**Documentation**

**Text classifier project**

1. Project Overview
2. Dataset Description
3. Tools and Technologies
4. Data Preprocessing
5. Model Development
6. Results and Analysis
7. Deployment
8. Conclusion

**1)-Project Overview**

The aim of this starting project is to build a classic text classifying model to detect if the SMS is spam or not . In this project we will try to train a traditional model without using any pretrained model from hugging face or any other repository.

The main objective is to learn how a traditional classifier is built , how can we finetune the classifier to get better results also learn about the various proprocessing techniques used to improve the final results .

The work will be done in a professional setup with regular 1 hour meetings on Mondays.

**2)- Dataset Description**

[**https://archive.ics.uci.edu/dataset/228/sms+spam+collection**](https://archive.ics.uci.edu/dataset/228/sms+spam+collection)

**https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset**

The dataset is a collection of 5574 text messages and its labelled ham or spam . The language is english . The dataset has a lot of punctuation and other stop words which needs to be filtered out.

Eg :

**Data:**