

Capstone Project Proposal



Olivier Polack

Business Goals

Project Overview and Goal

What is the industry problem you are trying to solve? Why use ML/AI in solving this task? Be as specific as you can when describing how ML/AI can provide value. For example, if you're labeling images, how will this help the business?

The industry problem I am trying to solve is to get valuable feedback from recorded phone calls in the form of customer satisfaction and quality of service indexes. Many service and goods providers may send automated emails to get customers feedback thanks to surveys customers have to fill in online. However the percentage of completed feedback surveys is limited.

Using Automatic speech emotion recognition (SER) systems were introduced several years ago but their main drawback is that there are almost no databases with real (non-acted) emotions appropriately rated by experts. It

Classification of Emotions and Evaluation of Customer Satisfaction from Speech in Real World Acoustic Environments

Classifying emotions and evaluating customer satisfaction from speech can help any businesses as they have more earnings from customers buying again goods or services rather than new customers.

has been shown that SER systems are suitable to help call-center managers in monitoring and optimizing the quality of service provided by their agents. These systems detect the emotional state of agents and/or customers and hence provide a quality of service index. Abnormal changes in service patterns like increasing number of angry customers can be detected. Labelling all phone calls with ML/AI regarding customer satisfaction and emotional feelings will help businesses spot the weaknesses and their

	strengths.
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring revenue, market share, customer happiness and/or other drivers of business success.	<p>A common practice in call-centers consists of recording and storing customer calls for posterior analyses. Those calls are listened to and evaluated with the aim of improving customer satisfaction and the quality of service (QoS). This procedure is usually hand made by randomly taking small samples from the total set of calls. There is a double cost for rating the calls: answering the calls and evaluating them; and only a very little number of call among the total set is evaluated.</p> <p>One of my first step working as an AI product managers for a given brand owner would be to ask the brand owner to estimate and consider costs, delays and outcome of the posterior analyses (listening and hand-made ratings). Then based on assumptions made one the number of phone calls totally received and the costs estimations of AI analysis on all of the recordings, the difference in costs can be assessed, and give a first hint on how relevant the use of AI will be.</p>
Application of ML/AI What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?	<p>The precise task ML/AI will accomplish is to give:</p> <ol style="list-style-type: none"> 1. At first a binary 'happy customer' / 'unhappy customer' classification. 2. Secondly potentially a rating on a scale from 1 (customer is unhappy) to 5 (customer is as happy as he could be) if it makes sense. 3. Thirdly a multi-labelled classification based one emotion responses such as 'customer is angry', 'customer is sad', 'customer is joyful', customer is 'annoyed'. <p>The business outcome is to give brand owners perceived quality and customer satisfaction indexes that can be very quickly monitored.</p> <p>By filtering the analysis results per types of products or services the brand/business owners can also get relevant clues on where the areas of improvements are, spotting which lineups yield lower as expected or lower as other lineups results.</p>

Success Metrics

Success Metrics

What business metrics will you apply to determine the success of your product? Good metrics are clearly defined and easily measurable. Specify how you will establish a baseline value to provide a point of comparison.

The success metrics are better when they are discussed in depth with end-users. In our case help-centers

Baseline can be established based on a benchmark of existing competitive products. But this kind of data especially if similar products hardly exist can be hard to get. So I would start with baseline metrics with figures agreed together with end-users.

Data

Data Acquisition

Where will you source your data from? What is the cost to acquire these data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?

Data should mainly be sourced from real-world helpdesk phone conversations between service or good customers and customer service agents. Existing emotional speech databases often consists of acted emotions that are mostly recorded under controlled-acoustic conditions, while the call-center database comprises recordings of real conversations of customers and service agents that are collected without any control over the recording process or the channel. Additionally, these helpdesk conversations are labeled by experts in customer service, which is the real way of evaluating these kinds of interactions in real industrial applications. Existing databases for emotionnal speech recognition are free like for example : <https://github.com/HLTSingapore/Emotional-Speech-Data>.

This is hard to know how much would cost real-world labeled or not helpdesk conversations. (Existing datasets are available at very varying prices – ie 50USD up to 6000 USD - from this [website](#) but these datasets are meant for speech recognition and not emotionnal reactions classifications). Not yet labelled conversations from the following top outsourced helpdesk US companies probably already exist. The idea would be to contact potential data recordings providers, ie helpdesk top agencies, advertise about potential benefits once the product is ready, so they agree to give access under agreed and acceptable terms and conditions. Some might already have labeled datasets or existing similar products running, hence this is one way to also get figures on potential existing competing products.

About PII, yes, as the customer service phone agents have to confirm the customer name and address there is a need to make sure audio recording file names are anonyms. Audio files should not include customers names and private information.

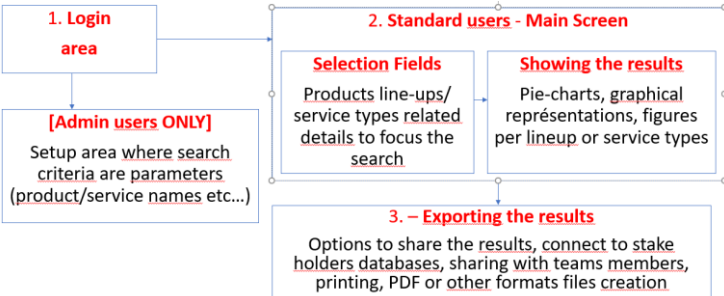
Once some data batches have been collected to

	<p>start prototyping, as long as helpdesks exist there is potentially more data available on regular bases. It is easier to get more data once the first few samples have been collected. Especially if prototypes early results can be shown to agencies so they are interested in the project.</p>
<p>Data Source</p> <p>Consider the size and source of your data; what biases are built into the data and how might the data be improved?</p>	<p>About the size, typical available emotional datasets seem to contain at least several thousands samples. So the starting point for prototype would be around 10 000 samples.</p> <p>To avoid biases, it is necessary to have various sets of teams and groups working on the data. For example, biases could arise if the data samples are mostly coming from a group of persons of the same gender, or the same ethnicity in majority. The more data is collected from diverse sources, the higher the chances biases are avoided.</p>
<p>Choice of Data Labels</p> <p>What labels did you decide to add to your data? And why did you decide on these labels versus any other option?</p>	<p>Depending on the market and the customers type there can be no point in trying to fit a fit a given model.</p> <p>I will start with basic “customer is happy” / “customer is not happy” binary classification prototypes. These early prototype models help assess datasets are ok and to show meaningful models results to share with data team and stakeholders so as to get their opinions.</p> <p>But emotional typical response types should be of course added later on:</p> <ul style="list-style-type: none"> -Customer is angry -Customer is annoyed -Customer is “neutral” -Customer is disappointed -Customer is satisfied -Customer is cheerful <p>The goal here is to get more detail on the type of response.</p> <p>What matters is to have real data classified in terms of customer response, but also customer needs to be able to filter out responses and compare responses of different lineups or product references or phone agencies.</p>

Model

Model Building How will you resource building the model that you need? Will you outsource model training and/or hosting to an external platform, or will you build the model using an in-house team, and why?	<p>I will initially do a quick prototype by building the model using Automated ML tools/services like Google Auto ML or Amazon SageMaker. Once I am able to observe that I am able to do what I wanted to do, I will go over the metrics and see if the model is achieving the performance. If it is then I go ahead with the model, else I will try to see where I can improve the model like adding more data etc. If I am still not satisfied with the performance of the model, then I try to build the model using in house team as Automated ML tools do not perform well in all use cases.</p>
Evaluating Results Which model performance metrics are appropriate to measure the success of your model? What level of performance is required?	<p>Common metrics like precision, recall, and F1-score will be used.</p> <p>Emotional classifications is not like medical diagnostics classifications, or anything hazardous. So between recall and precision, it is precision that matters the most. This means the number and the percentage of false positives would need to be attention to.</p> <p>As a baseline, at least 90% of the conversations should be classified correctly especially if we will be using other models to classify further.</p>

Minimum Viable Product (MVP)

<p>Design</p> <p>What does your minimum viable product look like? Include sketches of your product.</p>	<p>The minimum viable product should be discussed and agreed by stakeholders.</p> <p>The product would have a graphical user interface including a credentials page to login, an administrator page for administrators, at least one page/screen for selecting fields research criteria and the results based on those selected products/lineups or responses criteria.</p> <p>The is a need to export results to PDF, print document too.</p>  <pre> graph TD A[1. Login area] --> B[2. Standard users - Main Screen] B --> C[3. - Exporting the results] subgraph B [2. Standard users - Main Screen] D[Selection Fields Products line-ups/ service types related details to focus the search] --> E[Showing the results Pie-charts, graphical representations, figures per lineup or service types] end </pre> <p>1. Login area</p> <p>[Admin users ONLY] Setup area where search criteria are parameters (product/service names etc...)</p> <p>2. Standard users - Main Screen</p> <p>Selection Fields Products line-ups/ service types related details to focus the search</p> <p>Showing the results Pie-charts, graphical representations, figures per lineup or service types</p> <p>3. - Exporting the results Options to share the results, connect to stake holders databases, sharing with teams members, printing, PDF or other formats files creation</p>
<p>Use Cases</p> <p>What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?</p>	<p>The product is aimed at customers service managers and directors. It can decision makers among companies such as CGS Inc. or Global Help Desk Services, as well as brand owners actually delivering goods and looking after how their products and services are perceived.</p> <p>An Epic-level case of the product is when a car maker is launching a brand new car on the market and suddenly sees a growing dissatisfaction. He can identify quickly if the problem comes from by narrowing down the search criteria.</p> <p>Users access the product “on-line”, login and get the results. Depending the level of security and data governance policies, where and how data is stored will depend.</p>
<p>Roll-out</p> <p>How will this be adopted? What does the go-to-market plan look like?</p>	<p>The first phase will include beta testing within the Data science team and close stake holders, a small set of end-users. The goal is to gain insights whether the appropriate direction is taken for the tool.</p> <p>The second phase is to select the best prototype(s) and start test fields with the performance and</p>

accuracy of the artificial agent being analyzed and discussed with data team. Datasets improvements are critical at this stage, but also later on.

The third phase will be running product(s) after correcting all what can be in accordance with budgets, timing, resource and existing state-of-the-art available models and datasets.


For each phase, outcomes will be provided to the stakeholders so as to get their feedback.

Continuous improvement stage keeps on even afterwards, as models and ways to get better datasets are found.

At all stages, the question of data sensitivity (PII, PHI...) datasets taken from stakeholders must be addressed and agreed upon to make sure no privacy infringements exist.

Post-MVP-Deployment

Designing for Longevity How might you improve your product in the long-term? How might real-world data be different from the training data? How will your product learn from new data? How might you employ A/B testing to improve your product?	<p>The product might be improved in the long term by implementation of newer ML/AI models that might give faster and/or more relevant results.</p> <p>Real-world recorded phone calls might have unexpected background noises that somehow interfere and cause the model not to perform as good as on the training database. This is important to get real-world data from diverse sources as much as possible and retrain with new data if it makes to see if it improves the model. For instance, The product may have to include some labels with typical heavy background noises or other bad quality signals. These bad quality audio files would be added to some training datasets to see if model could filter out bad quality audio files and screen them out as bad quality recording.</p> <p>New data can come from recordings with other sampling settings, with other languages, with other industrial or customers products line-ups where the range of emotions and the ways they are expressed come to be different (ie recurring other training datasets)</p>
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	<p>I would look at the amount of false negatives and false positives relatively to each others. If one is significantly higher than the other, then the model might be biased for several reasons.</p> <p>Model bias is more to data team or AI models providers.</p> <p>Data bias will arise if datasets are unbalanced. As their might be more phone calls with customers claiming as they are phoning because they have issues, this aspect has to be taken into account.</p> <p>There might inherently more unhappy customers that pleased one, because happy customers might tend to less express themselves.</p> <p>One can think of splitting actual phone recordings into two parts, the first few moments where customer probably expresses disappointment, and the final few words where in the end this is mostly the perceived quality of service rendered by the</p>



helpdesk agent itself that is rendered in the emotional response of the customer. This timing aspect clearly means there is room for improvements and investigations.

Human bias will arise because human annotators, in our case hopefully service agents specialists. The pool of service agents annotators has to be taken into account (which industrial or customer area are they familiar will matter).