Project Report

Project G3: CHATBOT
Team members: Mark Tarnovski, Siim Tänavots, Tuule Tars
Repository: https://github.com/Si1mT/ChatBot

Identifying business goals

Background

We are participating in an introductory data science course in our university. Part of the course is a project wherein we need to go through the steps of doing a data science project. The project's objectives and scope must be somehow related to data science but otherwise can be freely chosen by us, according to how much we think we can accomplish within a few weeks.
For the project we have decided on creating a chatbot. Chatbots are programs that mimic conversation and have historically been fun toy programs to interact with. They can be built with machine learning algorithms and techniques which originate from data science. However recent improvements in those algorithms(along with the development of new ones) combined with the vast amounts of data that is available on the internet have created ChatGPT and later other commercial chatbots, which have millions of daily users, because they can help their users solve their problems, much unlike older toy chatbots.

Business goals

Our goal is relatively simple. We want to compare the performance of a machine learning algorithm when trained on human- vs AI-generated data to see how the different datasets affect the quality of a chatbot’s conversational skill. During that, we will also try to find the best machine learning algorithm to train the bot that suits our needs the best.

Business success criteria

The performance of our chatbots will be evaluated by ourselves. We will see how intelligible each chatbot is and based on that we will qualitatively rank each bot. We will also observe the effects of having more input data and if it improves our bots, it will be considered a success.

Assessing the situation

Inventory of resources

Our resources are the team members, the initial dataset we acquired from Kaggle, extra data from generative AI, an OpenAI API key and our personal computers.

Requirements, assumptions, and constraints

Requirements:

- Project and poster is finished before the poster session on December 13th
- OpenAI API key

Assumptions:

- Team members contribute to the completion of the project

There are no notable constraints.

Risks and contingencies

This aspect is irrelevant to our project.

Terminology

This aspect is irrelevant to our project.

Costs and benefits

A small monetary cost might be incurred by printing a big poster for the poster session.

Defining data-mining (machine learning) goals

Machine learning goals

Our machine learning goal is to identify the relevant machine learning algorithms used in creating chatbots and implementing them. Then, additional training with AI-generated data is done to create a hopefully better version of the bot. The user should be able to talk with both versions of the bot and get a response from them.

Success criteria

Minimum success is each chatbot version being able to generate a response for a prompt. Ideally each version responds in a different way and with enough responses, comparing the quality of the bots will be possible.

Gathering data

Outlining data requirements

To cover our project's machine learning goals, all the data needs to be in a human- and machine-readable text format. Considering the main language of our course, the data should be in English.

Verifying data availability

The first part of the data that will be used is already acquired from the Kaggle datasets and saved to the project repository, which means it will be available at all times.
The second part, which will be generated by another AI (our first choice is ChatGPT, but any other may be used) during the development process, should also be available as long as there are no connection issues with OpenAI servers. To verify the availability and possibility of such data creation method, a small test file was created using ChatGPT based on an example text from the first (Kaggle) dataset, which proved to be successful.

Defining selection criteria

The data sources used for the project are simple txt files with two columns (train and target data), separated by a tab. As they were specifically generated/compiled for the purpose of our project, all the data from these files is planned to be used. However, if during the learning process we notice that some topics repeat too often and/or get the model (chatbot) stuck in a loop, some filtering on the input data may be done to ensure normal workflow.

Describing data

The train data should consist of simple and concise questions (1-2 sentences long) on the topic of day-to-day life. For example: "So how have you been lately?". Declarative sentences (not questions) are also possible if they prompt some sort of answer in an everyday context. For example, the sentence "It's such a nice day", which could be followed by "Yes, it is". The target data should consist of answers to the according questions and/or sentences that support the conversation, which can also be questions themselves, also 1-2 sentences long.

Exploring data

While exploring the data we found that some modifications will have to be made in order to better suit the machine training algorithms. For example, writing out the

words that use apostrophes (“it’s” into “it is”) or separating periods (“.”) from the words.
As far as the topics of questions/answers provided by the Kaggle dataset are concerned, there does not seem to be any visible quality issues that might impact the machine learning process. However, some sentences appear to be based on specifically American topics (like living in California or studying in Pasadena city college), which might be modified to better suit the Estonian environment, though this is not a task of particularly high significance.

Verifying data quality

After the aforementioned modifications, the data provided by input files (Kaggle and AI-generated files) should be sufficient for our goals and meet the requirements.
Concerning the possibility of having too little or low-quality data, it is good that our project uses an expandable data source (AI), which means that we will always have the ability to gather new data or replace one that might not be suitable.

List of tasks

1. Data preparation. (15 h)
    a. Making necessary changes to the initial dataset from Kaggle to suit our needs for the training.
2. Loading the model for the chatbot. (2 h)
3. Training the chatbot on the initial dataset. (15 h)
    a. The model we plan to use is GPT3 via OpenAI API or BERT via Hugging Face Model Hub..
4. Testing the chatbot I (4 h)
    a. Feeding the chatbot a list of prompts and saving the outputs.
    b. If necessary, adjusting the training parameters and iterating.
5. Generating additional training data using ChatGPT. (2 h)
6. Training the chatbot on generated data. (15 h)
7. Testing the chatbot II (4 h)
    a. Feeding the chatbot a list of prompts and save the outputs.
    b. If necessary, adjusting the training parameters and iterating.
8. Ranking the bots based on their performance while trained on the initial dataset from Kaggle vs the generated dataset. (5 h)
9. Presenting results. (3x10 h) (10 person-hours per team member.)
    a. Formulating conclusions.
    b. Making a summary.
    c. Creating a poster.