

## Ticketing system Task

You are tasked with implementing a text-based machine learning classifier. The dataset is included in the email attachment. Note that the data is unlabeled, and part of the assignment involves labeling the data using a technique of your choice. Labels should be derived exclusively from the "Request Detail" or "Notes" fields. Ideally, the labels should highlight discrepancies between the "Request Detail" and "Notes" features in Arabic, but English labels are acceptable as well. Evaluation will be based on the range of coding techniques used, the professionalism of the results presentation, and the model's performance.

### Technical Requirements:

#### 1. Preprocessing

- 1) Data cleaning (e.g. remove punctuation, stemming, stopwords, etc.). please the emails thread and the bilingual nature of the contents.
- 2) Exploratory Data Analysis (EDA).
  - a. Document the 3-5 major insights from the data.
  - b. Show words most used in "Request Detail" feature and present it in a nice way.
  - c. Study correlation
  - d. Show distributions and demotions of data
  - e. Show statistical abilities and tendency metrics
- 3) Dimensionality reduction methods PCA, correlation analysis, etc
- 4) Feature importance (using algorithms).
- 5) Feature engineering.
- 6) Extract at least two columns of labels, such as from the features ("Request Detail") or ("Notes"), by identifying key terms that can serve as labels for the data. For example, classify tickets as "in scope" or "out of scope," or tasks as "challenging" or "non-challenging," etc. You are free to choose any labels that you think can be extracted from the data. Ensure that at least two columns of labels are created.
- 7) Make sure to use cutting edge NLP techniques as well as legacy solid techniques.

#### 2. Modeling

Provide results of the best three algorithms "classifiers" (you are encouraged to use stacking, deep learning algorithms, and transformers), and show different results from each algorithm based on parameter tuning.

#### 3. Evaluation

1. Suggest the best metric to evaluate the model performance.
2. Provide at least the following three matric for each result: Accuracy, Recall, and Precision

### Deliverables

After successfully implementing the above requirement, you are to submit two files:

- A report documenting your implementation to the above requirements.
- The source code (preferably written in Python, .ipynb) with documentation.

**Additional Credit:** Candidates who use modern software development tools (e.g., integrated development environments, automation scripts) and version control systems (e.g., Git) will receive extra credit. Moreover, leveraging the latest NLP techniques (e.g., transfer learning, pre-trained language models) is strongly encouraged.