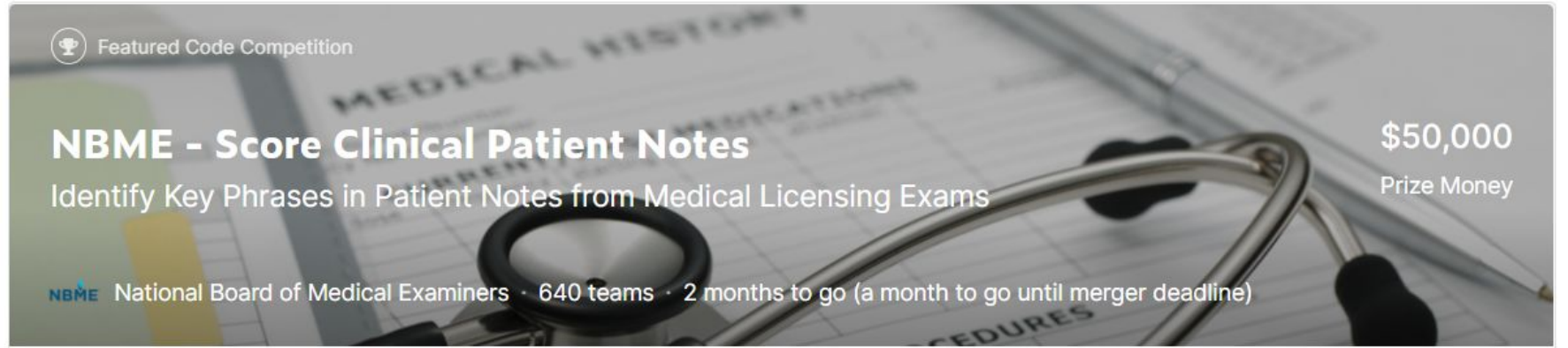




Projet n° 8: Participez à une compétition Kaggle

Céline Mendola

Présentation de la compétition: généralités



Featured Code Competition

NBME - Score Clinical Patient Notes

Identify Key Phrases in Patient Notes from Medical Licensing Exams

\$50,000
Prize Money

NBME National Board of Medical Examiners · 640 teams · 2 months to go (a month to go until merger deadline)

But : Développement d'un outil qui permet d'aider à l'évaluation d'un examen de médecine (United States Medical Licensing Examination[®] (USMLE[®]).

Présentation de la compétition: généralités

En quoi consiste cet examen ?

On a 10 cas cliniques (représentés par des patients). Chaque cas clinique possède entre 9 et 17 caractéristiques.

Les candidats doivent rédiger des notes de patient pour ces cas cliniques.

Notre but : Créer un modèle permettant de retrouver les caractéristiques du cas clinique dans la note de patient

Présentation de la compétition: généralités

Exemple

Note de Patient (donnée)

17 y/o m came to the clinic c/o heart pounding. started 2-3 mo ago. it started suddenly. does not recall any triggering events. it comes and goes, it happened 5-6 times since it started. it lasts 3-4 min, after than just goes away. he has also experiencing sob, pressure on her chest when he has this attack. he is a college student, experiencing some stress recently. \r\n denies cough, chest pain. \r\n ros neg except as above. \r\n pmh none. meds aterol, for his studies, sharing w his roommate. nkda. \r\n psh/ hosp/ travel/ trauma none. \r\n fh mom has thyroid problems. \r\n sh sex active w girlfriend, no stds, using condoms. smoke none. etoh only weekends. drug only once, 1 mo ago.

Caractéristiques du cas clinique (donnée)

Localisation ou "span" (à prédire)

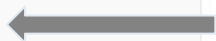
family history of thyroid disorder	→	[532 556]
family history of myocardial infarction	→	[]
chest pressure	→	[263 284]
17 year	→	[0 6]
male	→	[7 8]
intermittent symptoms	→	[131 145, 150 168]
...		...

Présentation de la compétition: évaluation

Un fichier de soumission à déposer. Après évaluation du fichier de soumission, notre modèle est évalué sur un fichier test “caché”.

id	location
Unique identifier for this instance, a feature within a patient note.	Character spans indicating the location(s) of the feature within the note.
5 unique values	[null] 40% 0 100 20% Other (2) 40%
00016_000	0 100
00016_001	
00016_002	200 250;300 400
00016_003	
00016_004	75 110

Note de patient n°00016,
feature n°002 (cas clinique 0)



Présentation de la compétition: évaluation

Mesure de la performance du modèle : Calcul du f1-score sur l'ensemble des caractères.

$$f1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Chaque caractère est défini comme :

TP s'il est dans la localisation réelle et dans la prédiction

FN s'il est dans la localisation réelle mais pas dans la prédiction

FP s'il est dans la prédiction mais pas dans la localisation réelle

Présentation de la compétition: évaluation

Exemple :

	Localisation réelle de la caractéristique	Localisation prédite
Ecriture en span	0 3 ; 3 5	2 5 ; 7 9 ; 2 3
Ecriture détaillée	0 1 2 3 4	2 3 4 7 8

$$TP = \#\{2\ 3\ 4\} = 3$$

$$FN = \#\{0\ 1\} = 2$$

$$FP = \#\{7\ 8\} = 2$$

$$f1 = 3 / (3 + 0.5 * (2+2)) = 0.6$$

Présentation de la compétition: évaluation

Exemple :

	Localisation réelle de la caractéristique	Localisation prédite
Ecriture en span	0 3 ; 3 5	2 5 ; 7 9 ; 2 3
Ecriture détaillée	0 1 2 3 4	2 3 4 7 8
Forme sortie du modèle	1 1 1 1 1 0 0 0 0 0 0 ...	0 0 1 1 1 0 0 1 1 0 0 ...

$$TP = \#\{2\ 3\ 4\} = 3$$

$$FN = \#\{0\ 1\} = 2$$

$$FP = \#\{7\ 8\} = 2$$

$$f1 = 3 / (3 + 0.5 * (2+2)) = 0.6$$

Les pistes

De nombreux kagglers qui ont partagé leur notebook s'appuient sur des réseaux pre-entraînés de transformers, qui sont un type de réseau de neurones qui ont révolutionné le NLP.

En fait, ces transformers se basent sur le mécanisme du self attention qui permet de garder une notion de contexte entre les mots, ce que nous n'avions pas avec les bags of words ni même les word embedding (type word2vec ou Glove).

L'un des plus connus est le réseau pré-entraîne BERT (Bidirectionnal Encoder representations Transformers) développé par google en 2018. Il a été entraîné sur Toronto book corpus et wikipedia.

Tâches qui ont permis d'entraîner BERT : masked language modeling et next sentence prediction.

Les pistes

La librairie transformers distribuée par HuggingFace est aussi très utilisée. HuggingFace propose des outils pouvant résoudre des tâches de NLP qui peuvent se rapprocher de notre problème comme :

- Le “question answering” : A partir d’une question et d’un contexte, on souhaite trouver la partie du contexte (sous forme de span) qui correspondra à la question
- Le “Named entity recognition” : Permet de détecter les parties d’un texte correspondant à certaines entités (lieu, personnes ...)

Les pistes

Dans le cadre de notre travail, nous nous sommes basés sur un kernel d'un candidat.

<https://www.kaggle.com/code/iamsdt/pytorch-bert-baseline-nbme/notebook>

Il se base sur le modèle pré-entraîné BERT ainsi que l'ajout de couches fully-connected. Nous allons par la suite détailler l'amélioration de ce kernel.

Présentation des données

patient_notes.csv : 42 126 notes de patient avec un identifiant et le numéro du cas clinique.

pn_num	case_num	pn_history
5000	22107	2 44 yo F c/o irregular periods for 3 years\r\nLMP : 2 month ago, cyces every 3 week t o every 4 months, lasting about 2-6 days, flow ranging from light to heavy, with abdominal discomfort, breast tenderness 2 days before last menstrual period. no dyspareunia, no postcoital bleeding, no vaginal bleeding, no vaginal discharge. no change in weight appetite. With occ hot flashes, vaginal drying and pruritus, sweating, palpitations. . no change in voice, no breast dischargfge, no visional changes. \r\nmed: HCTZ. NKDA. pmh: HTN 6 Y. Menarche: 14 . G2P2, uncomplicated vaginal birth.\r\nno smoking, occ ETOH, no rec drugs, no STD, sexually active with husband\r\n
5001	22108	2 This is a 44 year old G2P2 female presenting with a 3 year history of irregular periods. She notes that her periods have been unpredictable, and estimates she has been haivng 5-6 periods per year. The flow varies in heaviness. She also notes an episode 1 week ago of waking up drenched in sweat. She notes hot flashes and sweating in the daytime as well. She has had vaginal dryness during sex for the past year. She denies cramping, changes in weight, changes in hair/skin/nails, concentration or mood, or hot or cold intolerance. \r\nPMHx: HTN\r\nMedications: HCTZ\r\nSurgeries: None\r\nFamily history: noncontributory\r\nOb/gyn: 2 children, vaginal deliveries without complications. Pap smear 1 year ago was normal. Copper IUD was placed 4 years ago. Sexually active with husband.\r\nSocial: Non-smoker, drinks rarely, no drug use.
5002	22109	2 44 yo f G2P2A0, complains of irregular periods for the past 3 years. her periods can appear between 3 weeks to 4 months, varies between 5-6 pads on heavy days to 2-3 pads on low days and can last 2-6 days. she further mentions hot flushes and vaginal dryness. denies menstrual periods or abdominal periods. denies spotting between periods. unknown fibroid or any gynecological abnormality. PAP smear - a year ago and normal.mentions early morning awakening, denies mood change. she did not see an OBGYN for the past 3 years. she denies changes in bowel, GI, or urinary symptoms, and no change in weight or appetite. \r\nROS: None except above mentioned\r\nMEDS:HCTZ\r\nALLERGY: none\r\nPMH: HTN\r\nPSH: none\r\nSH: officer manager at an insurance company, rarely drinks, denies tobacco or drug use, sexually active with husband\r\nFH: non contributory, does not remember when her mother got menopausal.

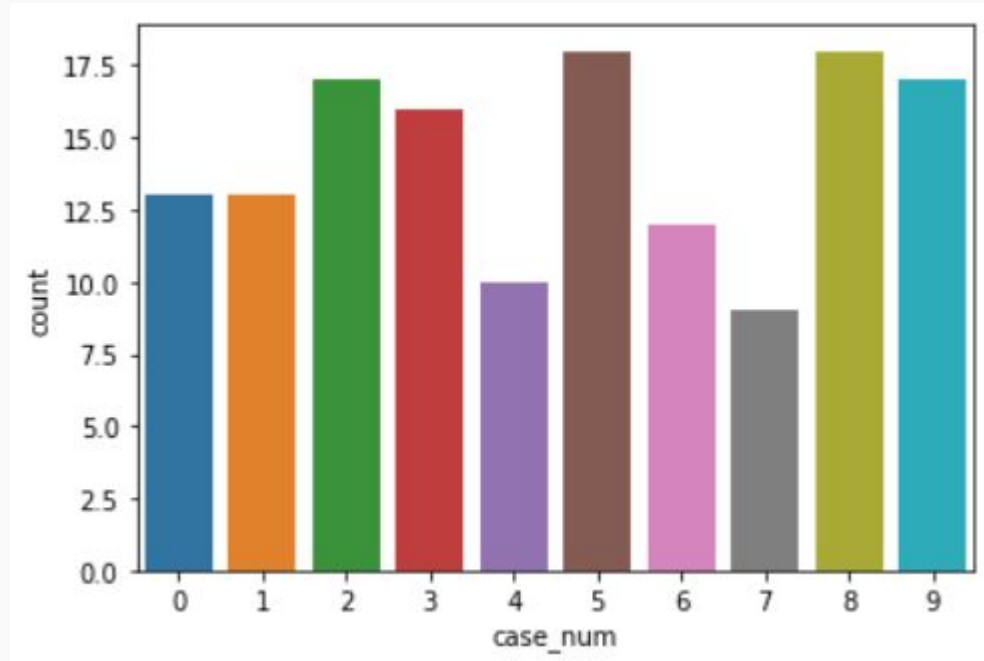
Présentation des données

features.csv : détails des caractéristiques de chaque cas clinique (143 lignes)

feature_num	case_num	feature_text
69	500	5 Onset-5-years-ago
70	501	5 Female
71	502	5 No-caffeine-use
72	503	5 Associated-SOB-OR-Associated-shortness-of-breath
73	504	5 Episodes-of-heart-racing
74	505	5 Recent-visit-to-emergency-department-with-negative-workup
75	506	5 No-chest-pain
76	507	5 No-illicit-drug-use
77	508	5 Associated-nausea
78	509	5 Increased-frequency-recently
79	510	5 Associated-feeling-of-impending-doom
80	511	5 Episodes-last-15-to-30-minutes
81	512	5 Associated-throat-tightness
82	513	5 Feels-hot-OR-Feels-clammy
83	514	5 Episode-of-hand-numbness-OR-Episode-of-finger-numbness
84	515	5 Fatigue-OR-Difficulty-concentrating
85	516	5 Increased-stress
86	517	5 26-year

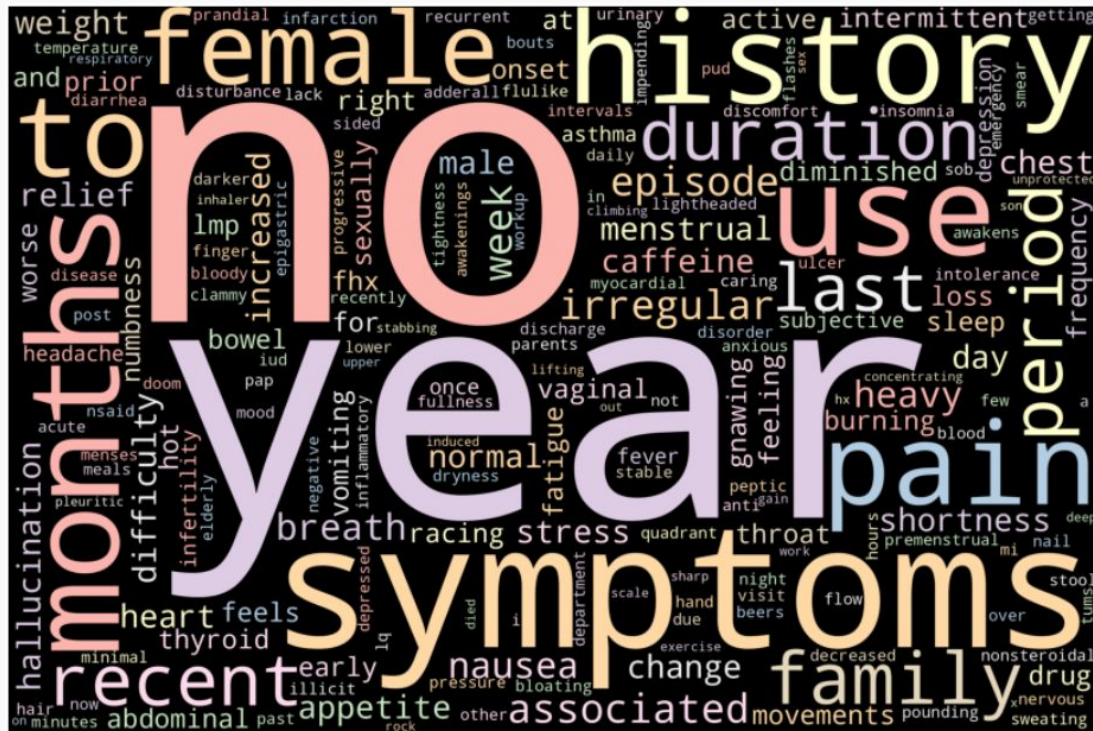
Présentation des données

features.csv : Nombre de caractéristiques par cas clinique



Présentation des données

Wordcloud des features



Présentation des données

train.csv : 14300 annotations des notes de patient avec 100 notes de patient par cas clinique.

	id	case_num	pn_num	feature_num	annotation	location
0	00016_000	0	16	0	['dad with recent heart attcak']	['696 724']
1	00016_001	0	16	1	['mom with "thyroid disease']	['668 693']
2	00016_002	0	16	2	['chest pressure']	['203 217']
3	00016_003	0	16	3	['intermittent episodes', 'episode']	['70 91', '176 183']
4	00016_004	0	16	4	['felt as if he were going to pass out']	['222 258']

On réalise un left join entre le train et les dataframes patient_notes et features. On passe également tout le texte en minuscule et on enlève les tirets.

Pre-processing

```
from transformers import AutoTokenizer
model_checkpoint = "bert-base-uncased"
tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)

feature = "no hair changes; no nail changes; no temperature intolerance"
patient_note = ("nothing makes his palpitations worse or better when they"
                " occur. ros is negative except for light headedness and sob."
                " he has not had any cold/heat intolerance, diarrhea, changes"
                " in voiding habits, or weight loss. worried about being able"
                " to play basketball.")

out = tokenizer(feature, patient_note, truncation='only_second',
               max_length=416, padding='max_length',
               return_offsets_mapping=True)

out.keys() → dict_keys(['input_ids', 'token_type_ids', 'attention_mask', 'offset_mapping'])
```

Pre-processing

out.offset_mapping	tokens	labels
(0,0)	[CLS]	
(0,2)	no	
(3,7)	hair	
(8,15)	changes	
...	...	
(37, 48)	temperat ure	
(49,53)	into	
(53,56)	##ler	
(56,60)	##ance	
(0,0)	[SEP]	

out.offset_mapping	tokens	labels
(0,7)	nothing	
(8,13)	makes	
	
(117,119)	he	
(120,123)	has	
(124,127)	not	
(128,131)	had	
(132,135)	any	
(136,140)	cold	
(140,141)	/	

out.offset_mapping	tokens	labels
(141,145)	heat	
(146,150)	into	
(150,153)	##ler	
(153,157)	##ance	
(157,158)	,	
(159,162)	diarrhea	
	...	
(245,255)	basketball	
(255,256)	.	
(0,0)	[SEP]	
(0,0)	[PAD]	

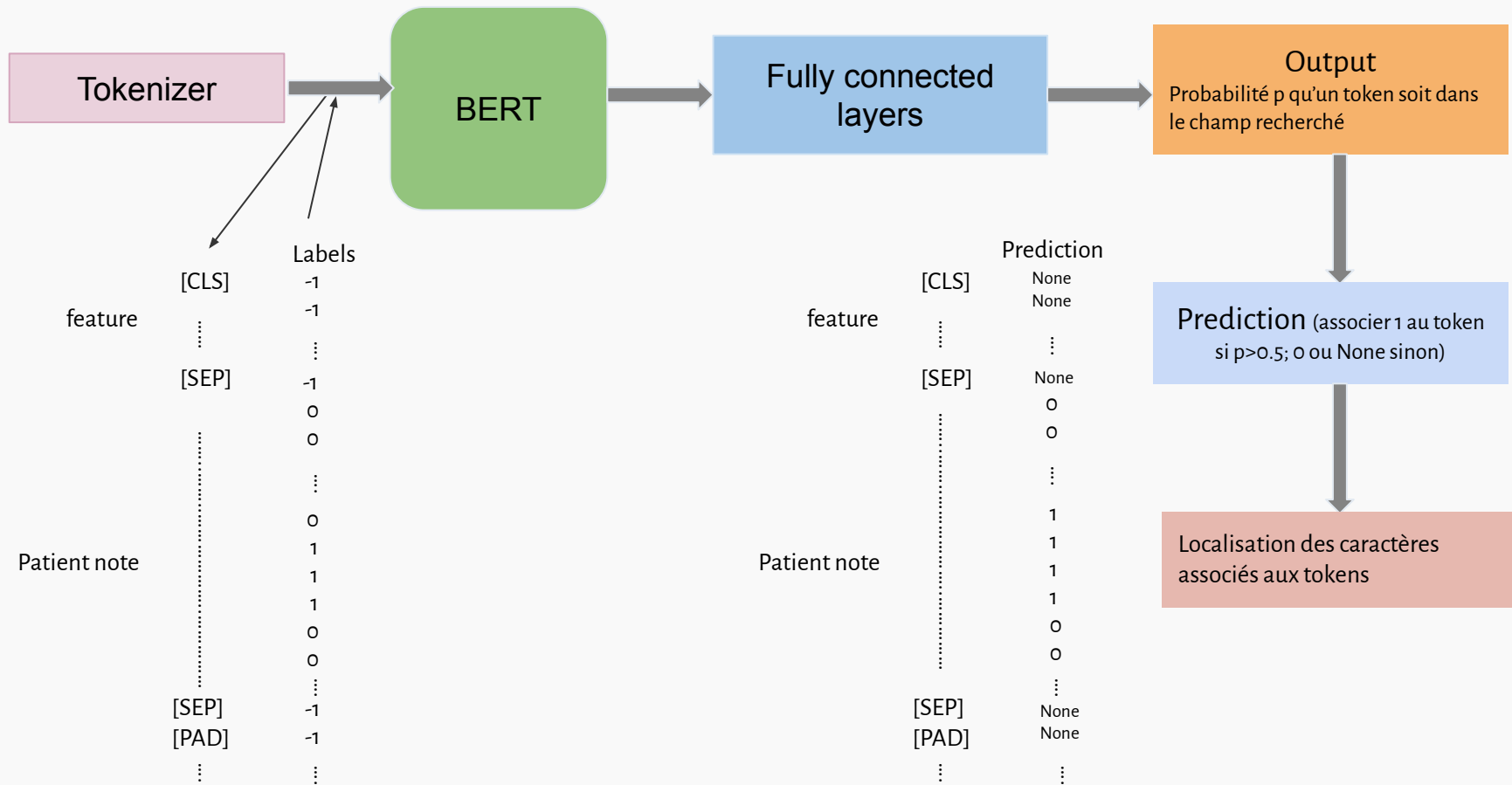
Pre-processing

out.offset_mapping	tokens	labels
(0,0)	[CLS]	-1
(0,2)	no	-1
(3,7)	hair	-1
(8,15)	changes	-1
...	...	-1
(37, 48)	temperature	-1
(49,53)	into	-1
(53,56)	##ler	-1
(56,60)	##ance	-1
(0,0)	[SEP]	-1

out.offset_mapping	tokens	labels
(0,7)	nothing	0
(8,13)	makes	0
	0
(117,119)	he	1
(120,123)	has	1
(124,127)	not	1
(128,131)	had	1
(132,135)	any	1
(136,140)	cold	1
(140,141)	/	1

out.offset_mapping	tokens	labels
(141,145)	heat	1
(146,150)	into	1
(150,153)	##ler	1
(153,157)	##ance	1
(157,158)	,	0
(159,162)	diarrhea	0
	...	0
(245,255)	basketball	0
(255,256)	.	0
(0,0)	[SEP]	-1
(0,0)	[PAD]	-1
(0,0)	[PAD]	-1

Modèle



hyperparamètres

Optimizer : AdamW

$lr = 1e-5$

Batch size = 8

Dropout = 0,2

Criterion : BCEWithLogitsLoss

Performances sur le jeu de test

	nb epochs	temps	f1 score
Baseline model (BERT base uncased + 3FC) pas de fonction d'activation sur les deux premières = oubli?)	6	64 min	0.795
BERT base uncased + 2 FC	6	63 min	0.803
BioBert + 2FC	6	63 min	0.800
PubmedBert + 2FC	6	62 min 30s	0.819

Versionning et choix du modèle

Afin de réaliser différents tests sur les modèles et leurs hyperparamètres nous avons créé différentes branches sur github. Nous avons réalisé un merge pour garder le meilleur modèle: le dernier entraîné avec PubmedBERT.

lien vers le repository : <https://github.com/CelineMendola/kaggle-competition>