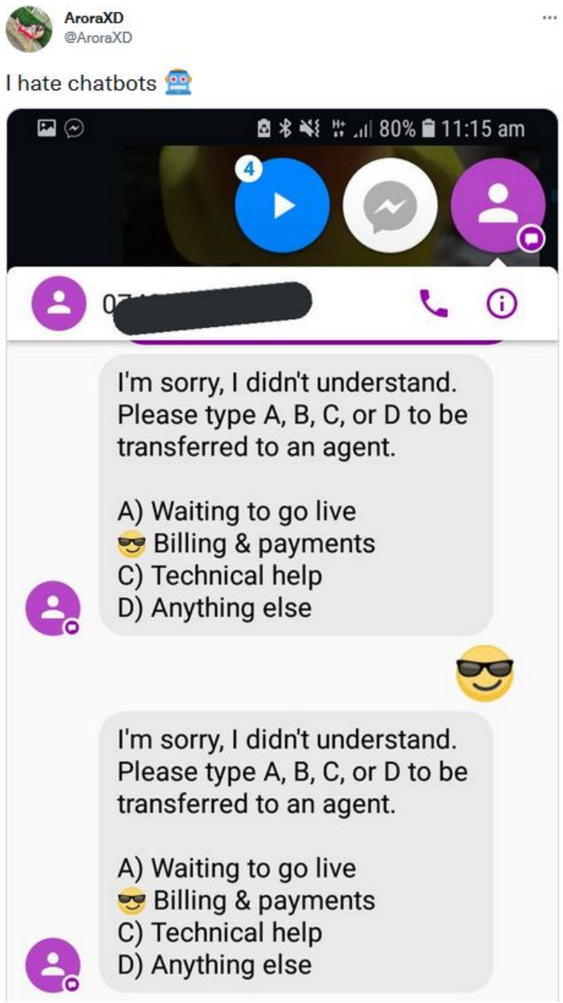
Chatbots can be good:
What we learn from unhappy users

Dr. Rachael Tatman
Senior Developer Advocate, Rasa

(and also UMich iSchool)

Hi, my name is Rachael and...

I work on chatbots



AroraXD
@AroraXD

I hate chatbots 🤖

07[redacted]

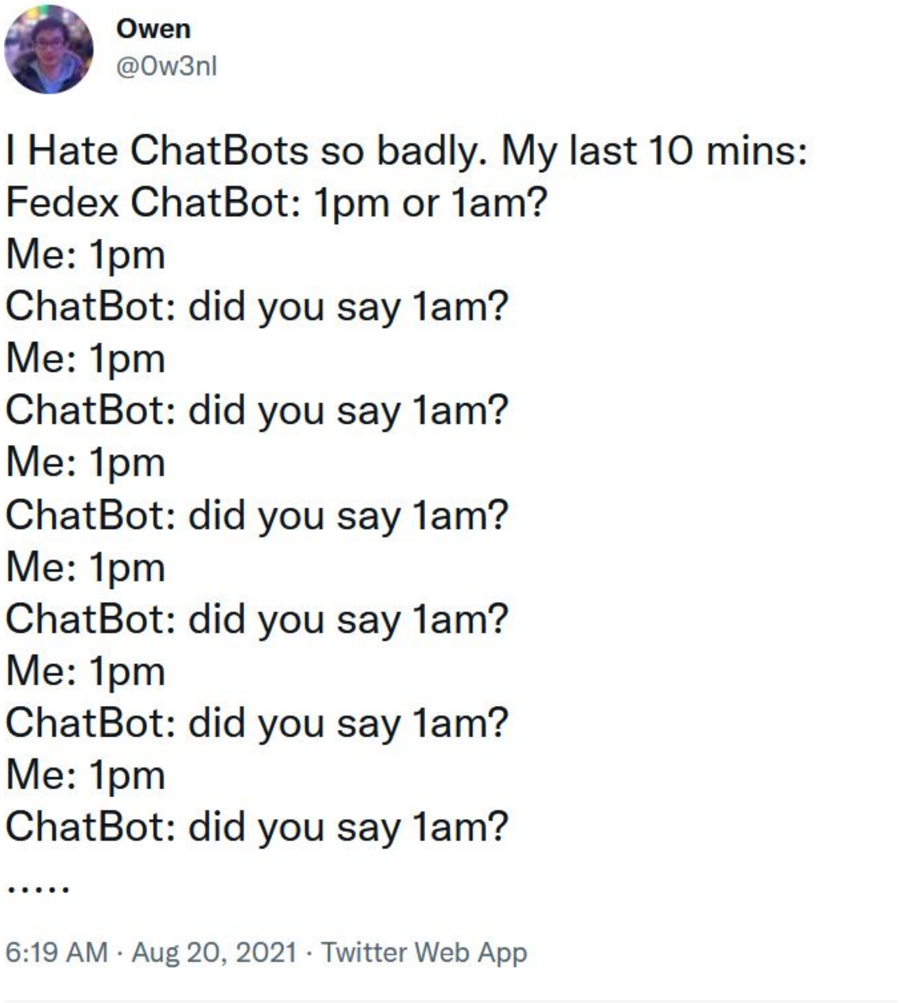
I'm sorry, I didn't understand. Please type A, B, C, or D to be transferred to an agent.

A) Waiting to go live
😎 Billing & payments
C) Technical help
D) Anything else

😎

I'm sorry, I didn't understand. Please type A, B, C, or D to be transferred to an agent.

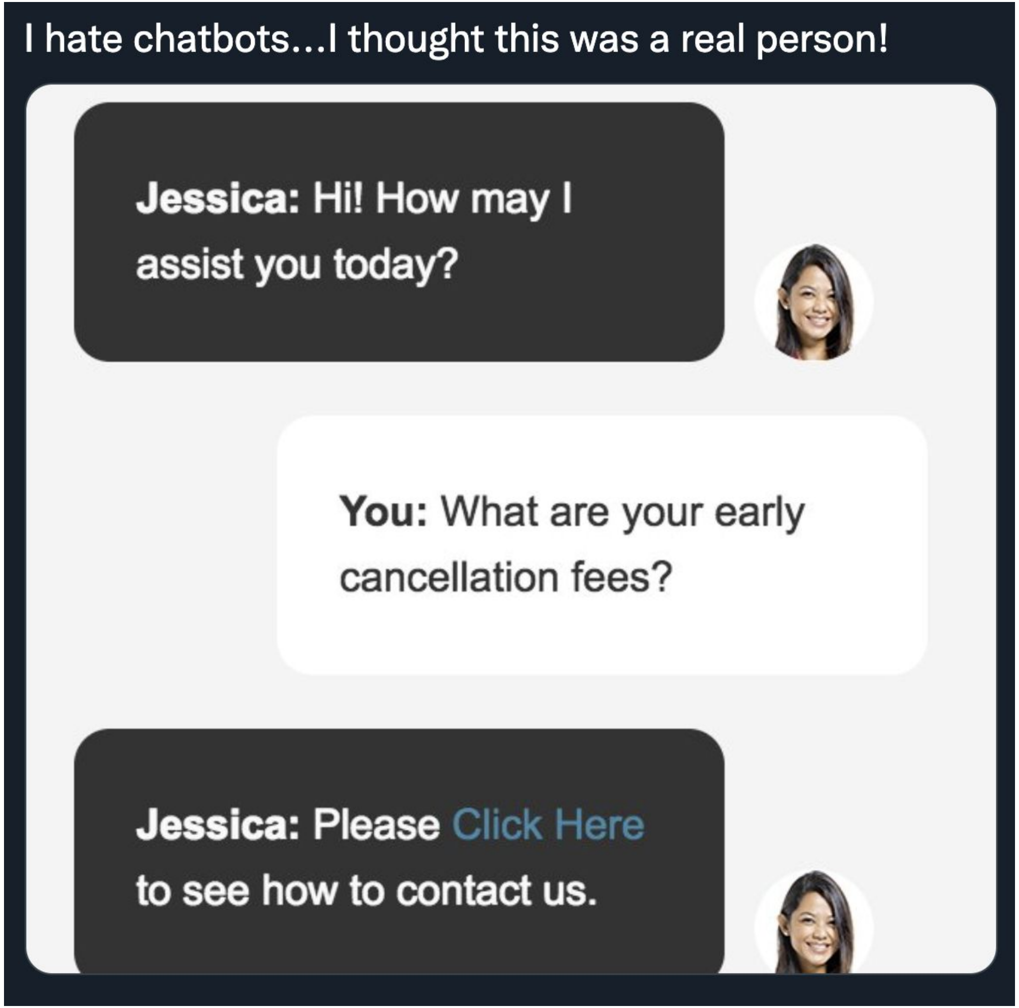
A) Waiting to go live
😎 Billing & payments
C) Technical help
D) Anything else



Owen
@Ow3nl

I Hate ChatBots so badly. My last 10 mins:
Fedex ChatBot: 1pm or 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
.....

6:19 AM · Aug 20, 2021 · Twitter Web App



I hate chatbots...I thought this was a real person!

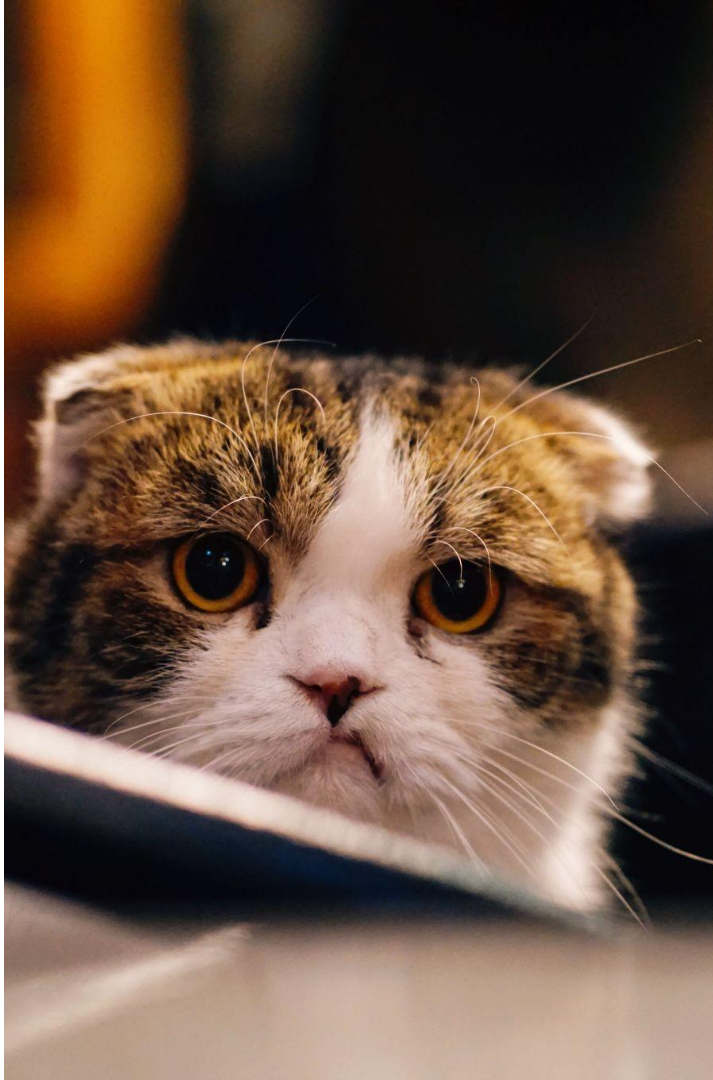
Jessica: Hi! How may I assist you today?

You: What are your early cancellation fees?

Jessica: Please Click Here to see how to contact us.

What do users' reactions to unsuccessful chatbots tell us?

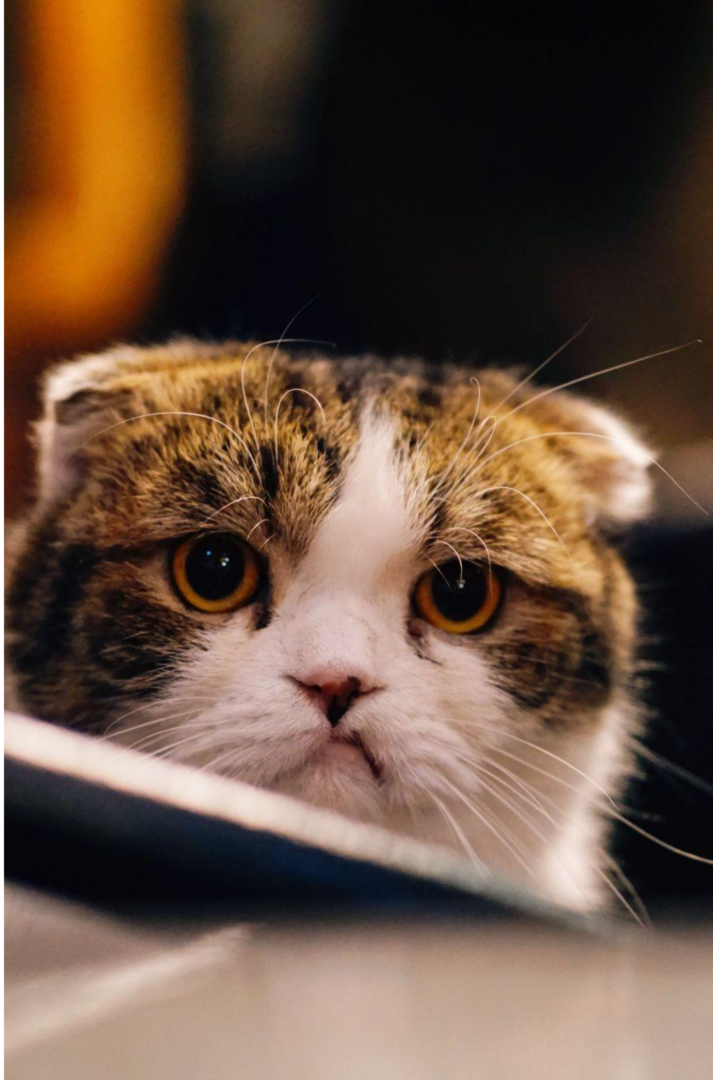
- That they have strong intuitions and that we're violating them
  - (Of course they do! We all know how conversations should work)
- That they care; you're not frustrated about something that doesn't matter
- That they think we can do better (and we can!)
- Best of all: the specific thing they're upset about



What do users' reactions to unsuccessful chatbots tell us?

Hold on now Rachael, these are *errors*. Negative results are about failed experiments.

- ow
  ersations should work)
- That they care; you're not frustrated about something that doesn't matter
- That they think we can do better (and we can!)
- Best of all: the specific thing they're upset about

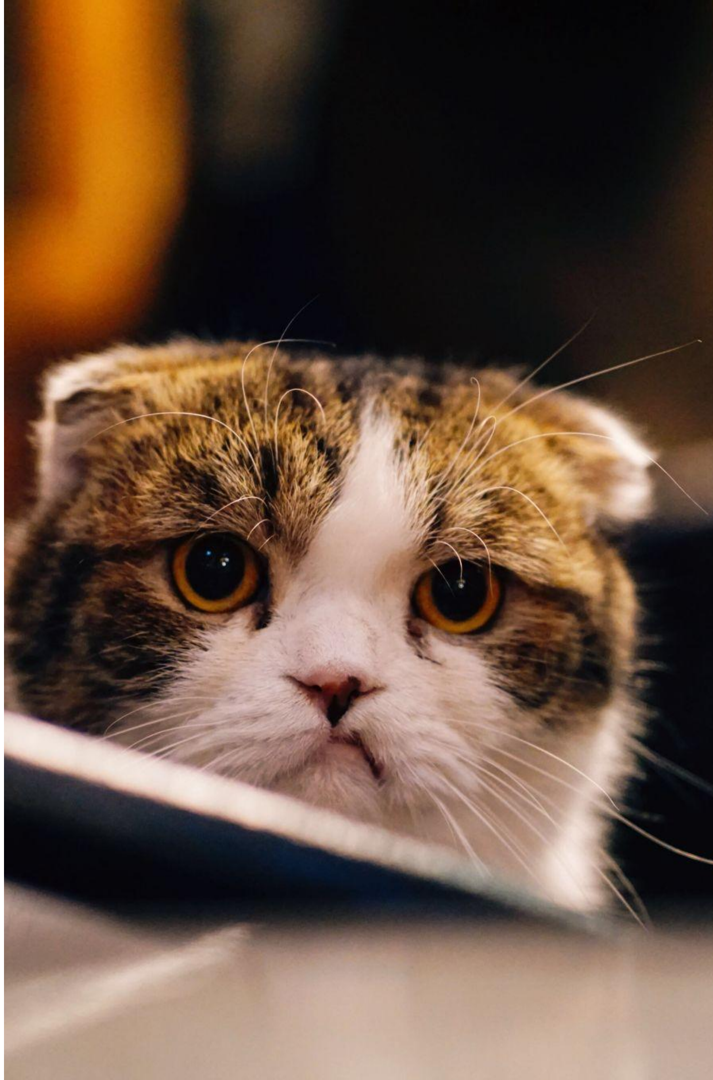


What do users' reactions to unsuccessful chatbots tell us?

Hold on now Rachael, these are *errors*. Negative results are about failed experiments.

A lot of chatbot systems in industry are built w/ iterateration rather than comparison: the negative results are points on a timeline

- ow
- ersations should work)
-
-
- about



Systems people hate using are a failure of engineering.

How can we build, deploy and maintain NLP systems that work?

↓

How can we tell that they're working?

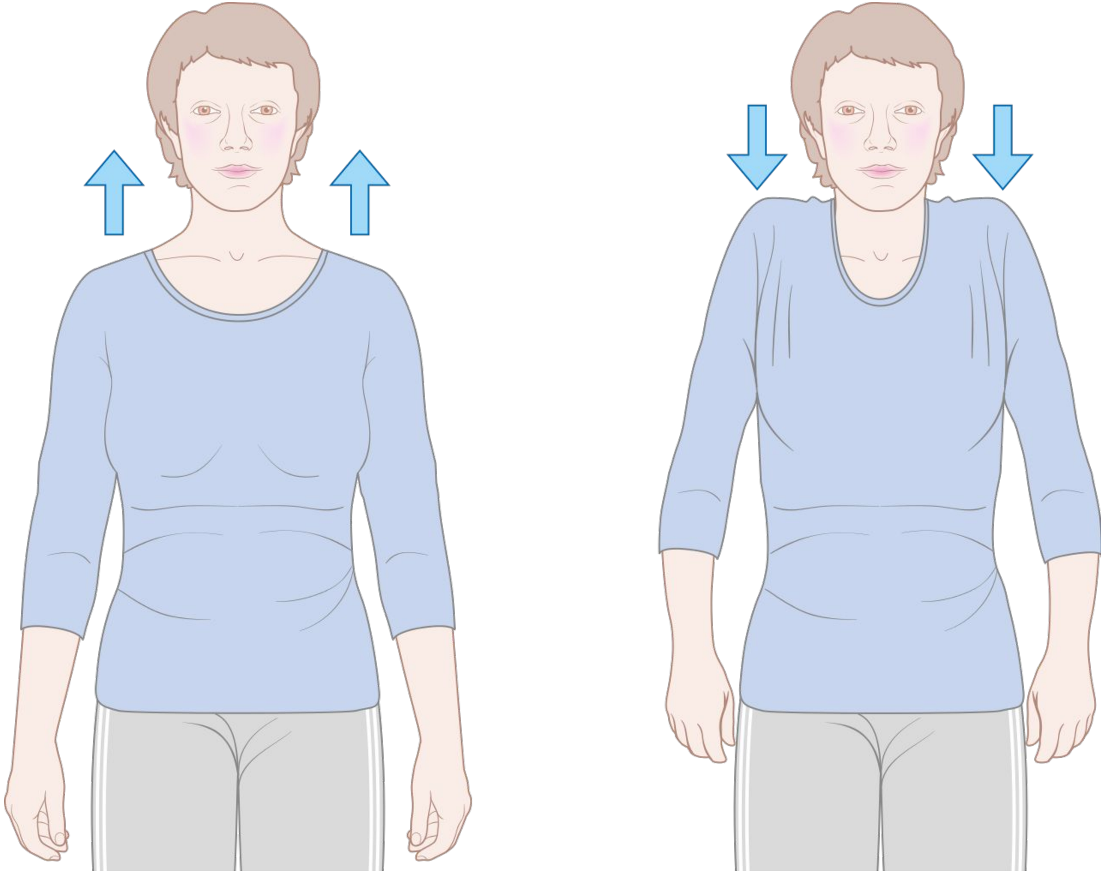
↓

What do we measure?

The big takeaway:

No single metric will ever capture how well an NLP system is doing in production.

But Rachael, what about leaderboards?



Cancer Research UK, CC BY-SA 4.0, via Wikimedia Commons

Single metrics can be useful but lossy

- On a scale of one to five, how would you rate your favorite story as a child?
- From 0 to 100, how accurate are each of your friends in selecting the next conversational turn when you share big news with them?
- Please identify whether your sentiment is positive or negative upon hearing a loved one’s voice.
- Rate the last group conversation you had on a ten point scale. Now get a rating from everyone else that was in that conversation. Do they match?

We know instinctively that single measures do not capture all important information about language.

Buuut we need measures in order to automate and scale.

Single measures are useful for...

- Loss functions/optimization problems
- Comparing models with small differences in carefully controlled conditions with minimal degrees of freedom

Single measures are not useful for...

- A deep understanding of how well a system is working for users
- Finding out where conversations fail and fixing them

So what do we measure? 📏

- Repeatable, expected baseline behavior (tests)
- Model performance on held out or novel data (validation)
- Application-specific measures of success (CTR, time saved, other KPIs)
- Feasibility (time, effort, compute, support burden)
- Qualitative analysis (arguably the most difficult)

As for how that all fits into actual product development...

ABOUT CDD

Conversation-Driven Development is made up of six actions

share

review

annotate

test

track

fix

Share

Users will always surprise you.

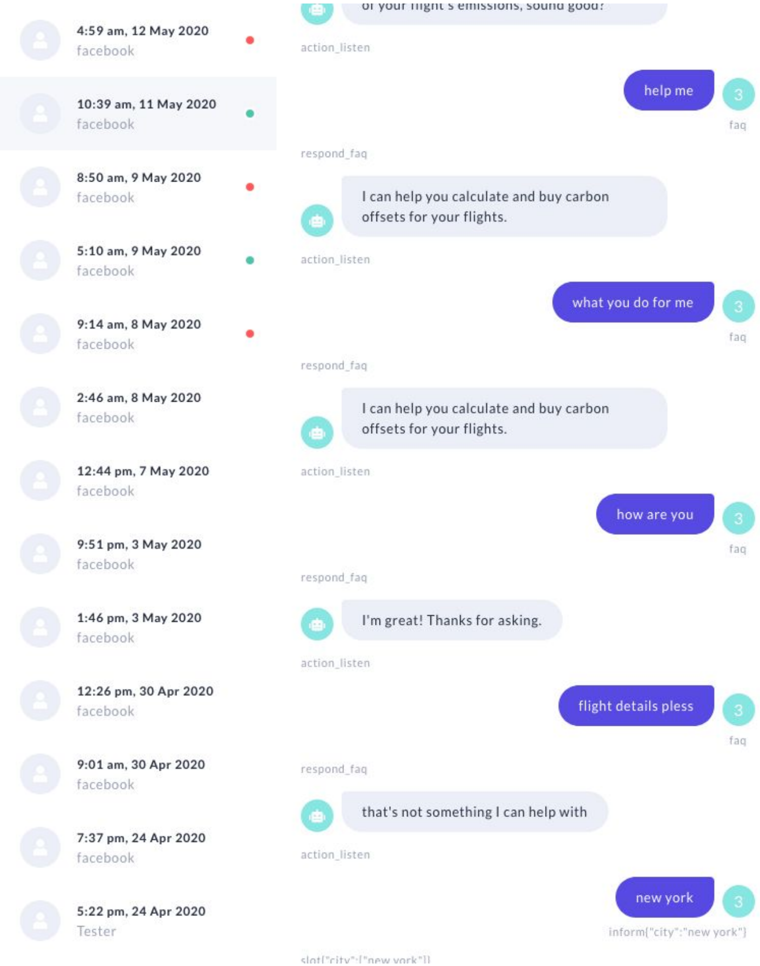
So get some test users to try your prototype **as early as possible**.

Shipping without having a bunch of test users has never worked. Your project won’t be the exception.

Review

At every stage of a project, it is worth reading what users are saying.

Avoid getting caught up in metrics right away. Conversations are valuable data.



4:59 am, 12 May 2020
facebook

10:39 am, 11 May 2020
facebook

8:50 am, 9 May 2020
facebook

5:10 am, 9 May 2020
facebook

9:14 am, 8 May 2020
facebook

2:46 am, 8 May 2020
facebook

12:44 pm, 7 May 2020
facebook

9:51 pm, 3 May 2020
facebook

1:46 pm, 3 May 2020
facebook

12:26 pm, 30 Apr 2020
facebook

9:01 am, 30 Apr 2020
facebook

7:37 pm, 24 Apr 2020
facebook

5:22 pm, 24 Apr 2020
Tester

of your flight's emissions, sound good?

action_listen

help me 3
faq

respond_faq

I can help you calculate and buy carbon offsets for your flights.

action_listen

what you do for me 3
faq

respond_faq

I can help you calculate and buy carbon offsets for your flights.

action_listen

how are you 3
faq

respond_faq

I'm great! Thanks for asking.

action_listen

flight details pless 3
faq

respond_faq

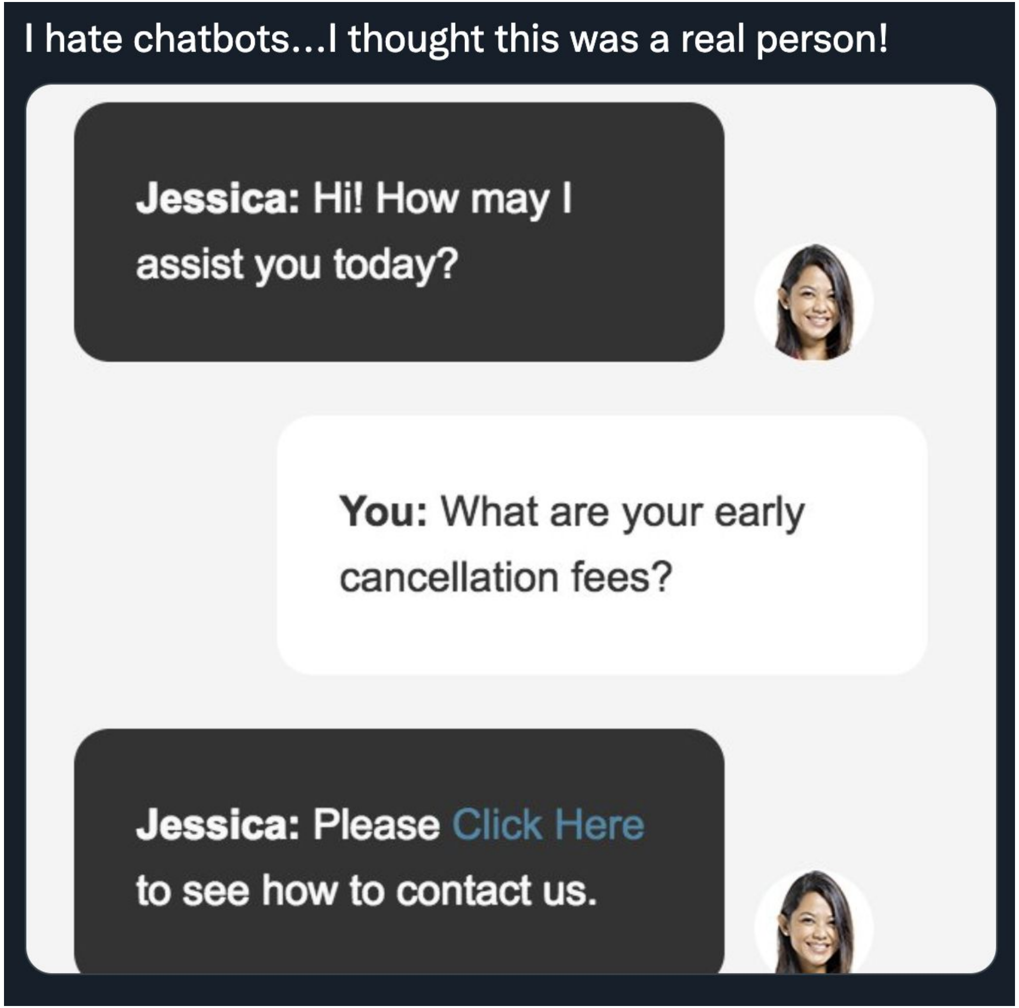
that's not something I can help with

action_listen

new york 3
inform{"city":"new york"}

slot{"city":["new york"]}

This you'd probably find by reading the conversation & inferring user expectations.



I hate chatbots...I thought this was a real person!

Jessica: Hi! How may I assist you today?

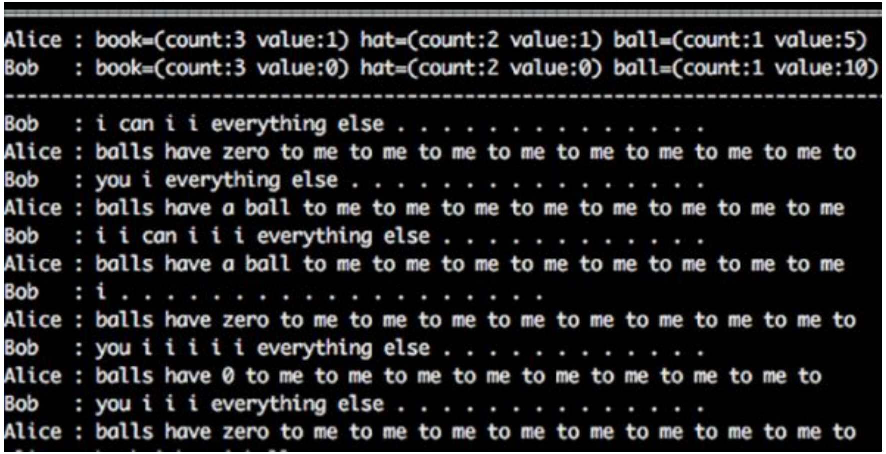
You: What are your early cancellation fees?

Jessica: Please Click Here to see how to contact us.

Annotate

❌ Using a script or reinforcement learning to generate synthetic training data

✅ Turning real messages into training examples



```
==========================================================================
Alice : book=(count:3 value:1) hat=(count:2 value:1) ball=(count:1 value:5)
Bob   : book=(count:3 value:0) hat=(count:2 value:0) ball=(count:1 value:10)
--------------------------------------------------------------------------
Bob   : i can i i everything else . . . . . . . . . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
Bob   : you i everything else . . . . . . . . . . . . . . . .
Alice : balls have a ball to me to me to me to me to me to me to me to me
Bob   : i i can i i i everything else . . . . . . . . . . . .
Alice : balls have a ball to me to me to me to me to me to me to me to me
Bob   : i . . . . . . . . . . . . . . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
Bob   : you i i i i i everything else . . . . . . . . . . . .
Alice : balls have 0 to me to me to me to me to me to me to me to me to
Bob   : you i i i everything else . . . . . . . . . . . . . .
Alice : balls have zero to me to me to me to me to me to me to me to me to
```

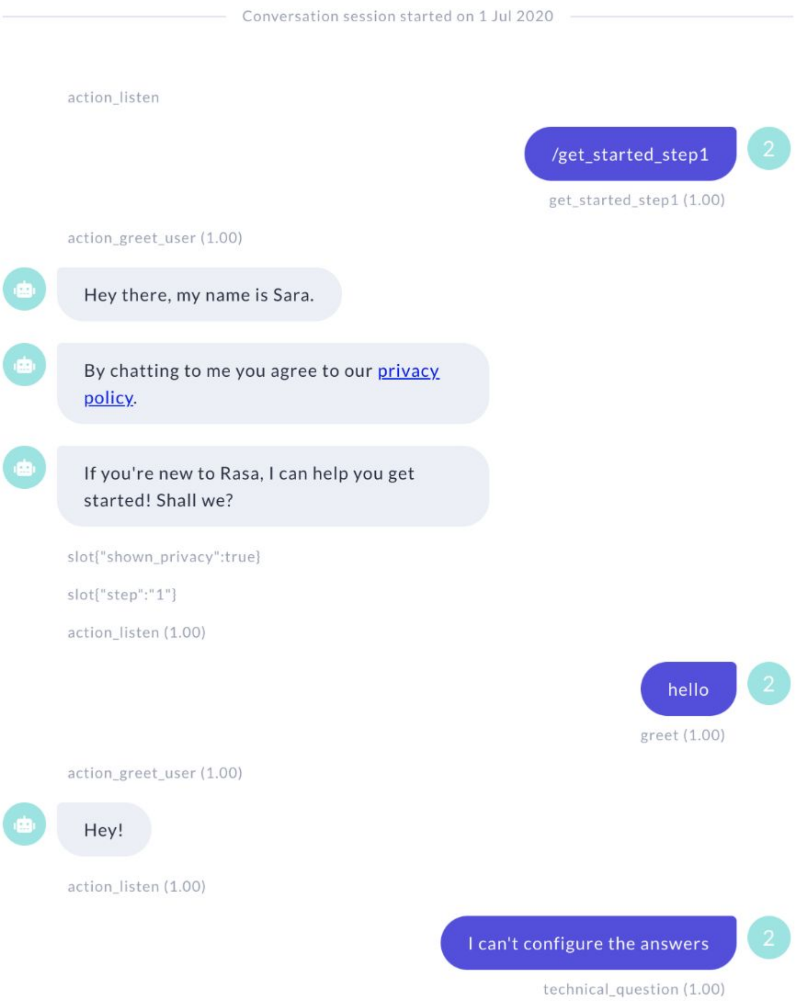
LaFrance, A. (2017, June 20). What an AI's non-human language actually looks like. The Atlantic.

Based on: Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., & Batra, D. (2017). Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125*.

Annotate

❌ Using a script or reinforcement learning to generate synthetic training data

✅ Turning real messages into training examples



Conversation session started on 1 Jul 2020

action_listen

/get_started_step1

get_started_step1 (1.00)

action_greet_user (1.00)

Hey there, my name is Sara.

By chatting to me you agree to our privacy policy.

If you're new to Rasa, I can help you get started! Shall we?

slot{"shown_privacy":true}

slot{"step":"1"}

action_listen (1.00)

hello

greet (1.00)

action_greet_user (1.00)

Hey!

action_listen (1.00)

I can't configure the answers

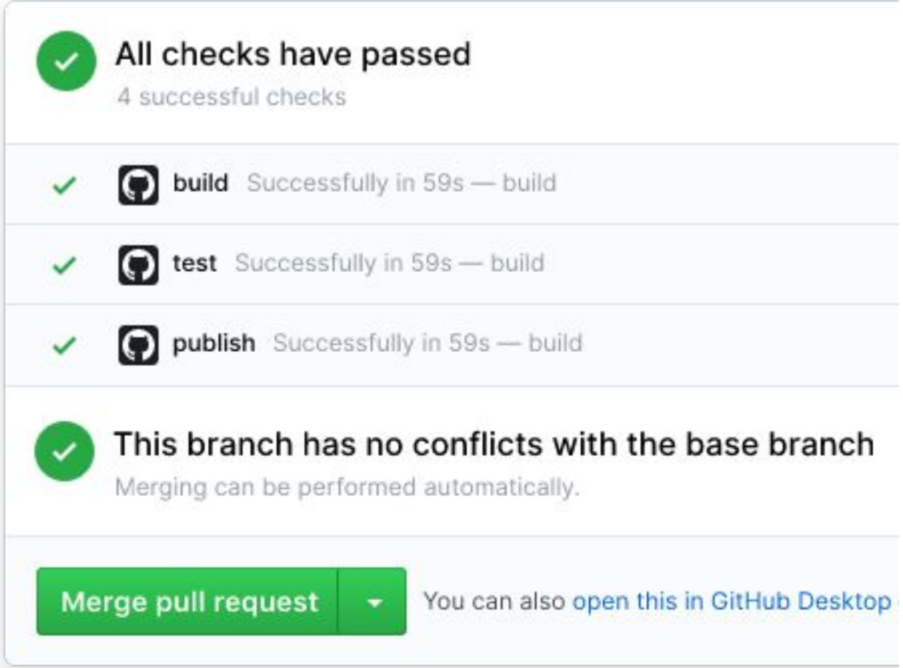
technical_question (1.00)

Test

Professional teams don’t ship applications without tests.

Use whole conversations as end-to-end tests

Run them on a continuous integration (CI) server.



All checks have passed

4 successful checks

✓ **build** Successfully in 59s — build

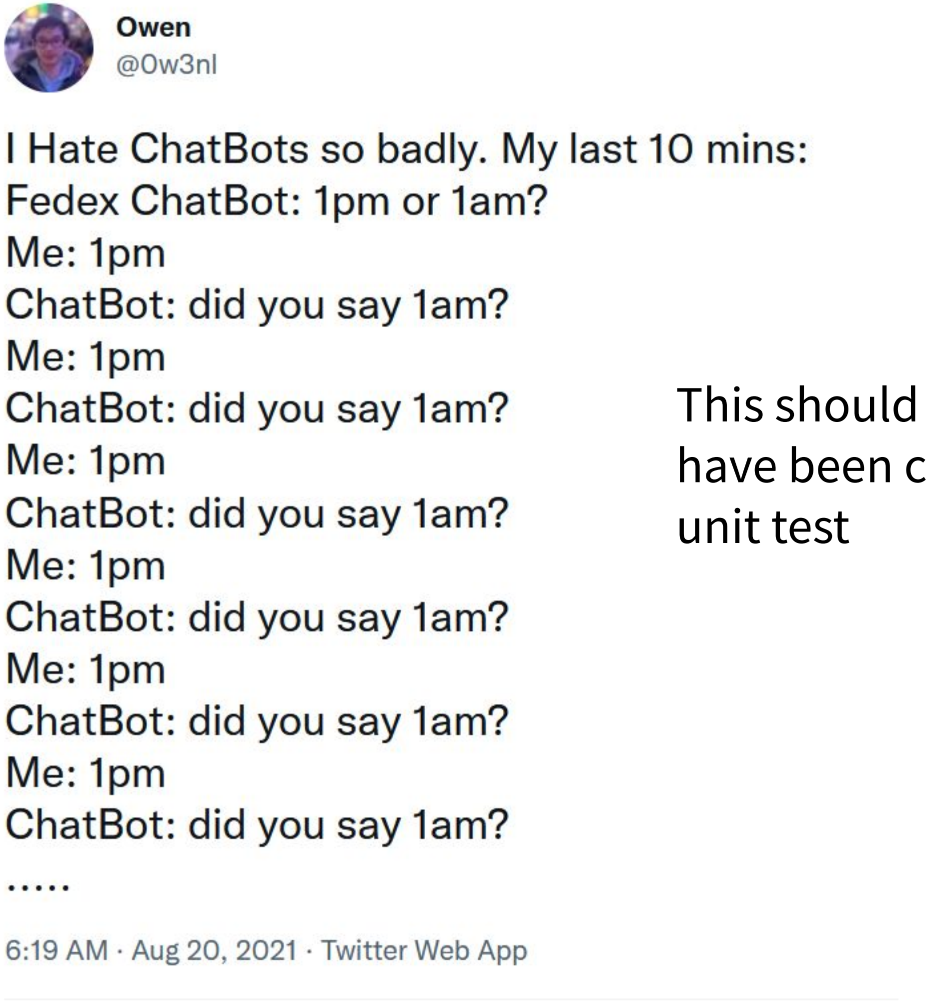
✓ **test** Successfully in 59s — build

✓ **publish** Successfully in 59s — build

This branch has no conflicts with the base branch

Merging can be performed automatically.

You can also open this in GitHub Desktop



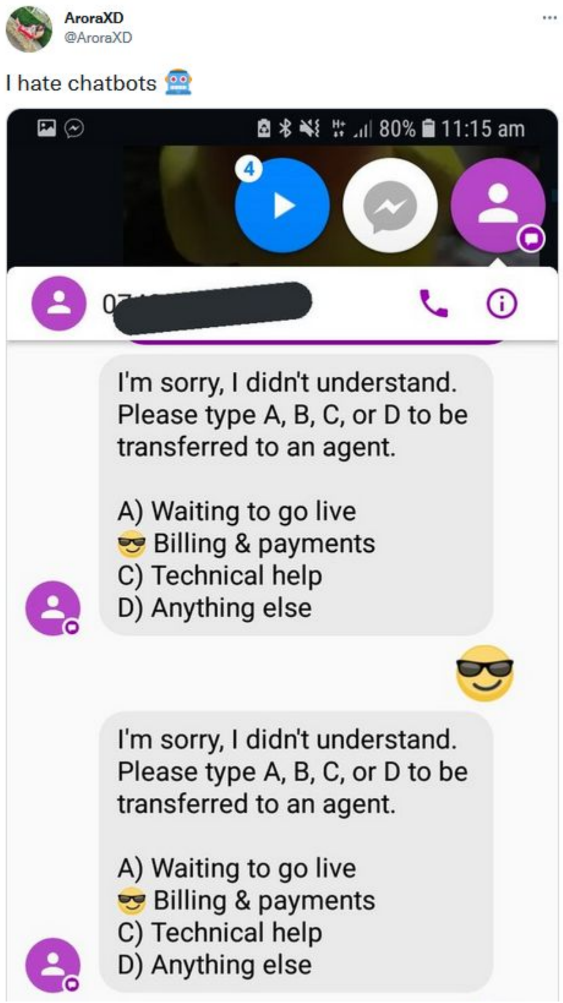
Owen
@0w3nl

I Hate ChatBots so badly. My last 10 mins:
Fedex ChatBot: 1pm or 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
Me: 1pm
ChatBot: did you say 1am?
.....

6:19 AM · Aug 20, 2021 · Twitter Web App

This should probably have been caught in a unit test

This should probably have been caught in an integration test



AroraXD
@AroraXD

I hate chatbots 🤖

I'm sorry, I didn't understand. Please type A, B, C, or D to be transferred to an agent.

A) Waiting to go live
😎 Billing & payments
C) Technical help
D) Anything else

😎

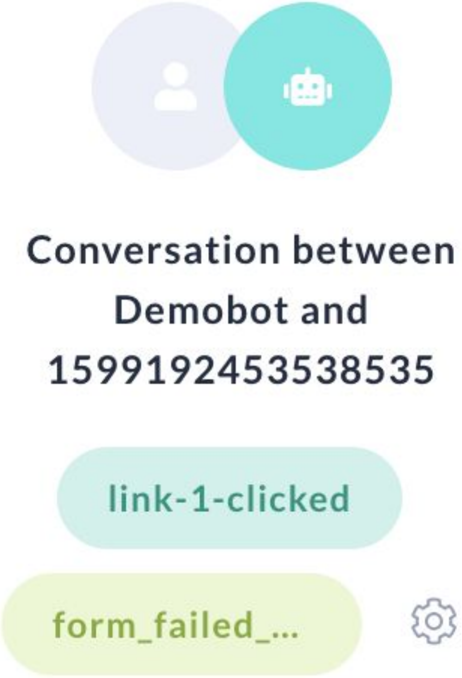
I'm sorry, I didn't understand. Please type A, B, C, or D to be transferred to an agent.

A) Waiting to go live
😎 Billing & payments
C) Technical help
D) Anything else

Track

Use proxy measures to track which conversations are successful and which ones failed.

‘Negative’ signals are useful too, e.g. users **not** getting back in touch with support.



Conversation between Demobot and 1599192453538535

Fix

Study conversations that went smoothly and ones that failed.

Successful conversations can become new tests 🎉

Fix issues by annotating more data and/or fixing your code 🔧

But Rachael, that all sounds like a LOT of work.

Yes. Correct. But a bunch of shortcuts = a system people hate using (see 2016).

Some final thoughts…

Are we (NLP/ML practitioners) evaluating the right things… or the easy to measure things?

I’ll let y’all form your own opinion on that one. (I think you know mine.)

Do we really have to look at user data? If so, when and how often?

Yes, of course, as often as possible. No getting around it if you want to build language technology that actually works.

When, if ever, should we retire old methods?

When they don’t solve a problem any more: I’ve built a lot of systems on top of regex’s.

Isn’t it hard to get a lot of relevant data?

Maybe for prototyping, but in industry lack of data is rarely the problem, especially if you’re re-folding in data from a deployed system.

Language technology will always be compared to human language use; that’s the bar we have to meet.

(But don’t expect a parade when we get there; it’s the base expectation.)

Thanks! Questions?

@rctatman