Capstone Project Proposal

Odili Charles Opute

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?	Poor access to health services and qualified personnel is a critical problem in Africa, resulting in millions of deaths from preventable and easily treatable diseases like Lower Respiratory Tract Infections (LRTI), which have replaced HIV/Aids as the number one killer in Africa. Machine Learning promises profund value in healthcare since it can power a product or service that combines very high quality LRTI prediction (diagnosis) and complementary medical advisory, delivered directly to patients or aid clinicians and semi-skilled caregivers provide scalable care without compromising quality. The impact of this includes timely access to personalised and affordable diagnosis and treatment of LRTI, even for Africa's poor patients, while aiding clinicians and driving overall ROI. **Ṣàìsàn** is the product we will build to address this problem. It means "sick" in Yoruba.
Business Case Why is this an important problem to solve? Make a case for building this product in terms of its impact on recurring	According to the WHO, Lower Respiratory Tract Infections (LRTI) have become the number one cause of death in low income countries The most common LRTI diseases are bronchitis and pneumonia, and in 2016, these infections accounted for almost 1 million deaths, surpassing deaths from HIV/Aids.

revenue, market share, customer happiness and/or other drivers of business success.	**Case In Point** Nigeria, which is largely a low income country with over 50% of its population reported to be living below the poverty line, has a doctor:patient ratio of 1:6000 against WHO's recommendation of about 1:1000. This means the majority of Nigerians have very poor access to health personnel and quality diagnosis of LRTI or just cannot afford one, and thus are at higher risk of death if infected. Solving the problem of acute doctor shortage as well as late or poor diagnosis and treatment of LRTI is one that a Machine Learning product layered over the ubiquity of mobile technology is very well suited for. Such a product will save lives, help clinicians perform better diagnosis and treatment of LRTI to more people, and bring in substantial revenues from the large footprint of patients and clinicians who are supported.
Application of ML/AI What precise task will you use ML/AI to accomplish? What business outcome or objective will you achieve?	Precisely, **Ṣàìsàn** will train and tune a multi-class classification model on a corpus of structured relational data of LRTI features (symptoms) and the infection (diagnosis) that the symptoms led to. It will then deploy the model into portable mobile devices for on-device prediction through a conversational mobile app. The product will also support online inferences for 3rd party integrated tools, through a REST API. Ultimately, a sick person, caregiver, or clinician will be able to provide symptom data into **Ṣàìsàn,** over a conversational interface, and the product will use the multi-class classification model to diagnose if there is no LRTI or which particular infection the sick individual is suffering from. As an AI product Ṣàìsàn's business outcome is to : *Improve Access To Quality LRTI Diagnosis And Treatment, Even For Africa's Poor*

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 for this product are: 1. Increase the number of patients who are able to access LRTI diagnosis and treatment by >= 5% YoY 2. Reduce the average symptom presentation time to diagnosis time for LRTI patients to not more than 72 hours We can establish a baseline that affects these metrics by collecting data about access and time to diagnosis from a substantial sample size of LRTI patients and caregivers while building the dataset for this ML product. That said, the above metrics have employed sensible targets that can suffice in the absence of conclusive baseline data E.g it is reasonable to expect diagnosis to happen with 72 hours of LRTI symptoms.

Data

Data Acquisition Where will you source your data from? What is the cost to acquire this data? Are there any personally identifying information (PII) or data sensitivity issues you will need to overcome? Will data become	Ṣàìsàn will be sourcing data from LRTI patient records from public and private hospitals and primary health care centers, leveraging any Government agency or non-profit as a central proxy where possible. The expectation is that cost of access to such data will be negligible even if present, and can be substantially offset by strategic partnerships with nonprofits working to reduce deaths from preventable diseases in Africa,

available on an ongoing basis, or will you acquire a large batch of data that will need to be refreshed?	especially for women and children. Where there is still cost for data, Ṣàìsàn is open to building partnerships with stakeholder hospitals or agencies, such that there is a mutual exchange of value for the data, inclusive of early access to Ṣàìsàn or flexible pricing of the deployed service. Ṣàìsàn will not be collecting or using any sensitive or personally identifying data, and will mostly be using data of symptoms and features enabling it to detect presence of LRTI or not. Ṣàìsàn will acquire a large batch of data that will need to be refreshed not later than every 18 months
Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	Considering that LRTI are generally easy to diagnose and treat but are responsible for the most deaths in low income countries, it is clear that these are deaths of mostly poor people who had little, late, or no treatment. All of this means there is a high chance that our data, to a certain degree, might not be inclusive and diverse enough to represent the poorer or more remote people who were not able to access treatment in a hospital for their symptom and treatment data to be documented, thus potentially skewing our model's learning and inference to favor the economic, social, and cultural class who have relatively better access to hospitals.
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels versus any other option?	Ṣàìsàn will use a multi-class labelling scheme so that it can detect cases of no infection as well as the various common types of LRTI, which comes down to common types of bronchitis and pneumonia. The output labels of the classification model will be: 1. Normal (no LRTI) 2. Acute Bronchitis 3. Chronic Bronchitis 4. Bacterial Pneumonia 5. Viral Pneumonia The product is opting for the above multi-class classification labels to be as prescriptively diagnostic as much as possible.

	A binary classifier only predicting LRTI and no-LRTI will be too high level and less helpful (even to clinicians) when LRTI is detected, hence the need to further classify the infection into pneumonia or bronchitis. Driving for even further discrete classification (e.g from just pneumonia into bacterial and viral pneumonia) is aimed at being more prescriptive to the bulk of Ṣàìsàn's users who are patients with very limited access to health personne or caregivers with limited access to infrastructure. Using a multi-class labelling scheme allows the product to diagnose several infection types with a reasonable level of specificity, but the downside of such a broad and quite generic multi-class model is that it not specialised and might not attain the best quality attainable is it was focused on detecting, say, only pneumonia.

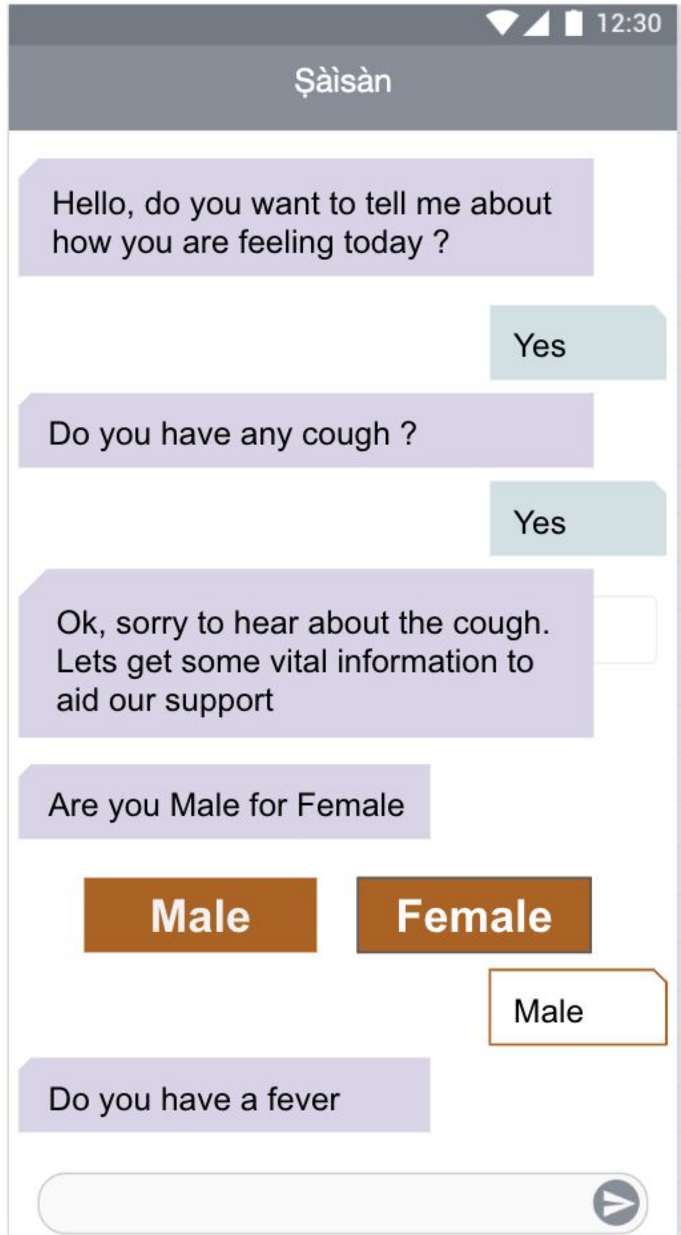
Model

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?	There were 916,851 global deaths from LRTI in 2016, which comes to 104 deaths every hour. Given it will be in the best interest of the business and stakeholders for Ṣàìsàn to drive early value and ultimately save lives, the product will build, tune and host its Machine Learning model on structured symptom data, using an Automated ML platform like Google AutoML Tables. The MVP will be built and managed by an initial small in-house team, which will expand to a robust Machine Learning and product team as time progresses.
Evaluating Results Which model performance metrics are appropriate to	To deliver high quality inference (diagnosis), Ṣàìsàn will optimize Precision and Recall as much as possible but will prioritise Recall over Precision. Ṣàìsàn's focal task is to provide quality primary

measure the success of your model? What level of performance is required?	diagnosis, allowing patients and clinicians to identify who is healthy or infected, and allowing for more targeted and robust secondary diagnosis before treatment. This implies that Ṣàìsàn, more than anything, needs to be able to surface as many instances of plausible infection, even if a few turn out to not have been infected after secondary diagnosis. This level of coverage and surfacing plausible cases of infection, over high precision of diagnosis to spot exact case of infection, is what prioritising Recall over Precision will achieve

Minimum Viable Product (MVP)

Design What does your minimum viable product look like? Include sketches of your product.	Ṣàìsàn's MVP will be a cross-platform Chatbot PWA (Progressive Web App) and a REST API that other tools can consult for diagnosis. The Chatbot PWA deploys a conversational user interface to capture LRTI symptoms which it then feeds into the on-device pre-trained model that the app would have downloaded and saved, and routinely updates Effectively, the app is able to deliver instant inference (diagnosis) through the Chatbot PWA, even without network connectivity, while the REST API allows 3rd party tools (e.g health management systems, USSD) to use the hosted online model for inference.

	
Use Cases What persona are you designing for? Can you describe the major epic-level use cases your product addresses? How will users access this product?	Below are the four (4) main personas and related use cases for **Ṣàìsàn**: 1. A sick adult using the product for quick primary diagnosis and advisory for LRTI 2. A caregiver, guardian or semi-skilled health care personnel using the product on behalf of a sick person, including children or the aged. 3. A clinician using the product for quick, quality,

	and scabale primary diagnosis of LRTI 4. An external tool presenting symptom data to the REST API and getting back diagnosis data
Roll-out How will this be adopted? What does the go-to-market plan look like?	The go-to-market plan will include 1. Strong online and offline presence, including the use of vibrant and persuasive digital and physical promotional materials, as well as socio-cultural influencers. 2. Strong and continued advocacy of the high death rate from LRTI and how Ṣàìsàn democratizes quality and timely diagnosis. 3. Strong and continued collaboration with Government, non-profits, influencers, hospitals and advocacy groups to ensure every smartphone-carrying parent, community, local health worker or health volunteer has agency of early diagnosis for themselves and their constituents / dependants 4. A community and an ambassador program where diagnosis and treatment success stories are highlighted and amplified such that survivors as well as their collaborators are spotlighted and incentivised.

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?	Ṣàìsàn will employ a tight feedback loop with hospitals, clinicians, and patients - incentivising 360 feedback, especially for cases where Ṣàìsàn's diagnosis was either very correct or very wrong. Ṣàìsàn will identify real world data differences (context) and learn from them by incentivising usage that reports them via the app, especially the additional context that makes such use case or data stand out. E.g If body mass index (BMI) is a highly correlated feature for predicting LRTI in children, is it impacted by

	their nutrition and does this affect the quality of inference if training data had more well fed children? In other words, can non-symptomatic real life data be used to improve the model ? Incremental improvements to the model will then be pushed to targeted subsets of users (A/B testing) to ensure the expected behaviour is attained before there is widespread rollout
Monitor Bias How do you plan to monitor or mitigate unwanted bias in your model?	Ṣàìsàn will actively monitor bias through periodic campaigns where select hospitals and clinicians are leveraged to conduct special use case diagnosis targeting identified or perceived underrepresented groups, and then reporting (with the relevant data) how Ṣàìsàn can evolve to better handle such forms of bias.