**Project report**

**Step 1 : data importation and label encoding :**

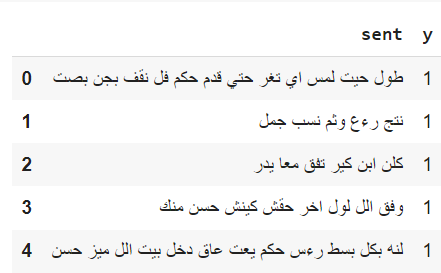
**results**

****

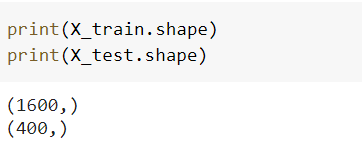
**Step 2 : data cleaning and preprocessing :**

**In this step I have removed diacritics , longations , english words, repetitions, spaces/tabulations/new lines …, digits, and stopwords**

**Finally I used ISRIStemmer to stem the data**

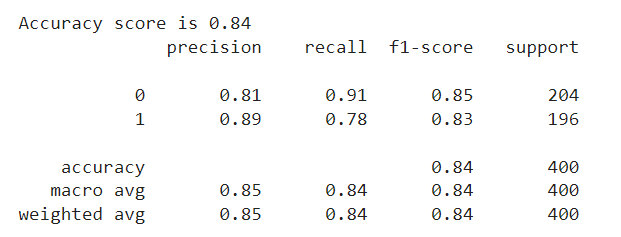
****

**Step3 : data splitting :**

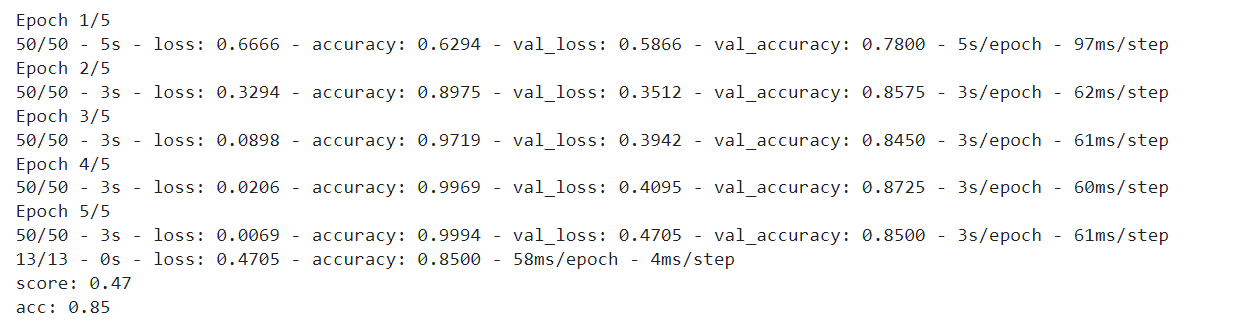
****

**Step3 : Modeling:**

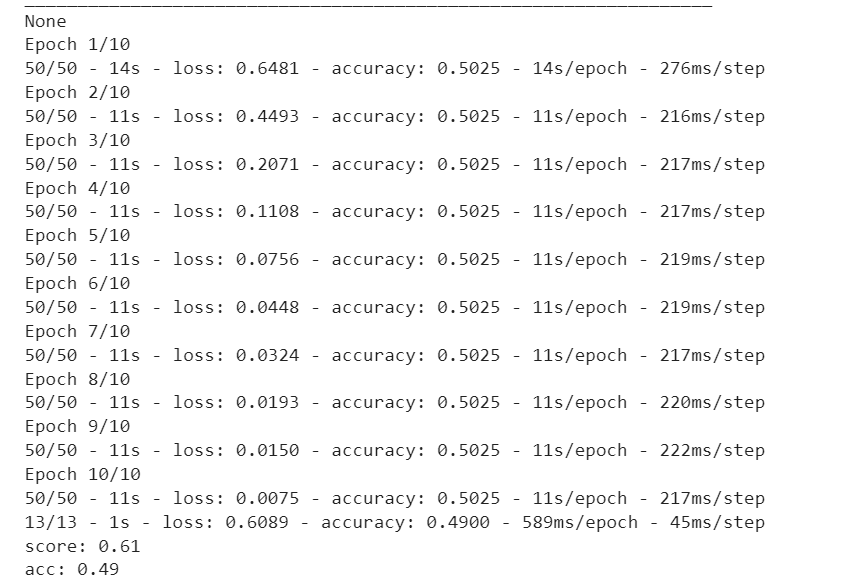
**First model : TF-IDF + Support vector classifier**

****

**Second model : CNN**

****

**Third model : LSTM**

****

**Comparaison**

The LSTM result after the training and the validation step gave us a poor accuracy compared to the CNN, the model was even slower. Thus this can be improved by tuning the hyperparameter, it can be even faster as well if we combine it with another model including CNN itself.

SVC gave us some interesting results with 0.83 accuracy

CNN outperformed the 2 models with 0.85 even without hyperparameter tuning.