Your Paper

Sebastian Sjöberg, Adam Sundqvist, Eddie Laser, Allan Salimi & Lucas Wallin

January 12, 2024

Abstract

The project is about modeling a virtual bartender using Furhat as the interaction interface. Furhat would be able to react to human emotions in somewhat convincing way and be able to carry out a scripted conversation. This would include taking inout from a human source, reacting to the input and respond accordingly.

The dataset that were used managed to create an algorithm with an almost 90% training. By using two different datasets with images of human faces, both AI-generated and real images of people.

The algorithm was sufficient enough to use for creating a video of a group member interacting with Furhat. The interaction was clear and the machine managed to hold a conversation, give a drink and also a fact about the drink.

1 Introduction

In this project we chose to make a virtual bartender using Furhat as the interaction interface. We wanted to make it be able to react to human emotions in a somewhat convincing way and carrying out a scripted conversation, including taking the human input and responding according to that. Our main goals were to set up a conversation with at least a few interactions back and forward, and to make the human feel like the answers (s)he gave had an impact on where the conversation is going, also to get a good prediction of the human emotion to incorporate that into the conversation.

The work division started with Lucas and Allan working on the conversation diagram, mapping out the flow of the conversation and what choices there would be and what they resulted in. Meanwhile Adam, Eddie and Sebastian started work on the coding. Sebastian made great progress with the CNN model for predicting emotions and Adam and Eddie went on with setting up the code structure. When we had a good structure and a working prediction model, the group worked together to join all parts with the conversation flow. We also added some mimicry of the Furhat bartender with the perceived emotion of the human.

2 Methodology

2.1 Overall system design

As mentioned in the introduction, we chose to make a virtual bartender instead of a mindfulness coach. The reason behind this choice was that we immediately saw some opportunities with making a rule based conversation. The benefit of choosing the bartender is that we felt that we could make the virtual bartender lead the conversation more than a mindfulness coach, which made our job easier since the bartender could give the human interacting with Furhat a few options, e.g. if you prefer beer or cocktails. Our goal was to have the conversation tree a few conversations deep, to get a satisfying interaction with Furhat, and to get a model that could predict human emotion with an accuracy that we felt was good enough.

To achieve these goals we designed the conversation to have a few opening pleasantries, followed by the bot asking if you prefer beer or cocktails and then what type of flavour profile one prefers. By dividing these steps up we achieved a bit of a longer conversation, and the answer for each question would then map to a specific drink, according to our emotion- and preference to drink map. To get the accuracy of the model right, we started with some of the previously used machine learning algorithms,

like Random Forest, Linear Regression and SVM, to predict the action units. These worked okay but we did not manage to tune them to get the accuracy we were looking for. We then started to explore the possibility of using a Convolutional Neural Network, in part to skip the middle step of using action units and just going from pictures to the four emotions we classified, and in part to see if we could train it to achieve a greater accuracy than was previously attained.

2.2 User Perception sub-system

2.2.1 Design

For this sub-system the main part was to perceive the users emotion, a task which at core can be quite difficult. There was a couple of choices that needed to be done to do this, firstly and perhaps most importantly, is the choice of the different emotions that can actually be perceived, to begin with, the group focused on seven emotion, but quickly decided on boiling it down to four emotion by aggregating. These four emotions are the following, aghast, furious, happy and melancholic, these corresponds in order to surprise+fear, angry+disgusted, happy and finally sad+neutral. Other than this there was a decision to be made in how to approach this from a machine learning standpoint, the group identified two possible ways, firstly using some smaller machine learning model that predicts emotions by using action units extracted trough some other models and the option the group actually settled on, using a Convolutional Neural Network that simply uses the pixel values from the face box extracted by OpenCV CascadeClassifier [1].

2.2.2 Implementation

When starting to develop the CNN we firstly started by finding a dataset, first we used the DiffusionFer set, where we extracted the biggest facebox using OpenCV CascadeClassifier, we also used all of the seven emotions mentioned above. To begin with we simply initialized a neural network with three convolutional layers followed by two dense layers, we quickly realized that the model had some problems, mainly that it was overfitting, while also having some problems getting above 70% accuracy on the validation test. To combat this the first step we took was to add some dropouts, which did help some in that it negated the overfitting, but the accuracy was still not quite to satisfaction. We then discussed with some TA's which pointed us to do as much data augmentation as possible as well as trying to expand the dataset. After this feedback we firstly found the FER2013 set which we decided to merge with the DiffusionFER, we also added some random flips, colour jitter as well as some gaussian blur to the training images, these augmentations were also reinitialized for each epoch. This yielded better results although it was quite slow and painful to train (a hundred epochs took about 12h), and ended up being the model we settled for. This specific model consists of three dense layers with size 16, 32, 48 respectively, with firstly relu activation as well as max pooling between each layer, there were also a dropout between the second and third layer, following this were two dense layers with size 40 and 4 with a relu activation on the second to last layer and an in theory softmax activation (Achieved trough cross entropy loss). This model was then combined with the OpenCV classifier to be able to firstly extract the face and secondly the emotion from a picture.

2.2.3 Results

This system ended up performing quite well, first of on the actual data set where it achieved good accuracy of 87.04%, 88.63% and 87.8% in order on the training, validation and test set. Furthermore what was actually nice to see was that it ended up performing quite good in production with it being able to quite accurately recognize emotions from an actual human face, although it is apparent that there are some specific things that the model ends up looking at that does mirror the correct emotion but can sometimes be a little over exaggerating for a human to actually express.

2.3 Interaction sub-system

2.3.1 Design

For the design of the interaction with the bartender, we decided on making a linear interaction flow consisting of six different interactions with Furhat. In the first interaction the bartender greets the

user asking for a name. This name is then saved for a response. In the background Furhat will simultaneously register the users emotion and base interaction two on what it can detect. The emotions are divided into four classes where each class will result in a different question from the bartender. However, we decided that continuing with four different interaction flows from this would be two much work, whatever the user responds on either of these questions, the bartender will respond with the same answer. The third and fourth interaction focuses on what kind of beverage the user would like. firstly there is the option of either beer or cocktail. For these, the user will have a choice of either bitter, sweet, strong or fruity. These choices are stored in two separate variables. Furhat will then use these saved variables, emotion, beer or cocktail and taste preference, and map a beverage for that. There are in other words 16 different types of beer and 16 different types of cocktails that the bartender can choose from. The chosen beverage will be presented in interaction five. In the final interaction, number six, the bartender will give some trivia about the chosen drink.

To see what the interaction-flow looks like, see the appendix.

2.3.2 Implementation

The main difference from the design is how the interaction functions correspond to each interaction int the flow-chart. In order to make the implementation easier there is a shift or/and an overlap in some functions. This subsection will however explain it for each interaction as the previous subsection.

To implement this interaction-flow in python we use interaction counter. This is a variable, *interaction_count* that can have the value 0-5. The counter works as a while loop in the *main.py* file. From *main.py* the programs calls a function in the file *furhatinteraction.py*. This function, *interaction*, uses pythons match and case where it matches the value of *interaction_count* to one of six cases where each case is an interaction. The Furhat itself is the *FurhatRemoteAPI* that we use to make the bartender to speak and react.

In order for the bartender to remember what the user had said and what it should say in the next interaction, we are using a variable that we call, *context*. In the first interaction the name of the user will be stored in *context["Name"]* and *context["Question"]* will be assigned the value "Name". In the next interaction the program will therefore know that the last question that was asked was regarding to the users name. *Context* will in other words be used for in if-statements in the interactions to help the program in what each interaction is supposed to do.

For speech recognition, we are using the python library *speech_recognition*. This library is used to capture the users voice and then convert the voice to text. The text will then be analysed in a way that suits the current interaction. For interaction one, the program uses the library *spaCy* which is a natural language processing model that we use to find the users name [2]. In the second interaction we use the model created by us to find the users emotion explained in section 2.2 of this report. In the third interaction we use a *SentimentIntensityAnalyzer* from the *vaderSentiment* library. Here we look for the sentiment in the users response, where the bartender asks only if the user wants a beer and not if the user wants a beer or cocktail. When only asking about beer, we can determine if the answer is positive i.e. the user want a beer, or negative i.e. the user wants a cocktail. In the fourth interaction we use a simple if substring in string to find what taste preference the user have to then do a sentiment analysis again. Since two of the preferences, bitter and sweet, have a sentiment themselves, we solve that by saving the preference and then removing that word from the text created from voice capture. In the final two interactions we first use the saved preferences in *context* to find a suitable beverage. After that the bartender will ask if the user would like to hear some trivia regarding the beverage, if so the bartender will give the trivia, otherwise the bartender will end the conversation.

2.3.3 Results

The interaction flow works well at this point. The bartender is however very limited with the responses which makes it very robotic and non human like. Which is to be expected of a project of this small scale. Part from that it is easy to use.

3 General Discussion

3.1 Discussion of the overall pipeline

Some of the challenges at the start was getting the model right. We started with trying out some of the machine learning models from the course, but then decided to switch to a CNN approach. That turned out to result in a much higher accuracy for predicting emotions and we were happy we made the switch. Another challenge was making the rule based diagram of the conversation. Early on we noticed that too many choices per interaction would make the conversation tree too large, and we then simplified by having fewer choices per interaction but a deeper interaction tree.

3.2 Challenges faced

(what has worked and what has not) There were never any really great challenges when doing the project. What ended up being difficult was making Furhat gestures work by using the Python API. Mainly making Furhat show emotions and doing small gestures that corresponded well with the conversation.

3.3 Use of ChatGPT or similar tools

We used ChatGPT for some parts of this project. One part was the schedule of mapping the emotions and preferences of the human to the drinks. We told it that have the emotions "Furious", "Aghast", "Melancholic" and "Happy" and the preferences "Sweet", "Fruity", "Bitter" and "Strong" and that we wanted all combinations to correspond to a certain drink. We also said that we wanted all combinations to correspond to one drink and one beer each, which resulted in 32 different beverages that we could map the emotion and preference of the user to a certain drink. After that we also used ChatGPT to come up with some fun facts about each drink, which resulted in the last conversation segment where the bartender asks you if you would like to know something about the drink. For example, if you get an Indian Pale Ale, and agree to the fun fact, Furhat will respond "Indian Pale Ale, or IPA, originated in England and was later adapted by British brewers for export to India. The extra hops helped preserve the beer during the long sea voyage." This may of course lead to some misinformation since it is known that ChatGPT can make up information but we felt that it was a fun ending to the conversation, even if ChatGPT made some stuff up.

3.4 Ethical issues

The dataset that we used first was AI developed images that in our opinion were ethically correct used. Which means we did not think that the images were unfairly treated since they are all imaginary and don't resemble anyone. Nevertheless, it's important to know the consequences of using images that are based on different human faces with stereotypical facial structures.

For our second dataset we used images of real humans to train our algorithm on. This dataset was FER2013, a dataset of images from the website Kaggle. The dataset was public data and therefore we argued that it's alright to use it for learning purposes in academical work.

4 Conclusion

The project turned out great and the group is happy with the choices we made and the implementation was satisfying. The focus was mainly on learning how Furhat works and what components were important for making a great, interacting dialog. By taking it step by step the group managed to complete the assignment with satisfactory results, achieving a high training, test and validation percentage. This was done by applying our knowledge from the work that was done on the previous assignments. Since the group was considered of 5 students, there were a lot of opportunities to divide the project in to smaller parts for a more in depth job. This also freed up the group to try different approaches, for example using CNN instead of random forest classification, which was used in favour for our algorithm.

References

[1] OpenCV. Cascade classifier. https://docs.opencv.org/4.x/db/d28/tutorial$_c ascade_c lassifier.html$, 2024. $Accessed : Jan12, 2024$.

[2] spaCy. en$_c ore_w eb_s m.https : //spacy.io/models/en$, 2024. $Accessed : Jan9, 2024$.

Appendix

A Interaction

A.1 Interaction flow-chart

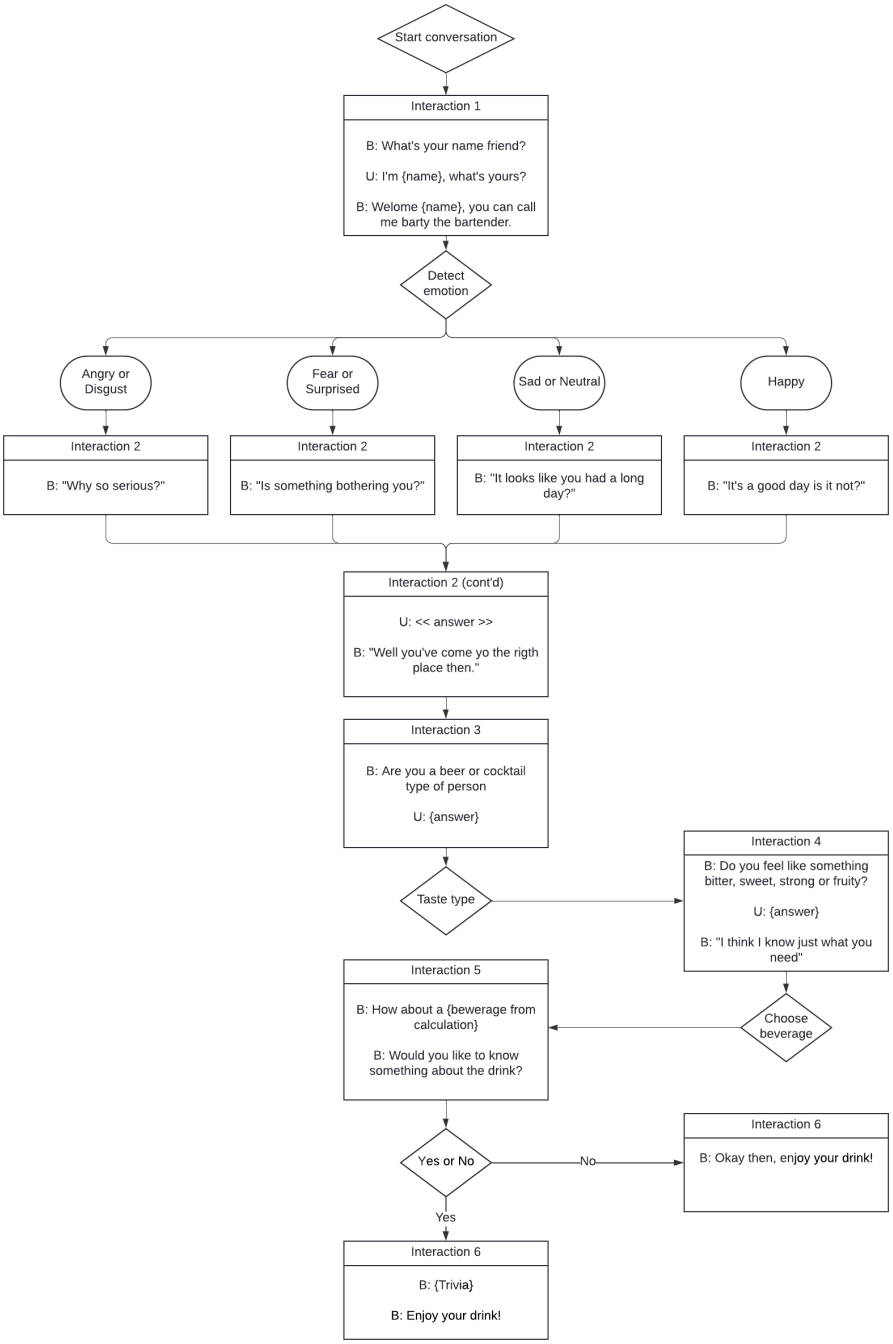

Figure 1: Interaction flow-chart

A.2 Example of mapping of drinks and trivia

```
    drinkdict = {
    "Indian Pale Ale": ["Furious", "Bitter", True],
    "Belgian Double": ["Furious", "Sweet", True],
    "Russian Imperial Stout": ["Furious", "Strong", True],
    "Rasberry Fruit Labmic": ["Furious", "Fruity", True],
    "American Pale Ale": ["Aghast", "Bitter", True],
    "Honey Wheat Ale": ["Aghast", "Sweet", True],
     ...,
    "Negroni": ["Furious", "Bitter", False],
    "Bitter lemon drop": ["Furious", "Sweet", False],
    "Zombie": ["Furious", "Strong", False],
    "Rasberry Mojito": ["Furious", "Fruity", False],
    "Espresso Martini": ["Aghast", "Bitter", False],
    "Blue Lagoon": ["Aghast", "Sweet", False],
    "Long island iced tea": ["Aghast", "Strong", False],
    ...,
}
```

```
    drinkinfodict = {
    "Indian Pale Ale": "Indian Pale Ale, or IPA, originated in England and was later
    adapted by British brewers for export to India. The extra hops helped preserve the
    beer during the long sea voyage.",
    "Belgian Double": "Belgian Double is a rich and malty beer brewed by Belgian
    Trappist monks, known for its deep flavors and higher alcohol content.",
    "Russian Imperial Stout": "Russian Imperial Stout, a favorite of Catherine the
    Great's court in 18th-century Russia, is a robust and dark beer with origins in England.",
    "Rasberry Fruit Labmic": "Raspberry Fruit Lambic is a Belgian beer with a
    fruity twist. Lambics are fermented through exposure to wild yeast and
    bacteria, resulting in a unique and refreshing brew.",
    "American Pale Ale": "American Pale Ale, a hop-forward beer style, emerged in
    the United States during the craft beer revolution, showcasing the vibrant
    flavors of American hops.",
    "Honey Wheat Ale": "Honey Wheat Ale is a sweet and smooth beer that often
    incorporates honey during the brewing process, adding a touch of natural sweetness.",
    ...,
}
```

article graphicx [left=2.5cm, right=2.5cm, top=2cm]geometry
IIS - Project specification
Group - MINT Eddie Laser, Allan Salimi, Sebastian Sjöberg, Adam Sundqvist, Lucas Wallin
November 2023

A Context

A.1 Elevator pitch

The purpose of this project is to create an emotionally aware Furhat bartender that is capable of recommending the user a drink from a set menu depending on the users emotions and answers to a couple of questions.

A.2 Objectives

The problems we are trying to solve are programming an interactive robot that can reliably predict human emotions and respond to this in a convincing manner. One part of this consists of detecting emotions, which we will constrict to a few basic emotions to ease run-time performance, another part is interacting with the human counterpart and saying appropriate lines based on what the input is. Finally we want to build a machine learning model that can recommend a drink based on the detected emotion and human inputs.

A.3 Deliverables

Deliverables for this project will be to build the different components for the bartender. Initially we will sketch the architecture for the system so that we have a better idea of how to build the bartender. Then each deliverable will focus on different components.

- Create the initial architecture.
- Visual emotion detection.
- Emotion detection based on questions.
- Create a set of questions.
- Base questions on the emotions detected by camera.
- Connect emotions with cocktails.
- Save whether the drink suggestion was correct or not for future training.
- Create a demo system.
- Create a demo video for the presentation.
- Project report.

A.4 Success metrics

To track the progress of the project the group will try to set deadlines for each objective and have certain goals to achieve ahead of every feedback session. To know if we are keeping our time schedule we will check where we are progress-wise before the feedback sessions and determine if we need more sessions or not. The emotional prediction of the human face will be considered satisfactory if the human interacting feels that Furhat accurately responds to the human emotion by mirroring the human, for example if the human is happy, Furhat shall express happiness, if the human is sad, Furhat shall express emotions of compassion. Therefore the success of the project lies in the eyes of the human interacting.

A.5 Potential issues

The purpose of the project is to create an Furhat bot that, based on data, offers a drink for an individual. The issue will be to accurately collect and train the data for the purpose. Since no data has been collected the model has to be trained on data that the group will collect from scratch. It's really important that the collected data is relevant to the Furhat bot. For the sake of the project it's not necessary that the Furhat bot has to many labels, so that it does not get too complex. It's

also a potential problem that the recommendation of a drink is not really a task that a ML approach can solve, this could also be an interesting result tough in proving that the bartender profession is dependant on humans.

B Project Breakdown

B.1 Feedback Session 1 - 24/11

Have an idea and a complete project specification

B.2 Feedback Session 2 - 6/12

Have a reasonably well working emotion detection and started work in both data collection and ML method for drink recommendation as well as mapping out the interactions

B.3 Feedback Session 3 - 15/12

Be in principle done with the emotion detection and mapping of the interaction. Have a working implementation of the interaction regarding asking questions ans such and also have a working ML model for the drink recommendation as well as a growing dataset.

B.4 Presentation - 10/1

Finished the presentation, reasonably finished report and a working Furhat based system.

B.5 Report Submission - 12/1

Completely finished report as well as making sure the system is one hundred percent working.