Machine Learning Task

Introduction

Many countries speak Arabic; however, each country has its own dialect, the aim of this task is to build a model that predicts the dialect given the text.

Guidelines

- You are given a dataset which has 2 columns, id and dialect.
- Target label column is the “dialect”, which has 18 classes.
- The “id” column will be used to retrieve the text, to do that, you need to call this API by a **POST** request. https://recruitment.aimtechnologies.co/ai-tasks
- The request body must be a **JSON** as a **list of strings**, and the size of the list must **NOT** exceed **1000**.
- The API will return a dictionary where the keys are the ids, and the values are the text, here is a request and response **sample**.

Request body sample:

[

"1055620304465215616", "1057418989293485952"

]

Response sample:

{

"1055620304465215616": "@MahmoudWaked7 @maganenoo ايه اسمه جمبه الي والمركز بهيج مركز مطروح طريق في😂😂",

"1057418989293485952": "@mycousinvinnyys @hanyamikhail1 ده المحل فين الهالوين شكولاته دي متهيالي"

}

- The dataset and the dialect identification problem were addressed by Qatar Computing Research Institute, moreover, they published a paper, feel free to get more insights from it, https://arxiv.org/pdf/2005.06557.pdf
- You must use python.
- Choose the most suitable data pre-processing techniques.

- Train **two** models, a machine learning model and a deep learning model, then compare the results (You are free to choose any ML algorithm and any DL architecture)
- Use Flask or FastAPI or any suitable web framework to deploy the model locally.
- Push the source code to a GitHub repo, it's preferred to be a well-documented repo.

Deliverables

1. A GitHub repo link, which has the following structure:
    a. Data fetching script/notebook
    b. Data pre-processing script/notebook
    c. Model Training script/notebook
    d. Deployment script/notebook
    e. Any additional files/documentation you need
2. A PowerPoint presentation summarizing your approach, data pre-processing, model architecture, evaluation metrics and results. (**Max 8 slides**)

Notes

- Make sure that the GitHub repo is public.
- In the assessment phase, you'll be asked to run your models locally, furthermore, you'll be asked in any technical decision/implementation you've made, so be well prepared, and avoid over-complicated approaches you don't fully grasp.
- The deadline is 12 days from the day you've received the task, specific dates are written in the email.
- Early submission doesn't affect your grade, take your time.
- To submit the task, reply to the email you've received with the deliverables.