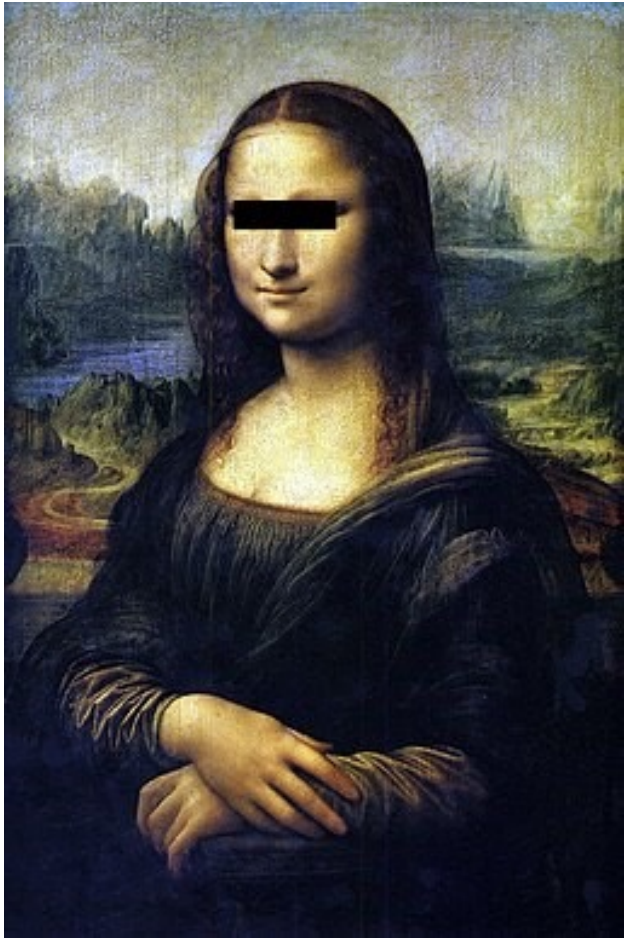


Algorithme de pseudonymisation des compte-rendus cliniques

Bernoulli Lab

17 juin 2024

Perceval Wajsbürt, PhD
Datascientist

DSN, AP-HP

Contexte

Constat

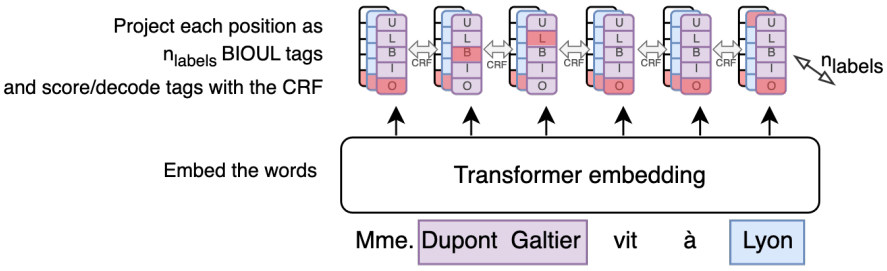
- Les textes de l'EDS contiennent de nombreuses **informations identifiantes**
- La CNIL demande aux EDS de **suivre l'état de l'art** pour retirer le plus de ces infos
- On sait que le **ML** permet de meilleures performances de pseudonymisation

Enjeux

- Développer un algorithme de pseudonymisation des CRs cliniques à l'état de l'art
- **Mise en prod.** pour l'intégration quotidienne des textes cliniques à l'EDS
- **Valider** ses résultats ⇨assurer une transparence vis-à-vis des **risques résiduels**
- **Open-sourcer** les développement réalisés pour ouvrir le projet aux collaborations

Modèle hybride

1. **Deep learning**: Transformer + CRF



2. **Règles**
   - Statiques (regex ++)
   - Dynamiques: recherche à partir des données structurées
3. **Fusion** des résultats ML + Règles

Implémentation


```python
import edsnlp

nlp = edsnlp.blank('eds')
nlp.add_pipe(
  'eds.ner_crf', # à entrainer
  config={
    "embedding": {
    "@factory": "eds.transformer",
    ...}
})
nlp.add_pipe('eds.dates')
nlp.add_pipe('eds_pseudo.adresses')
nlp.add_pipe('eds_pseudo.misc_rules')
nlp.add_pipe('eds_pseudo.context')
```

D'où viennent les textes du jeu annoté ?

Répartition documents:

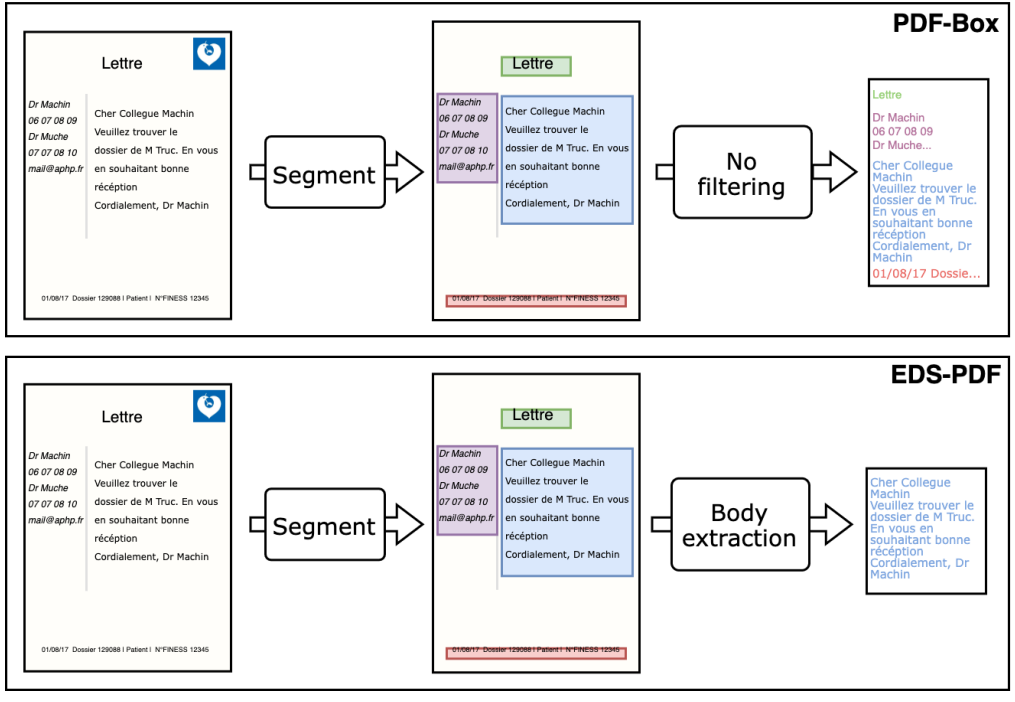
- Train: documents post-2017
- Test: sampling aléatoire

Répartition extractions PDF train:

- 90% edspdf
- 10% pdfbox

Répartition test:

- 50% edspdf
- 50% pdfbox (mêmes docs)

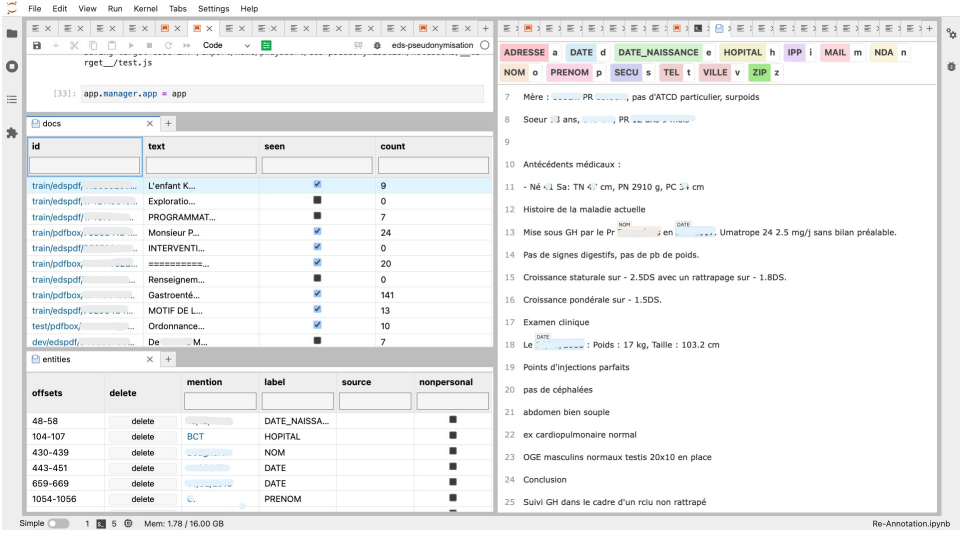


Phase d'annotation

Campagne en deux étapes

- Première phase en dec 21 – jan 22
- Vérification en dec 22 – jan 23

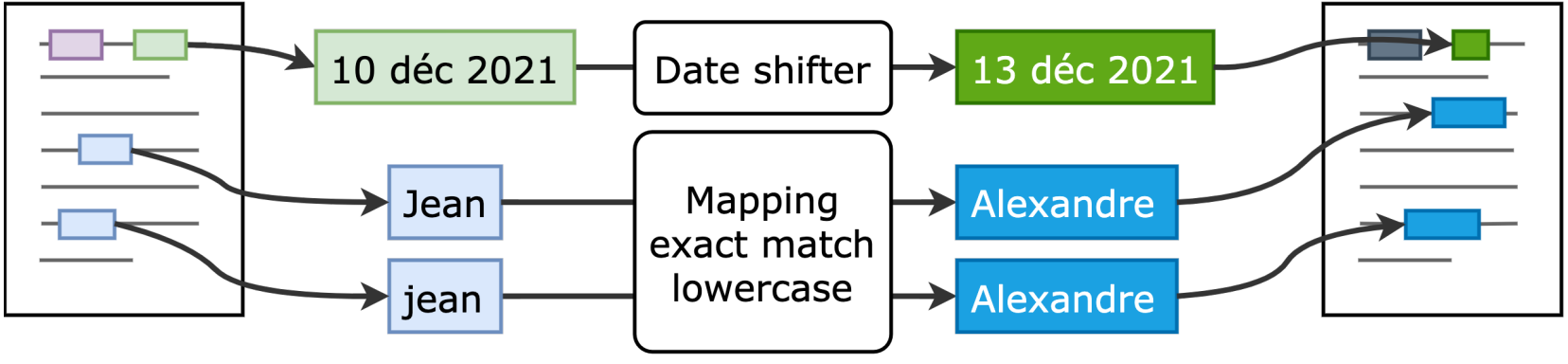
	ENTS	DOCS
train/edspdf	27135	3025
train/pdfbox	16071	348
dev/edspdf	1615	200
dev/pdfbox	967	22
test/edspdf	**3491**	348
test/pdfbox	**16793**	348



2e phase d'annotation (metanno)

Remplacement des entités identifiantes

On ne masque pas mais on remplace, ce qui permet de dissimuler les erreurs parmi les données correctement remplacées.



- Remplacement par exact match en minuscules
- ⚠ Il faut décaler les indices si détection d'entités en amont

Métriques

- Niveau du mot (pas exact match)
- Redact: caviardage en % de mots
- Full: % de docs avec redact à 100%

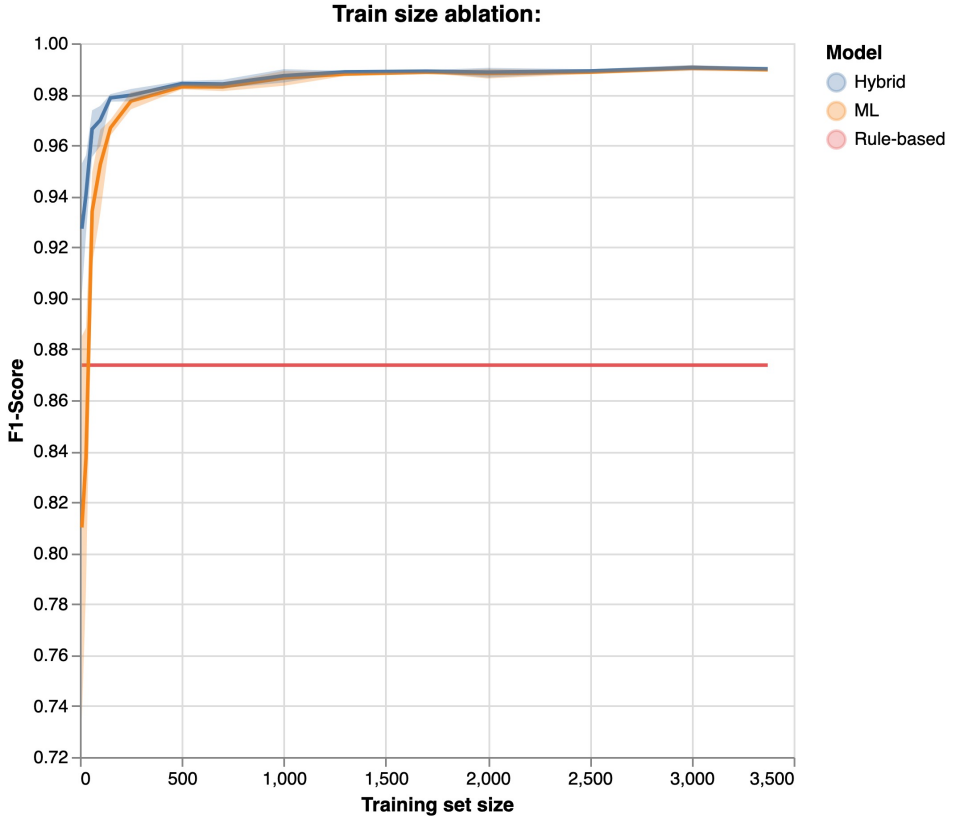
Résultats

- 99 en métrique F1
- et 99.4 en redact
- 86.2% de docs entièrement caviardés
- VISIT ID difficile car formats divers et source dans les données structurées

Label	P	R	F1	Redact	Full
ADDRESS	99.0	98.4	98.7	98.5	98.4
BIRTHDATE	98.2	98.2	98.2	99.8	99.7
CITY	98.0	98.8	98.4	98.8	98.2
DATE	99.7	99.3	99.5	99.6	95.4
EMAIL	98.9	99.9	99.4	99.9	99.9
FIRSTNAME	98.8	98.4	98.6	99.4	97.4
LASTNAME	98.6	98.6	98.6	99.6	97.2
NSS	88.0	98.9	93.1	100.	100.
PATIENT ID	99.0	94.0	96.4	98.2	99.1
PHONE	99.6	99.7	99.7	99.7	99.0
VISIT ID	91.5	89.4	90.4	90.4	98.3
ZIP	99.9	99.9	99.9	99.9	99.9
ALL	99.0	98.9	**99.0**	99.4	86.2

Quelle performance si on a moins de documents pour entrainer ?

- Modèle ML est très vite meilleur que les règles
- La performance croît avec le nombre de documents
- On peut s'arrêter autour de 1500 documents



Quel effet de la méthode d'extraction des PDFs ?

Rappel: EDS-PDF retire environ **80%** des entités identifiantes !

PDF extraction	P	R	F1	Redacted	Full
edspdf	99.1 ± 0.1	98.8 ± 0.1	98.9 ± 0.1	99.2 ± 0.1	**93.1 ± 1.0**
pdfbox	99.1 ± 0.0	98.9 ± 0.2	99.0 ± 0.1	99.4 ± 0.1	**75.7 ± 3.0**

On observe :

- la même performance selon les métriques micro-moyennées (P, R, F1, Redact)
- mais le nombre de documents entièrement caviardés gagne **18.6%** !

Quel effet du choix du modèle d'embeddings ?

- **camembert**: Pré-entrainement sur corpus général français (OSCAR/FR)
- **finetuned**: Finetuning de Camembert sur 28M CR originaux, 1 epoch
- **scratch**: Pré-entrainement sur 28M CR pseudonymisés, ~15 epochs

Model	P	R	F1	Redact	Full
finetuned	97.8	97.7	97.8	98.2	75.5
camembert	96.8	96.9	96.8	97.4	68.9
scratch	97.3	97.2	97.3	97.6	69.0

- **finetuned** est significativement meilleur
- Et plus rapide à entrainer que **scratch**

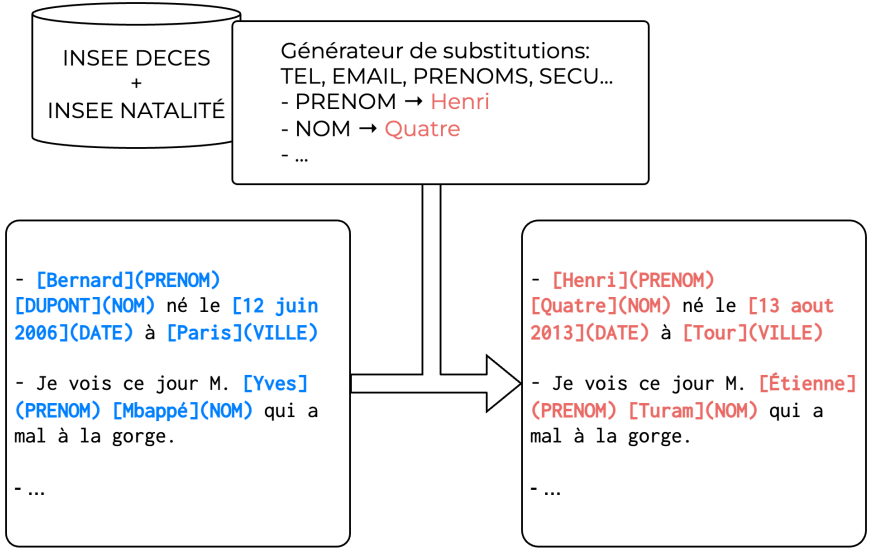
NEW !

Un modèle public

Modèle : hf.co/AP-HP/eds-pseudo-public

Démo : eds-pseudo-public.streamlit.app

Génération d'exemples fictifs :



~ **230** patrons fictifs

~ **35000** documents fictifs

Label	P	R	F1	Redact	Full
ADDRESS	98.2	96.9	97.6	97.6	96.7
BIRTHDATE	97.5	96.9	97.2	99.3	99.4
CITY	96.7	93.8	95.2	95.1	91.1
DATE	99.0	98.4	98.7	98.8	85.9
EMAIL	96.1	99.8	97.9	99.8	99.7
FIRSTNAME	93.5	96.6	95.0	99.0	93.2
LASTNAME	94.4	95.3	94.8	98.2	89.5
NSS	88.3	100	93.8	100	100
PATIENT ID	91.9	90.8	91.3	98.5	99.3
PHONE	97.5	99.9	98.7	99.9	99.6
VISIT ID	92.1	83.5	87.6	87.4	97.2
ZIP	96.8	100	98.3	100	100
ALL	97.0	97.8	**97.4**	98.8	63.1

Questions ?

EDS-PSEUDO — MODÈLE PRIVÉ VS MODÈLE PUBLIC

Label	P	R	F1	Redact	Full
ADDRESS	99.0	98.4	98.7	98.5	98.4
BIRTHDATE	98.2	98.2	98.2	99.8	99.7
CITY	98.0	98.8	98.4	98.8	98.2
DATE	99.7	99.3	99.5	99.6	95.4
EMAIL	98.9	99.9	99.4	99.9	99.9
FIRSTNAME	98.8	98.4	98.6	99.4	97.4
LASTNAME	98.6	98.6	98.6	99.6	97.2
NSS	88.0	98.9	93.1	100.	100.
PATIENT ID	99.0	94.0	96.4	98.2	99.1
PHONE	99.6	99.7	99.7	99.7	99.0
VISIT ID	91.5	89.4	90.4	90.4	98.3
ZIP	99.9	99.9	99.9	99.9	99.9
ALL	99.0	98.9	**99.0**	99.4	86.2

Label	P	R	F1	Redact	Full
ADDRESS	98.2	96.9	97.6	97.6	96.7
BIRTHDATE	97.5	96.9	97.2	99.3	99.4
CITY	96.7	93.8	95.2	95.1	91.1
DATE	99.0	98.4	98.7	98.8	85.9
EMAIL	96.1	99.8	97.9	99.8	99.7
FIRSTNAME	93.5	96.6	95.0	99.0	93.2
LASTNAME	94.4	95.3	94.8	98.2	89.5
NSS	88.3	100	93.8	100	100
PATIENT ID	91.9	90.8	91.3	98.5	99.3
PHONE	97.5	99.9	98.7	99.9	99.6
VISIT ID	92.1	83.5	87.6	87.4	97.2
ZIP	96.8	100	98.3	100	100
ALL	97.0	97.8	**97.4**	98.8	63.1