Shaamer Kumar
Ryusuke Suehiro

What is the main goal of the project?

Current chatbots have four main issues: a lack of personalisation, contain misunderstandings, and have a lack of empathy. Firstly, there's a big lack of personalization in their responses. Chatbots tend to offer very generic, unhelpful answers, failing to recognize and adapt to the unique circumstances and needs of a customers' concerns. Secondly, a significant challenge with chatbots is their propensity for misunderstandings and miscommunications. Due to common dependence on keyword recognition instead of more intelligent techniques, chatbots can easily misinterpret customer intentions or miss the nuances of human language. Finally, tying all these together is the issue that chatbots have a large lack of empathy. When facing personal issues, many chatbots respond in tones inappropriate for the corresponding situation, which is essential for a satisfactory customer service experience. All of these as a combination leave customers feeling undervalued and more frustrated from when they started, demanding an audience with a customer representative which ends up wasting everyone's time in the end. According to Zendesk's 2022 CX Trends report, around 60% of customers see disappointment when dealing with chatbots. The goal of our project is to create a chatbot that adds empathy in responses specifically for professional use i.e the airline industry, and behaves as close to how a human representative would, providing appropriate responses according to the situation at hand.

What are the main claims?

Our claim is that by fine tuning this Dialo-GPT model, pretrained on conversational text like Reddit, we can customize it to make it a great airline customer service chatbot. We believe that with our methodology of creating a custom loss function that penalizes responses to common airline customer complaints and questions with bot-like text, it will reduce severe customer anger that often leads to them never using that airline ever again. Since poor airline customer support is a super relevant issue that has not yet been addressed much, and with airlines losing customers and therefore revenue through this, we believe our fine-tuned chatbot can help tackle this issue

What are the experiments?

We are going to be fine tuning the dialoGPT model and training it with more formal, customer-query-specific data in order to achieve our goals. We will be comparing the dialoGPT original model with our resulting models across different learning rates, batch sizes, and combinations of datasets.

What is the evaluation protocol?

Creating a loss function for this type of problem is tricky. That is why we spent hours upon hours looking for customer queries to human-response datasets. At the moment we plan to use cross-entropy losses from what the model outputs to what the "correct" empathetic response would be in order to train our model. Further steps here would include adding custom components to this loss function i.e making sure the length of response isn't too short or too long, making sure we don’t include certain phrases that are demeaning like “you should have” or negative wording.

What is the data?
The data we are using is from Hugging Face titled Customer-Report-Responses. It has detailed responses to customer queries like I can't find the item I'm looking for. With a response of We're here to help. Can you please provide a description or product name of the item you're looking for so we can assist you? Evidently, this can take the output responses up a huge notch by making them more prompting, clear, and problem-specific as opposed to a generic reply that most chatbots would output i.e looking for keywords etc.

What is the task?
The task at hand, as mentioned earlier, is to fine tune dialoGPT which achieves conversational-type responses and make it slightly more formal, and respectful when dealing with airline-type issues like customer complaints.

How do the experiments support the goal/claims of the paper?
By directly training with customer complaints, we tailor the model and teach it how to deal with more advanced questions and complaints. By experimenting over different parameters i.e datasets, batch sizes, and learning rates we hope to see large improvement in the current model’s responses for these types of issues.

Are any of the limitations discussed in the paper?
The limitations we have faced so far are largely with regards to data reformatting and training large sums of data. Since we are using Google CoLab right now, we are exploring a potential switch to VSCode and connecting to a remote server that has a lot of GPU.

What are the strengths of the paper?
We believe the biggest strength of our paper lies in the replicability and the big degree of customization we place in addressing one core issue at hand, which is mitigating severe customer anger. Our custom loss function specifically ensures that our pretrained model is improved through fine-tuning by exponentially penalizing responses similar to automated bots.

What are the weaknesses of the paper?
The biggest challenge right now is the lack of relevant results to revise our methodology, but this will be worked upon in the coming days and addressed by the next peer review.

Provide a suggestion for improving the paper.
One suggestion for improving the plan for what the final paper should look like would be to include our own data and see how that will train. There were close to no datasets that we could

find directly relating to airline issues, just general customer support queries. Thus, I suggest we add about 100-200 queries with human-responses for airline specific content. This way our model can learn how to deal with customers while also staying within our broader context of airline chatbots.

What is the relevant related work?
The relevant related work is all stored here :
https://huggingface.co/microsoft/DialoGPT-medium?text=Hey+my+name+is+Julien%21+How+are+you%3F

This is the dialoGPT model that we are basing our work from and it includes any and all coding done prior to this project.

Is the paper reproducible?
Yes

Can you rerun the experiments?
Yes

Can you reproduce the results in the paper?
Yes

Is the English in the paper correct and clear?
Yes

Do you have any feedback on any TODOs that the authors have left at this stage?
We have quite a few TODOs left for this project. Due to the nature of extenuating circumstances with severe sickness, and how we only have had 2 people in our team, my partner and I tried the best we could given our personal circumstances. We have initial codes and frameworks set up, however, we do not yet have a working model. We have, however, put a lot of thought into the methodology and implementation which we have described in the paper. We will get our model to train for this in the coming days. Next is to plot loss curves, once our model has finished training and we have different versions with different parameters, we need to plot out the different loss curves for each of these and see what performs the best. Is it the original DialoGPT? Or tailored? One big TODO would be in my additional suggestion, where we can create our own data or append it to the larger dataset we got online to include airline-specific queries and how to handle them like a human would. One way to test this effectively would be to call an airline service with a list of questions and see how the operator would respond to it. Finally, we need to enhance the customized loss function we have right now. It is relatively

barebones, only including a few terms that are negative to add a negative loss to and if the sentence is too long or short we will add to the loss too.