Natural Language Processing (ENSAE 3A-MS)-Project

ENSAE 2021/2022

Professeur: Benjamin Muller

Mastère Spécialisé - Data Science

Oscar Fossey Jeffrey Verdière

Lab small group professor: Gaël Guibon

Pôle Emploi Job offer classification

Contents

1 The problematic

[1] Pôle emploi receives around one thousands of job offer per day on their plateform and new to classify according to very rigid nomenclature.

B			**Arts et Façonnage d'ouvrages d'art**	
B	**11**		**Arts plastiques**	
B	**11**	**01**	**Création en arts plastiques**	
B	11	01	Aquarelliste	11101

Figure 1: Nomenclature of a job offer

As show on Figure 1, each job offer has to be classified following a ROME code with different levels. The first level is the letter, the second level is a job category which corresponds to a two digits number.

The aim of our project: is to find the best predictor for a job offer description which predicts the first level of the ROME Code. To be accurate, we want to predict for a chain of character the label between A to N (14 labels).

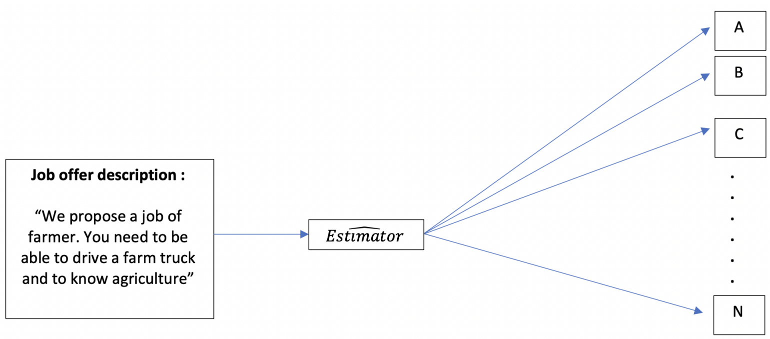


Figure 2: Data pipeline of our project

2 Methodology

2.1 Data collection

We collected the data on pôle emploi website API. It enables to collect 64 402 job description texts with the ROME code associated.

Table 1: Data Collection Organization

The total number of samples is	64 402 samples
The train dataset is	38640 samples
The validation dataset is	12 881 samples
The test dataset is	12 881 samples

Also our data set, has the following repartition for each labels.

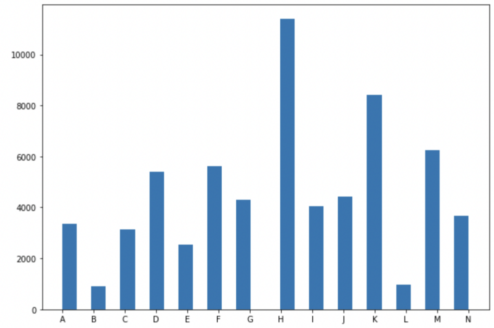


Figure 3: Repartition for the labels of our Data Set

2.2 Model litterature review for text classification

According to [3], the most common models for text classification are TF-IDF, LSTM and Transformers.

2.3 Models training and benchmarking

[2] explains that the best methods to solve the unbalanced data set problem released in our case with the label H are oversampling, undersampling or random weighted sampling.

For sampling techniques, we have tested four sampling techniques to build our data loaders and we desplayed the results in the following tab.

Table 2: Model Accuracy on validation test for different sampling during training

Model	Undersampling	Oversampling	Weighted sampling	Normal Data Set
TD-IDF	Accuracy: 64%	Accuracy: 69%	Accuracy: -	Accuracy: 47%
Camembert	Accuracy: -	Accuracy:-	Accuracy:72%	Accuracy: 76%
LSTM	Accuracy: -	Accuracy:-	Accuracy:-	Accuracy: 69%

2.4 Model testing

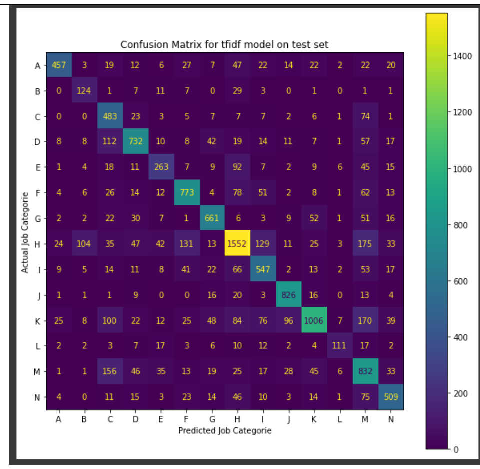
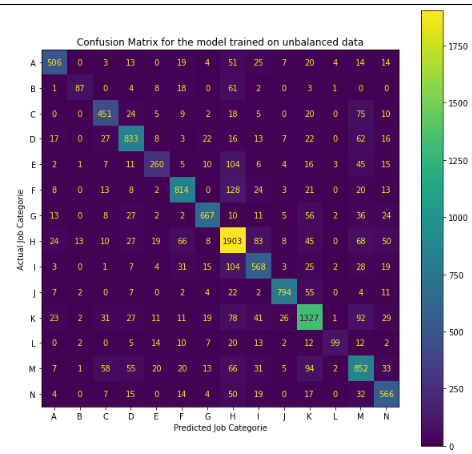
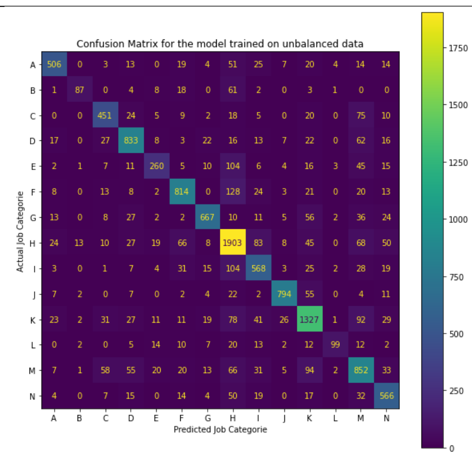
Model	Accuracy	Confusion matrix
TD-IDF	Accuracy: 69%	
Camembert	Accuracy: 76%	
LSTM	Accuracy: 70%	

Table 3: Models Accuracy on validation set and confusion matrix

2.5 Model picked

Despite the large amount of parameters for the camembert models, it is the one which performs the best and can draw predictions for on text in less than a second which is fast enough. So we keep this model for our product delivery.

3 Product delivery

The product delivery is github link: https://github.com/oscarfossey/NLP-Job-classifier-based-on-description.

The data and the models were registered on huggingface cloud plateform for convinient reasons. Link to data: $https://huggingface.co/datasets/oscarfossey/NLP_{P}ole_{e}mploi/tree/main$
Link to models: $https://huggingface.co/oscarfossey/job_{c}lassification$

On this repository, you can find all the used notebooks. Particulary, we divided our works in three main differents files: scrapping, trainings of the models and pipelines.

The "main" notebook should be run using a google collab GPU.

References

[1] Romarik Le Dourneuf. “Le nombre d’offres explose sur Pôle Emploi, voici les secteurs qui recrutent le plus”. In: (1982).

[2] Gao. “Data Augmentation in Solving Data Imbalance Problems”. In: (2020).

[3] Cambria Nanyang Nikzad Chenaghlu Gao Minaee Kalchbrenner. “Deep Learning Based Text Classification: A Comprehensive Review”. In: (2020).

List of Tables

List of Figures