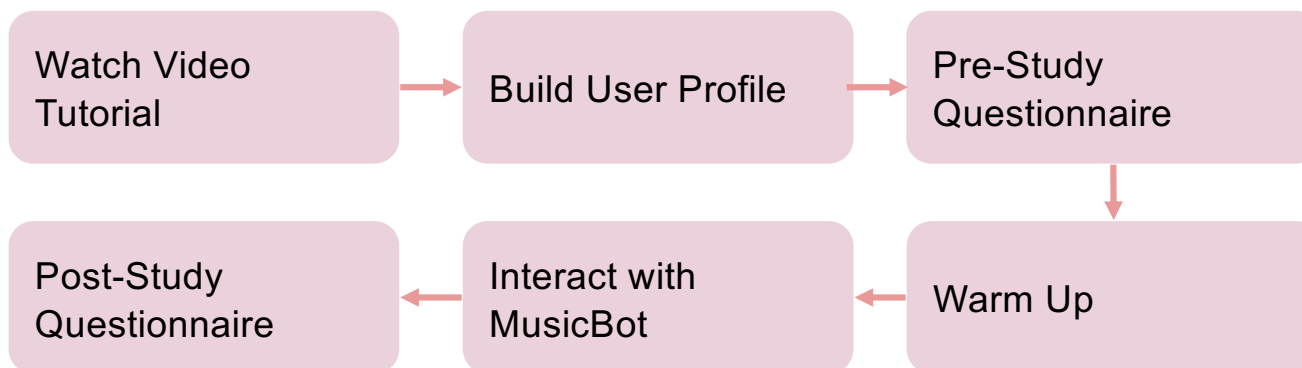
subway

UC



party

HC



# Measurements

## Question items

- Q1: The items recommended to me matched my interests.  
Q2: I easily found the songs I was looking for.  
Q3: Looking for a song using this interface required too much effort (reverse scale).  
Q4: The songs recommended to me are diverse.  
Q5: I found it easy to inform the system if I dislike/like the recommended song.  
Q6: I felt in control of modifying my taste using *MusicBot*.  
Q7: I am confident I will like the songs recommended to me.  
Q8: I like to give feedback on the music I am listening.  
Q9: This music chatbot can be trusted.  
Q10: I found the system easy to understand in this conversation.  
Q11: In this conversation, I knew what I could say or do at each point of the dialog.  
Q12: The system worked in the way I expected in this conversation.  
Q13: I will use this music chatbot again.  
Q14: Overall, I am satisfied with the chatbot.

- Rating (stars) for the selected songs
- Completion time
- Dialog turns
- Listened songs
- Button clicks
- Messages by typing
- Messages by voice
- Words per utterances
- Unknown utterances

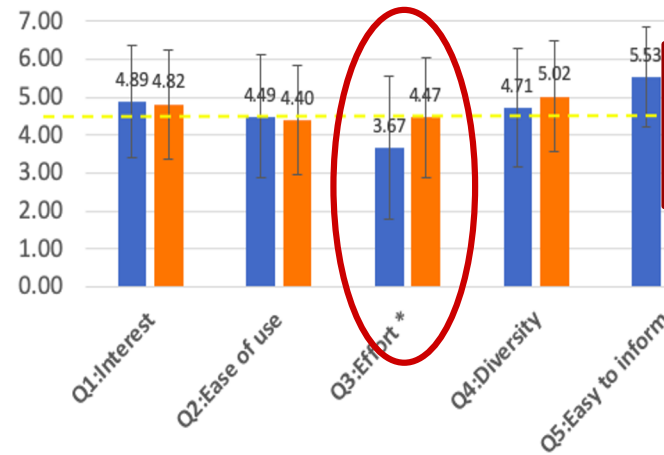
**ResQue: User-centric evaluation framework for recommender systems**  
(Pu et al., 2011)

**PARADISE: Evaluation framework for spoken dialogue agents**  
(Walker et al., 1997)

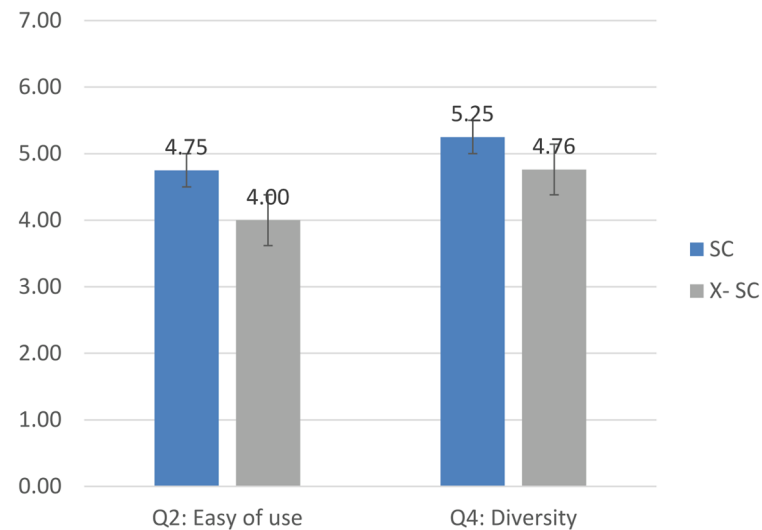
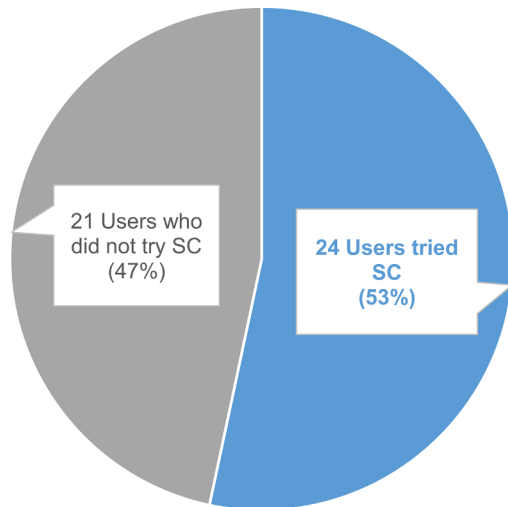
**Objective behavioral variables**

## UC: User-initiated Critiquing

## HC: Hybrid Critiquing (UC + SC)



Interaction metrics	UC (mean,sd)	HC (mean,sd)
Rating (stars)	(4.05, 0.47)	(4.08, 0.44)
Completion time* (minutes)	(5.40, 4.19)	(6.98, 4.16)
#Listened songs**	(10.67, 4.99)	(13.13, 6.09)
#Turns(times)**	(12.29, 8.21)	(16.11, 9.35)
#Btn(times)***	(9.18, 3.38)	(12.64, 7.07)
#Typing(times)	(3.09, 4.78)	(3.07, 4.21)
#Voice(times)	(1.24, 7.90)	(0.71, 2.97)
#Words	(2.13, 1.92)	(2.28, 1.84)
#Unknown utterances	(1.78, 6.46)	(0.78, 1.80)



Users who tried SC tend to **perceive higher ease of use and diversity.**

## Effect of personal characteristics on user perceptions

PC	Q1:Interest	Q2:Ease of use	Q3:Effort	Q4:Diversity	Q5:Easy to inform	Q6:Control	Q7:Confidence
CE	0.15 (0.33)	0.14 (0.37)	0.07 (0.66)	0.03 (0.84)	-0.03 (0.86)	0.11 (0.46)	0.05 (0.73)
TS	-0.01 (0.98)	-0.13 (0.40)	<b>0.36 (0.02)*</b>	0.10 (0.51)	-0.08 (0.59)	-0.19 (0.21)	-0.12 (0.43)
MS	<b>0.40 (0.01)*</b>	0.25 (0.10)	-0.22 (0.14)	0.17 (0.26)	0.10 (0.53)	<b>0.31 (0.04)*</b>	0.29 (0.05)
DFC	0.23 (0.14)	0.03 (0.84)	0.13 (0.41)	0.24 (0.11)	0.22 (0.15)	<b>0.35 (0.02)*</b>	0.25 (0.10)

PC	Q8:Feedback	Q9:Trust	Q10:Understand	Q11:Difficulty	Q12:Expected	Q13:Intent to reuse	Q14:Satisfaction
CE	0.06 (0.70)	-0.01 (1.00)	-0.07 (0.65)	0.02 (0.88)	0.06 (0.69)	0.21 (0.17)	0.10 (0.52)
TS	0.16 (0.29)	0.07 (0.66)	-0.12 (0.42)	-0.04 (0.77)	0.04 (0.78)	-0.12 (0.42)	-0.19 (0.10)
MS	<b>0.55 (&lt;0.001)**</b>	<b>0.37 (0.01)*</b>	0.09 (0.57)	0.13 (0.38)	0.23 (0.14)	<b>0.31 (0.04)*</b>	0.22 (0.15)
DFC	0.06 (0.68)	0.16 (0.29)	<b>0.30 (0.04)*</b>	<b>0.38 (0.01)*</b>	0.22 (0.14)	0.28 (0.06)	0.20 (0.19)

**Music Sophistication (+):** Interest matching, Control, Trust, Intention to Give Feedback and Reuse

**Desire For Control (+):** Control, Easy to Understand and Use



# Summary

- Combining UC and SC in a conversational user interface may **increase user engagement** and likelihood of **finding more (diverse) songs**.
- Designers should **consider MS and DFC** as key personal characteristics in interaction design for critiquing-based music recommenders.
- ***Limitations***
  - Small-scale user data
  - Not “smart” enough to understand user intentions

# Our Focuses

Empirical study: User perception of and interaction with critiquing support in DCRS

Classification of user intents for dialogue-based conversational recommendations

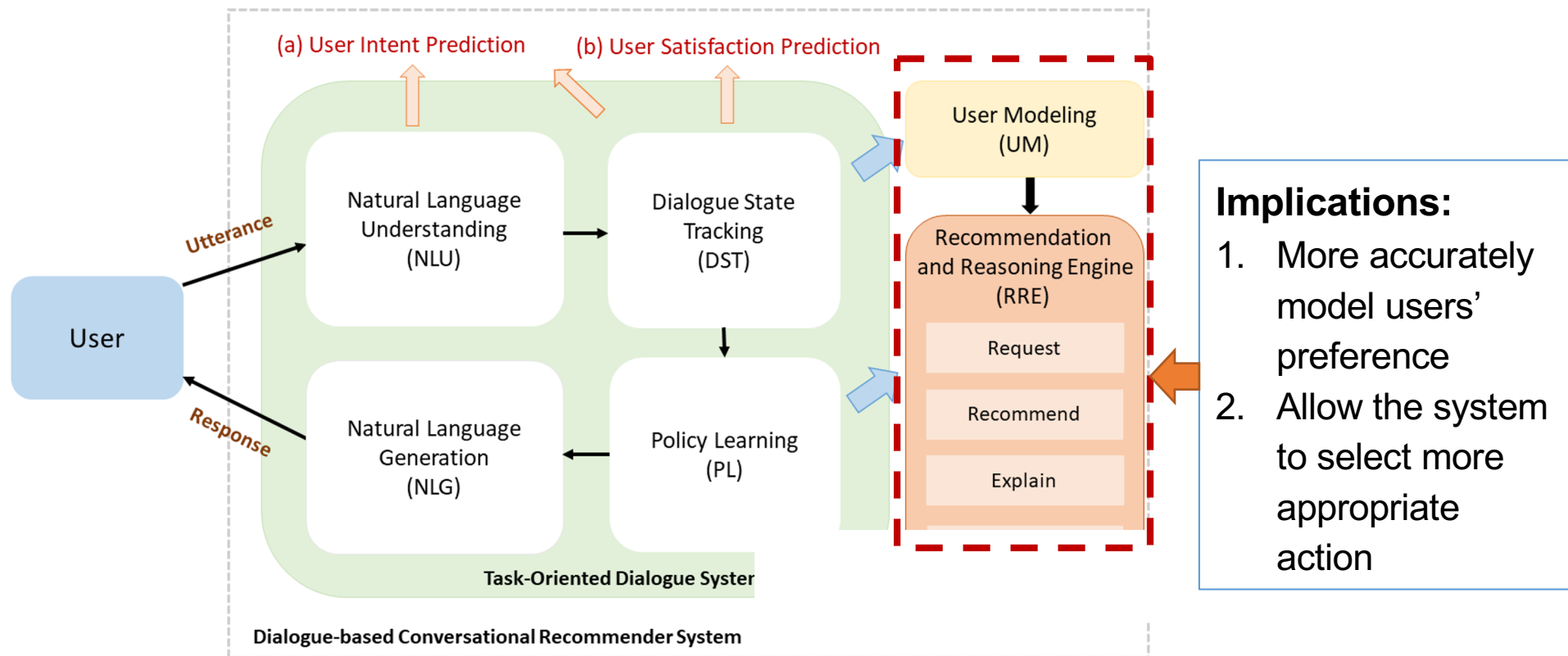
Prediction of user intents and satisfaction

Wanling Cai and Li Chen. Predicting User Intents and Satisfaction with Dialogue-based Conversational Recommendations. In *Proceedings of 28th Conference on User Modeling, Adaptation and Personalization (UMAP'20)*, pages 33–42, July 14-17, 2020. [Best Student Paper Award]

# User Intent and Satisfaction Prediction

**User intent** indicates the **goal** or **intention** that users have during their interaction with the system (Rose and Levinson, WWW 2004)

**User satisfaction** indicates **if the user's goal is fulfilled** to some extent (Hashemi *et al.*, CIKM 2018)



# Recommendation Dialogue Data

**Recommender:** Hi how are you today? I heard you might be interested in a movie. Any particular genre?

**Seeker:** Hi, I'm good, just looking for a nice horror movie. Nothing too gory, I liked Beetlejuice.

**Recommender:** hmm. I don't know too many horror movies. I did watch The Birds.

**Seeker:** Yeah I've seen the birds it was okay but I felt like it was too old for my tastes.

**Recommender:** border line with suspense might be something like Hannibal or The Silence of the Lambs.

**Seeker:** I didn't like any of those movies, too much talking.

**Recommender:** okay. Well, how about Saw ?

**Seeker:** Something more like Final Destination.

**Recommender:** Do you like any other genres?

**Seeker:** The Saw was okay, I felt like it was too violent. I really love like fantasy horror, maybe Ghost.

**Recommender:** I've heard that is a good one. Have you seen Signs ?

**Seeker:** I heard about that but didn't watch it.

**Recommender:** Mel Gibson in it. I've heard it is excellent.

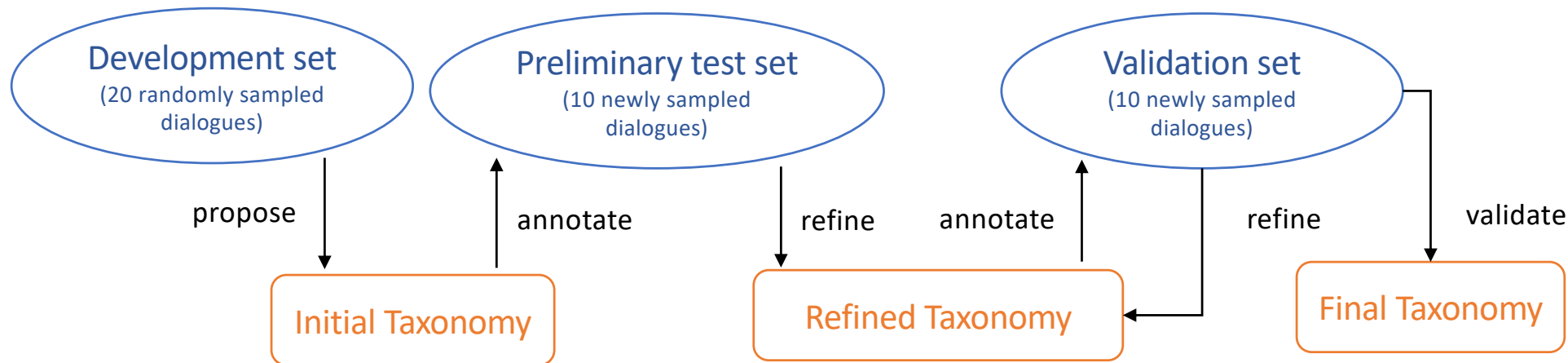
**Seeker:** okay, great I will check it out. thank you.

**ReDial Dataset**  
**human-human dialogues**  
 centered around movie  
 recommendations  
 (Li *et al.*, NIPS 2018)

ReDial dataset: <https://redialdata.github.io/website/>

## Statistics of our selected dialogue data

Items	SAT-Dial (with user-satisfied recommendation)	unSAT-Dial (without user-satisfied recommendation)
# Conversations	253	83
# Human seekers	125 (# utterances: 1,711)	59 (# utterances: 550)
# Human recommenders	151 (# utterances: 1,747)	68 (# utterances: 575)
# Suggested movies per dialogue	4.57	4.51
# Turns per dialogue	mean=6.58, min=3, max=19	mean=6.49, min=3, max=12
# Words per utterance	mean=11.29, min=1, max=72	mean=10.72, min=1, max=69



Intent (Code)	Description	Percentage
<b>Ask for Recommendation</b>		<b>18.26%</b>
Initial Query (IQU)	Seeker asks for a recommendation in the first query.	12.91%
Continue (CON)	Seeker asks for more recommendations in the subsequent query.	3.10 %
Reformulate (REF)	Seeker restates her/his query with or without clarification/further constraints.	1.50%
Start Over (STO)	Seeker starts a new query to ask for recommendations.	0.84%
<b>Add Details</b>		<b>18.58%</b>
Provide Preference (PRO)	Seeker provides specific preference for the item s/he is looking for.	12.30%
Answer (ANS)	Seeker answers the question issued by the recommender.	4.91%
Ask Opinion (ASK)	Seeker asks the recommender's personal opinions.	2.39%
<b>Give Feedback</b>		<b>61.92%</b>
Seen (SEE)	Seeker has seen the recommended item before.	21.14%
Accept (ACC)	Seeker likes the recommended item.	18.89%
Reject (REJ)	Seeker dislikes the recommended item.	11.50%
Inquire (INQ)	Seeker wants to know more about the recommended item.	6.55%
→ Critique-Feature (CRI-F)	Seeker makes critiques on specific features of the current recommendation.	6.50%
→ Critique-Add (CRI-A)	Seeker adds further constraints on top of the current recommendation.	5.35%
Neutral Response (NRE)	Seeker does not indicate her/his preference for the current recommendation.	4.29%
→ Critique-Compare (CRI-C)	Seeker requests sth similar to the current recommendation in order to compare.	1.55%
<b>Others</b>	Greetings, gratitude expression, or chit-chat utterances.	14.55%

# User Intent Prediction

- Multi-label Classification Problem

*E.g., “I did see that one, but I didn’t really like it. I do love 80s movies though.” -> two intents: Reject and Critique-Add*

- *Classification Models*

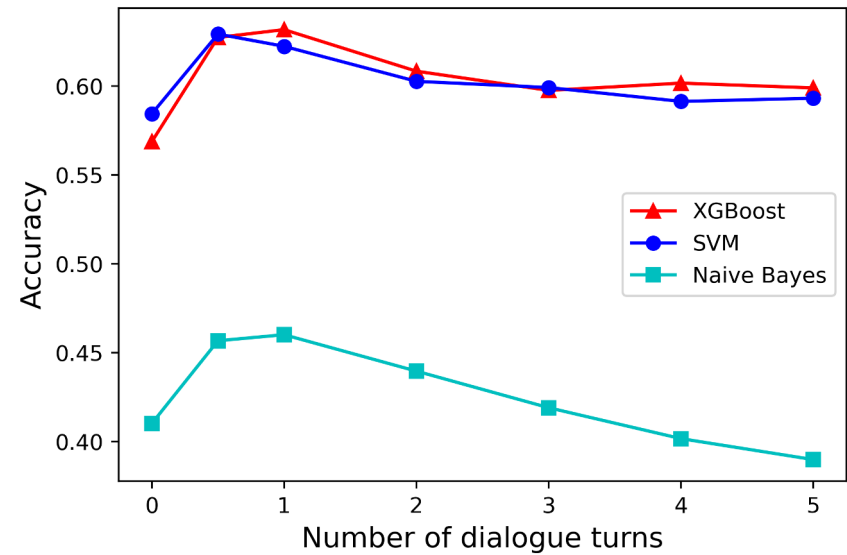
- 8 Machine Learning Models (e.g., LR, SVM, Naive Bayes, XGBoost, MLP, etc.) and 2 Deep Learning Models (CNN and Bi-LSTM)

- Features

<i>Category</i>	<i>Features</i>
<b>Content</b>	TF-IDF, Name Entity, # Relevant Items
<b>Discourse</b>	POS, 5W1H Question, Question Mark, Exclamation Mark, Utterance Length
<b>Sentiment</b>	Thanks, Sentiment Score, Opinion Lexicon
<b>Context</b>	Absolute Position, Utterance Similarity, Previous user intents & recommendation actions

Methods	Binary Relevance				Classification Chain				Label Powerset			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
Logistic Regression	0.5796	0.7160	0.6148	0.6612	0.6111	0.6898	0.6322	0.6596	0.6198	0.6791	0.6053	0.6400
SVM	0.5597	0.6701	0.6047	0.6332	<b>0.6293</b>	0.7179	0.6340	0.6730	0.6048	0.6004	0.6123	0.6056
Naive Bayes	0.4438	0.5137	0.5705	0.5400	0.4567	0.5137	0.5793	0.5439	0.5365	0.5989	0.5542	0.5755
Decision Tree	0.5264	0.5187	0.6778	0.5871	0.5356	0.5513	0.6325	0.5887	0.4515	0.4706	0.4755	0.4729
Random Forest	0.5742	0.5962	<b>0.7029</b>	0.6449	0.5968	0.6372	<b>0.6817</b>	0.6583	0.4794	0.4748	0.5096	0.4913
XGBoost	<b>0.5970</b>	<b>0.8169</b>	0.6007	<b>0.6919</b>	0.6274	<b>0.7957</b>	0.6268	<b>0.7010</b>	<b>0.6199</b>	<b>0.6868</b>	<b>0.6109</b>	<b>0.6463</b>
MLP	0.4773	0.7922	0.4743	0.5928	0.5079	0.7780	0.5045	0.6115	0.6157	0.6837	0.6029	0.6407

	Cont	Disc	Sent	Context	Acc	Prec	Rec	F1
1 Category	✓				<b>0.4726</b>	<b>0.7165</b>	<b>0.4868</b>	<b>0.5793</b>
		✓			0.3918	0.5224	0.3841	0.4426
			✓		0.3407	0.5020	0.3343	0.4011
				✓	0.1993	0.3241	0.2044	0.2498
2 Categories	✓			✓	<b>0.5603</b>	<b>0.7669</b>	<b>0.5627</b>	<b>0.6488</b>
		✓		✓	0.5438	0.6946	0.5346	0.6039
	✓	✓			0.5291	0.7381	0.5350	0.6201
	✓		✓		0.4921	0.7289	0.5067	0.5972
			✓	✓	0.4587	0.6209	0.4518	0.5229
		✓	✓		0.4268	0.5553	0.4208	0.4787
3 Categories	✓	✓		✓	<b>0.6119</b>	<b>0.7913</b>	<b>0.6112</b>	<b>0.6896</b>
	✓		✓	✓	0.5870	0.7760	0.5887	0.6692
		✓	✓	✓	0.5698	0.7188	0.5569	0.6275
	✓	✓	✓		0.5415	0.7418	0.5500	0.6313
All	✓	✓	✓	✓	<b>0.6274</b>	<b>0.7957</b>	<b>0.6268</b>	<b>0.7010</b>



**Context features** can help boost the prediction performance

# User Satisfaction Prediction

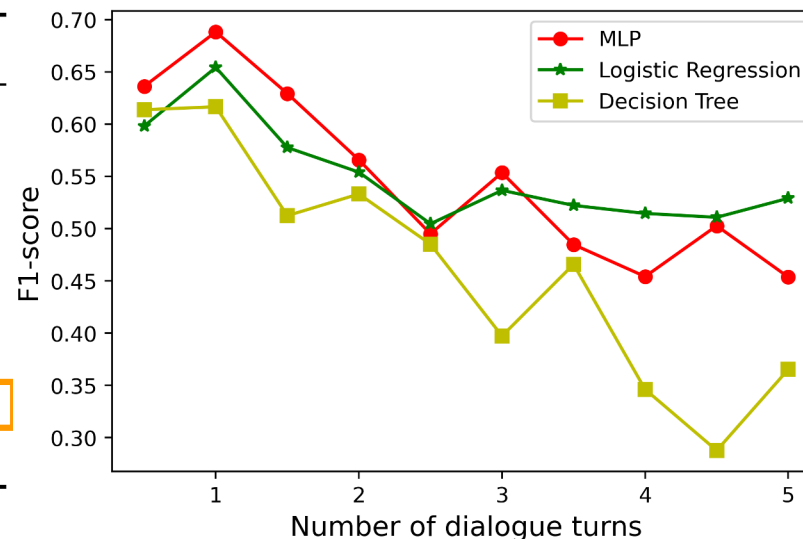
- Binary Classification Problem
- Classification Models
  - 8 Machine Learning Models: LR, SVM, Naive Bayes, XGBoost, MLP, etc.
- Features
  - Dialogue behavior features (i.e., user intents and recommender actions)
  - Utterance-level features (i.e., content, discourse, and sentiment features)

<i>Category</i>	<i>Features</i>
<b>Content</b>	TF-IDF, Name Entity, # Relevant Items
<b>Discourse</b>	POS, 5W1H Question, Question Mark, Exclamation Mark, Utterance Length
<b>Sentiment</b>	Thanks, Sentiment Score, Opinion Lexicon



Methods	Cont	Disc	Sent	Dial	Prec	Rec	F1
Logistic Regression	✓	✓		✓	0.8488	<b>0.5806</b>	0.6795
SVM		✓		✓	0.8778	0.5556	0.6629
Naive Bayes				✓	0.8833	0.5556	0.6651
Decision Tree				✓	0.7109	0.5528	0.6167
Random Forest				✓	0.8862	0.5306	0.6503
XGBoost				✓	0.7897	0.5653	0.6426
<b>MLP</b>				✓	<b>0.8990</b>	0.5681	<b>0.6884</b>
KNN			✓	✓	0.8850	0.5181	0.6427

### Comparison of Classification Models



Method	Cont	Disc	Sent	Dial	Prec	Rec	F1
				✓	<b>0.8990</b>	<b>0.5681</b>	<b>0.6884</b>
MLP	✓				0.6551	0.4944	0.5501
		✓			0.5570	0.3486	0.4122
			✓		0.6067	0.2681	0.3606
	✓	✓	✓	✓	0.7995	0.5444	0.6292

### Comparison of Feature Categories

- Classification Models: MLP (best precision & F1)
- Effective Features: **Dialogue behavior features** (i.e., user intents and recommender actions)

# Summary

- **Taxonomy** established for user intents in dialogue-based CRS
- **User intent prediction**: XGBoost and SVM can achieve outperforming accuracy by unifying four feature categories (i.e., content, sentiment, discourse, and context)
- **User satisfaction prediction**: Leveraging both user intents and recommender actions enables some model like MLP to achieve competitive accuracy

# Future Work

- **Prediction**

- To verify the taxonomy's generalizability to other domains
- To measure the performance of deep learning (DL) methods with more labelled dialogue data
- To investigate the temporal sequence of utterances/responses within a dialogue, for further improving the prediction accuracy

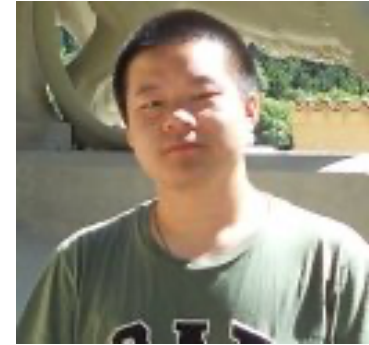
- **System design**

- To integrate more feedback/critiquing aids to match to users' intents
- To study how system-suggested critiques could guide users to explore (diverse) items

# Thanks!



Ms. Wanling Cai



Dr. Yucheng Jin

**Contact info:**

Dr. Li CHEN

[lichen@comp.hkbu.edu.hk](mailto:lichen@comp.hkbu.edu.hk)

Homepage:

<http://www.comp.hkbu.edu.hk/~lichen/>



Intent Annotation of Recommendation Dialogue (**IARD**) dataset is publicly available at:

[https://www.comp.hkbu.edu.hk/~lichen/download/IARD\\_dataset.html](https://www.comp.hkbu.edu.hk/~lichen/download/IARD_dataset.html)