

# Projet Science des données

## Contents

<b>Remerciements</b>	<b>1</b>
<b>Analyse du problème</b>	<b>2</b>
Classification . . . . .	2
Fairness vis à vis de <b>Gender</b> . . . . .	2
<b>Analyse du jeu de données :</b>	<b>3</b>
Structure & Complexité . . . . .	3
<b>Réalisation du projet</b>	<b>3</b>
Méthodologie et découpage du projet : . . . . .	3
Prétraitement . . . . .	3
Augmentation des donnees . . . . .	3
Représentation . . . . .	4
Classification . . . . .	4
Deep learning . . . . .	4
Nos architectures . . . . .	4
Différents Modeles . . . . .	4
<b>Résultats &amp; Performances</b>	<b>4</b>
<b>Conclusions</b>	<b>4</b>
<b>References</b>	<b>4</b>

## Remerciements

## Analyse du problème

Ce projet s'inscrit dans le cadre du Défi IA, organisé par l'INSA Toulouse. Cette édition du Défi IA porte sur le traitement automatique du langage.

La tâche est simple : attribuer la bonne catégorie d'emploi à une description de poste. Il s'agit donc d'une tâche de classification multi-classes avec 28 classes.

Les données ont été extraites de CommonCrawl. Les données sont donc représentatives de ce que l'on peut trouver sur la partie anglophone de l'Internet, et contiennent donc un certain biais. L'un des objectifs de ce concours est de concevoir une solution qui soit à la fois précise et équitable.

## Classification

La précision sera mesurée par le macro F1 score :

$$F_{1_{macro}} = \frac{1}{C} \sum_{k=1}^C 2 \cdot \frac{\text{precision}_k \cdot \text{recall}_k}{\text{precision}_k + \text{recall}_k}$$

\$\$\$\$

où la précision est définie tel quel :

$$\text{precision}_k = \frac{tp}{tp + fp}$$

et le recall est défini tel quel :

$$\text{recall}_k = \frac{tp}{tp + fn}$$

L'équité sera mesurée par le Disparate Impact:

$$DI_k = \frac{\mathbb{P}(\text{Job} = k | \text{Gender} = F)}{\mathbb{P}(\text{Job} = k | \text{Gender} = M)}$$

## Fairness vis à vis de Gender

## Analyse du jeu de données :

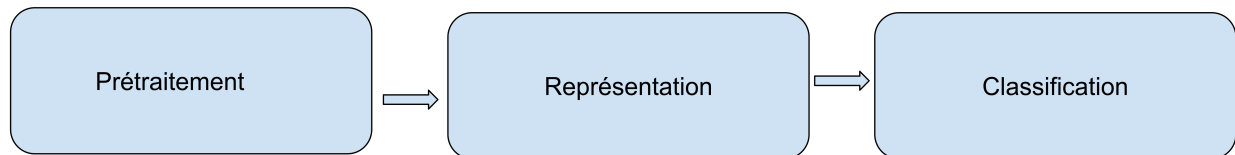
### Structure & Complexité

Extrait de descriptions : - Diversité des descriptions : - cas normaux représentatifs - cas pathologiques (classes où il y a peu de descriptions, fautes d'orthographe)

## Réalisation du projet

### Méthodologie et découpage du projet :

Nous avons séparé notre projet en 3 parties : la préparation et le prétraitement, la représentation et la classification.



### Prétraitement

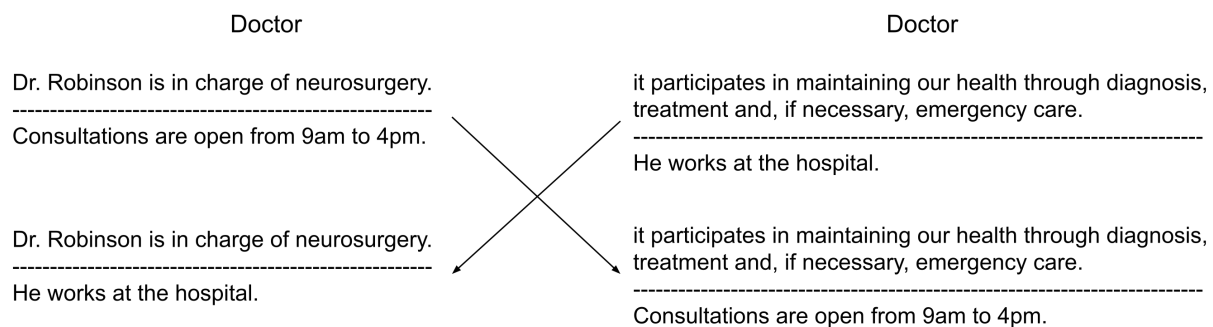
Nous avons utilisé plusieurs prétraitements : lesquels ?

#### Augmentation des données

Pour rendre notre modèle de classification plus robuste, nous avons augmenté le jeu de données. Nous avons donc utilisé plusieurs techniques (Ces techniques sont inspirées de (BOURDOIS 2020)):

- Augmentation par crossover :

Nous sélectionnons des descriptions faisant référence au même métier. Nous découpons les descriptions en phrases et nous créons des exemples en mélangeant les phrases des deux descriptions.



- Augmentation par substitution/insertion
- Augmentation par rétro-traduction :

Nous traduisons une description dans une langue étrangère (en chinois, par exemple) puis nous retraduisons dans la langue originale.



This is the description #XXX

这是描述 #XXX

This is a description of #XXX

- Augmentation par échange

## Représentation

Comme representation nous avons utiliser plusieurs vectoriseurs :

- CountVectorizer
- TfidfVectorizer
- WordEmbedding

## Classification

### Deep learning

Parmi les modeles de langage les plus connu nous retrouvons :

- Modele recurrent : LSTM , BLSTM
- Modeles Transformer Based :
  - BERT
  - GPT2 - GPT3
  - T5

### Nos architectures

- Rappel sur BERT

BERT pour **Bidirectional Encoder Representations from Transformers** (Devlin et al. (2019)) est un modèle de langage. L'architecture BERT est un empilement d'encodeurs de Transformers (Vaswani et al. (2017)).

- Implémentation et fine-tuning de BERT

### Differents Modeles

## Résultats & Performances

## Conclusions

## References

BOURDOIS, Loïck, trans. 2020. "L'AUGMENTATION DE DONNEES EN NLP." *Loïck BOURDOIS*. <https://lbourdois.github.io/blog/nlp/Data-augmentation-in-NLP/>.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv:1810.04805 [Cs]*, May. <http://arxiv.org/abs/1810.04805>.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. “Attention Is All You Need.” *arXiv:1706.03762 [Cs]*, December. <http://arxiv.org/abs/1706.03762>.