# Omdena local chapter Ethiopia: NLP for Ethiopian languages

## Problem statement

Amharic, one of the Ethiopian languages, is low a resource language( doesn't have lots of data to practice NLP).

Even though the language is the only one in the continent with its alphabet, the lack of NLP data and tools has been a bottleneck to leveraging the technology which is progressing at light speed in the rest of the world.

## Motivation

Ethiopia has usually termed the Land of Origins. As it has been archeologically proven our earlier ancestors of humans wandered off to the rest of the world from Ethiopia. Technically, we are all Ethiopians, and you are, in a sense, solving also your ancestor's problem

## Goals

Our goals for this project are:

* Compile and organize the Amharic dataset, ( existing, and collect new)
* Develop a model to classify Amharic text (at minimum we can iterate and go beyond just classification as Omdena encourages a bottom-up approach )

## Tasks

The overall project is divided into tasks and carried out in teams. Separate teams will be working and delivering these tasks to make the final output. Through discussion, these tasks can be reorganized, decomposed, unveiled, etc as we go along the project. There will be task leads in charge of the task in their team, their names will be announced in the slack channel one identified.

#Task 0: Knowledge: In charge of documenting the resources throughout the project

#Task 1: **Data collection**: provide data (collected, downloaded, translated, etc)

#Task 2: **Data preprocessing**: Everything from data to model including (cleaning, tokenization, lemmatization, etc)

#Task 4: **Model (ML/DL/ Transformer)** : building models for classification at first, then iterating and improving starting from simple models, towards more advanced ones if we manage to.

#Task 5: **Research team:** In general, out of its projects, Omdena aims to get some sort of tools such as a website that will be of use for the problem or a research paper that will be impactful. So, since this is the first project in Ethiopian NLP, it makes sense to meter/ measure it first. Therefore, we will aim to output a research paper on ‘the state of NLP in Ethiopia’. Our aim is ICAS, the international conference on the Advancement of Science and technology whose deadline for submission is on April 30, 2023: https://icast-conf.eai-conferences.org/2023/submission/?utm\_source=sendinblue&utm\_campaign=Call%20for%20Papers%20-%20ICAST%202023%20-%20n1&utm\_medium=email

## Time schedule:

The duration of the project is 4 weeks (yes, it is enough)!

**Member Rules & Guidelines**

* Main tasks are predefined and everyone will be free to join the task(s) they find best fit for them.
* New sub-tasks can be created if necessary after consulting the chapter lead(s)/product owner(s). (In the #tasks channel)
* Task leaders will self-identify or be identified for each task to present the progression in the weekly meeting
* English is the language to be used for any communication about the project (Slack,mail,...)

## FAQ

**How do I start?**

Apply to join whichever task you want. But it is recommended that you join the one you want to learn and contribute the most.

**What are the tools we use?**

We will use: Google colab, GitHub (I will add you so that you can contribute to the project github page), notion to keep track of the project progress

**Where do we get data?**

This is the task of the data collection team, but in general, there are news sites we can target for already labeled datasets, and also collected datasets.

**Who are the task leads?**

They will be announced once identified. The team is really self-organizing and anyone who thinks can do it can dm me, if it's still vacant…or apply to be a co-lead for a task.

**Any benefits?**

* Learn how real-world data science projects are solved by getting hands-on experience
* Develop your DS/AI skills and showcase them through open-source solutions
* Network with chapter members and leads from 50+ different countries
* Preferential selection to the Omdena core challenges and future access to Omdena members' [benefit platform](https://omdena.com/collaborator-tiers/). (access to paid version )