PARTICIPEZ À UNE COMPÉTITION KAGGLE !

OPENCLASSROOMS - INGÉNIEUR MACHINE LEARNING

THIBAUD GROSJEAN - MARS 2022

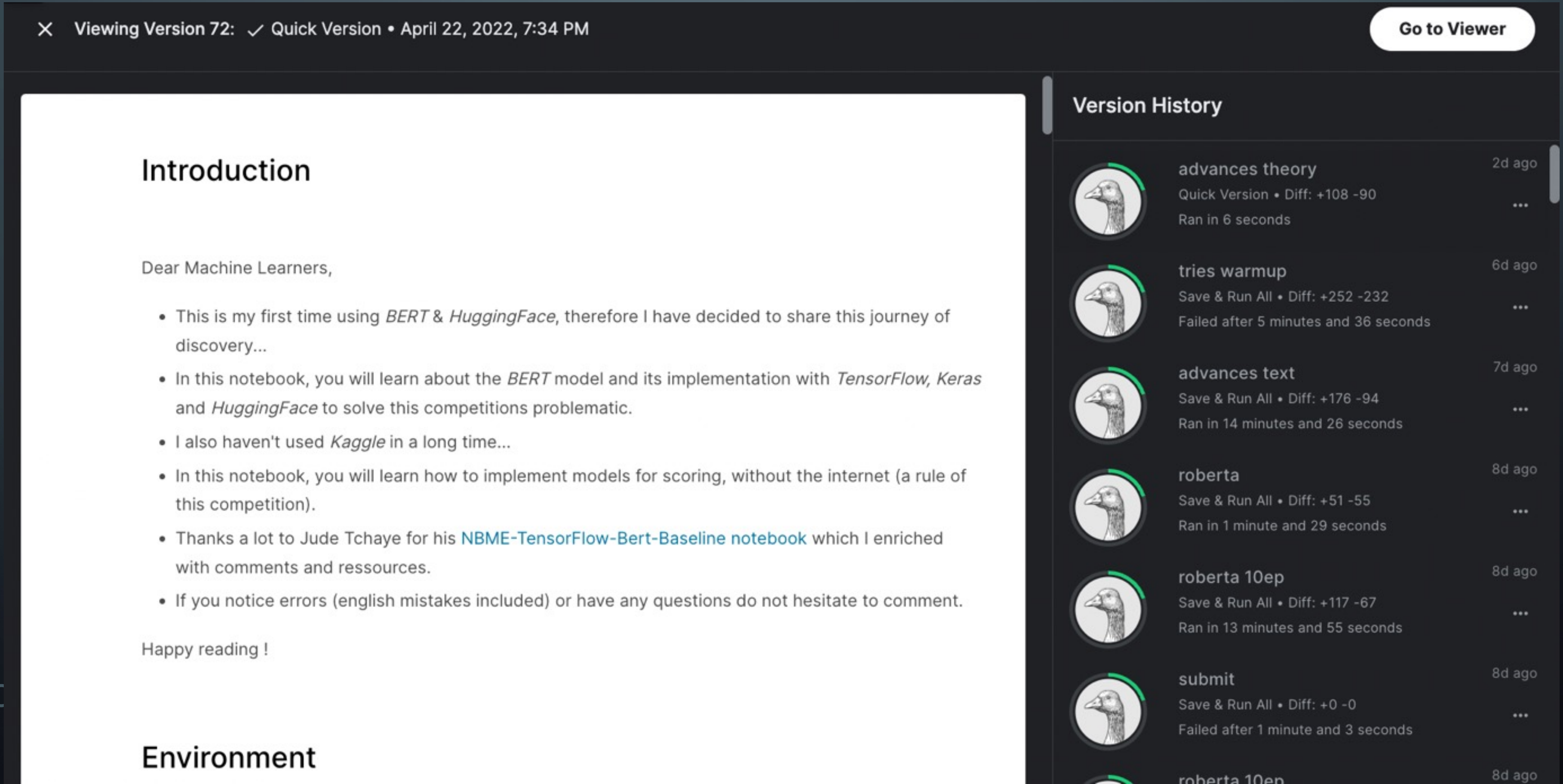
OBJECTIFS DU PROJET

- Participer à une compétition *Kaggle* (“*The House of Data Science*”)
- Utiliser les ressources partagées par la communauté
- Partager un élément intéressant avec la communauté
- Expliciter le modèle selectionné

COMPÉTITION

- **NBME - Score Clinical Patient Notes**
  - *United States Medical Licensing Examination® (USMLE®)*
  - Examen de compétences cliniques
  - Simulation de consultation
  - Historique du patient

VERSIONNAGE



✕ Viewing Version 72: ✓ Quick Version • April 22, 2022, 7:34 PM

Introduction

Dear Machine Learners,

- This is my first time using *BERT* & *HuggingFace*, therefore I have decided to share this journey of discovery...
- In this notebook, you will learn about the *BERT* model and its implementation with *TensorFlow, Keras* and *HuggingFace* to solve this competitions problematic.
- I also haven't used *Kaggle* in a long time...
- In this notebook, you will learn how to implement models for scoring, without the internet (a rule of this competition).
- Thanks a lot to Jude Tchaye for his NBME-TensorFlow-Bert-Baseline notebook which I enriched with comments and ressources.
- If you notice errors (english mistakes included) or have any questions do not hesitate to comment.

Happy reading !

Environment

Version History

advances theory 2d ago
Quick Version • Diff: +108 -90
Ran in 6 seconds

tries warmup 6d ago
Save & Run All • Diff: +252 -232
Failed after 5 minutes and 36 seconds

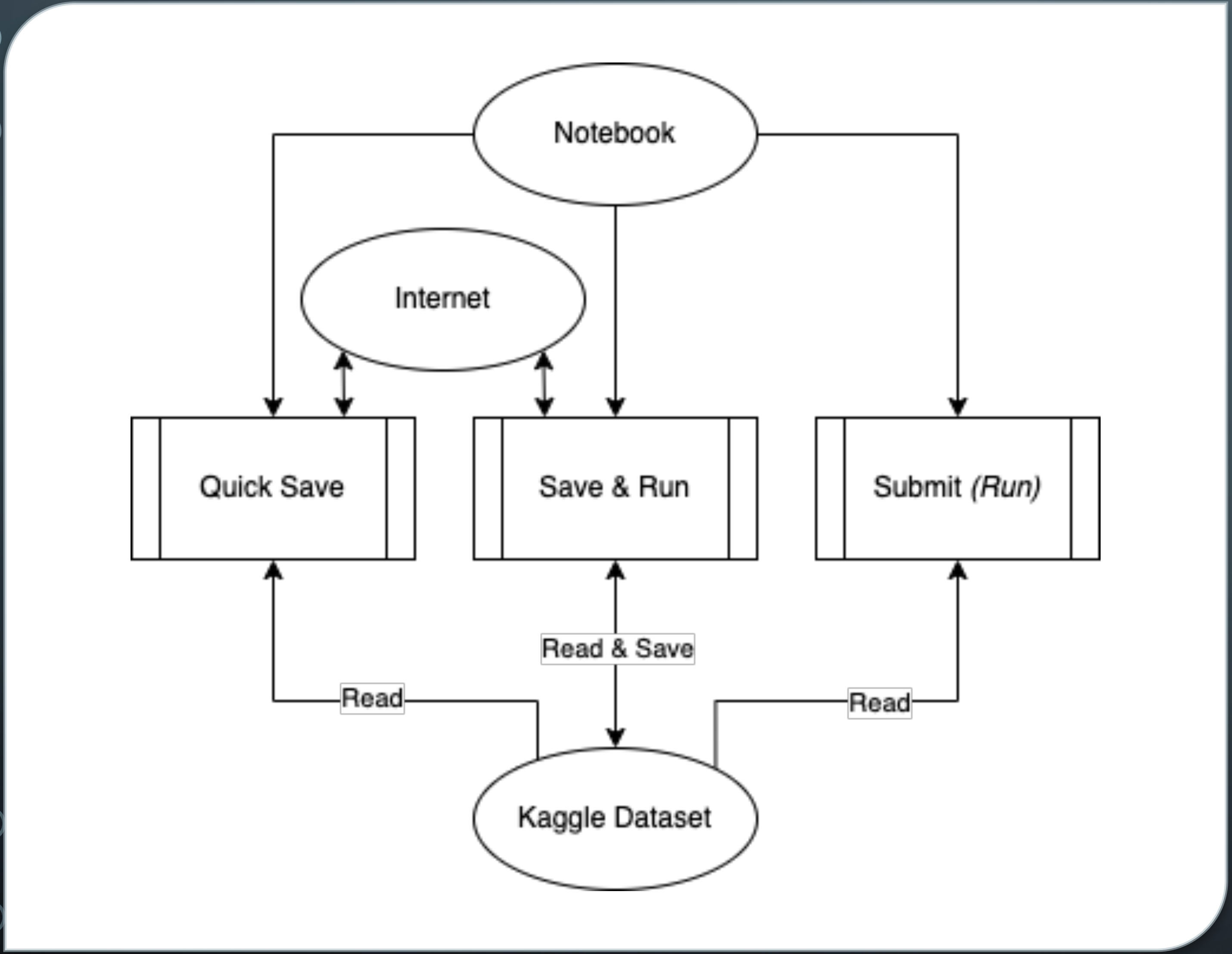
advances text 7d ago
Save & Run All • Diff: +176 -94
Ran in 14 minutes and 26 seconds

roberta 8d ago
Save & Run All • Diff: +51 -55
Ran in 1 minute and 29 seconds

roberta 10ep 8d ago
Save & Run All • Diff: +117 -67
Ran in 13 minutes and 55 seconds

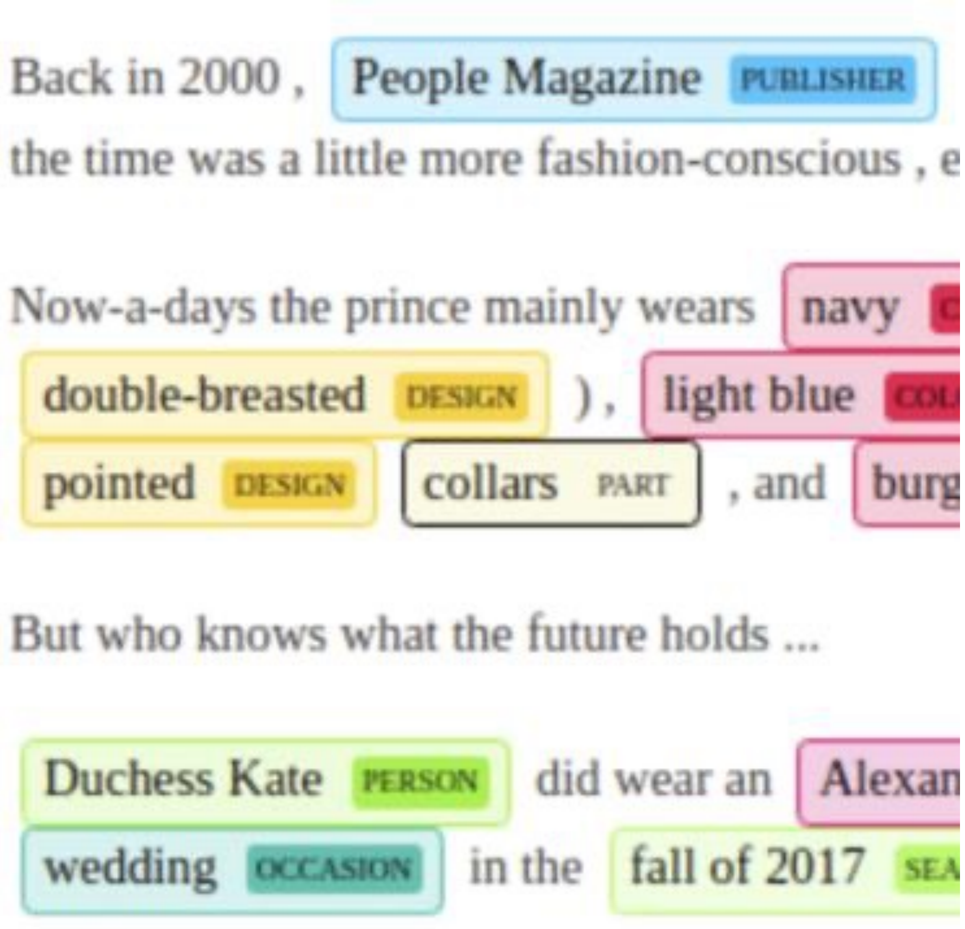
submit 8d ago
Save & Run All • Diff: +0 -0
Failed after 1 minute and 3 seconds

roberta 10ep 8d ago



ENVIRONNEMENT

- *Kaggle Notebooks*
- *Internet*
- *Parallelisation*



Back in 2000 , People Magazine PUBLISHER
the time was a little more fashion-conscious , e

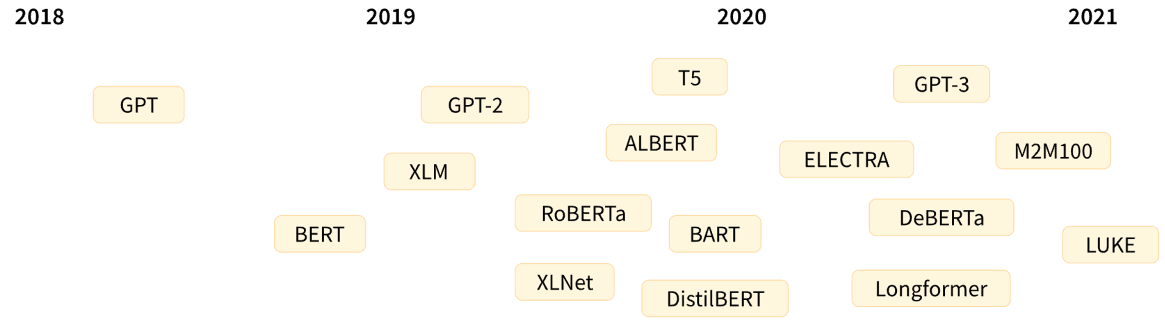
Now-a-days the prince mainly wears navy C
double-breasted DESIGN) , light blue COL
pointed DESIGN collars PART , and burg

But who knows what the future holds ...

Duchess Kate PERSON did wear an Alexan
wedding OCCASION in the fall of 2017 SEA

OBJECTIF

- *Natural Langage Processing (NLP)*
- *Named-entity Recognition (NER)*
- Evaluation *(score F1 micro-moyenné)*

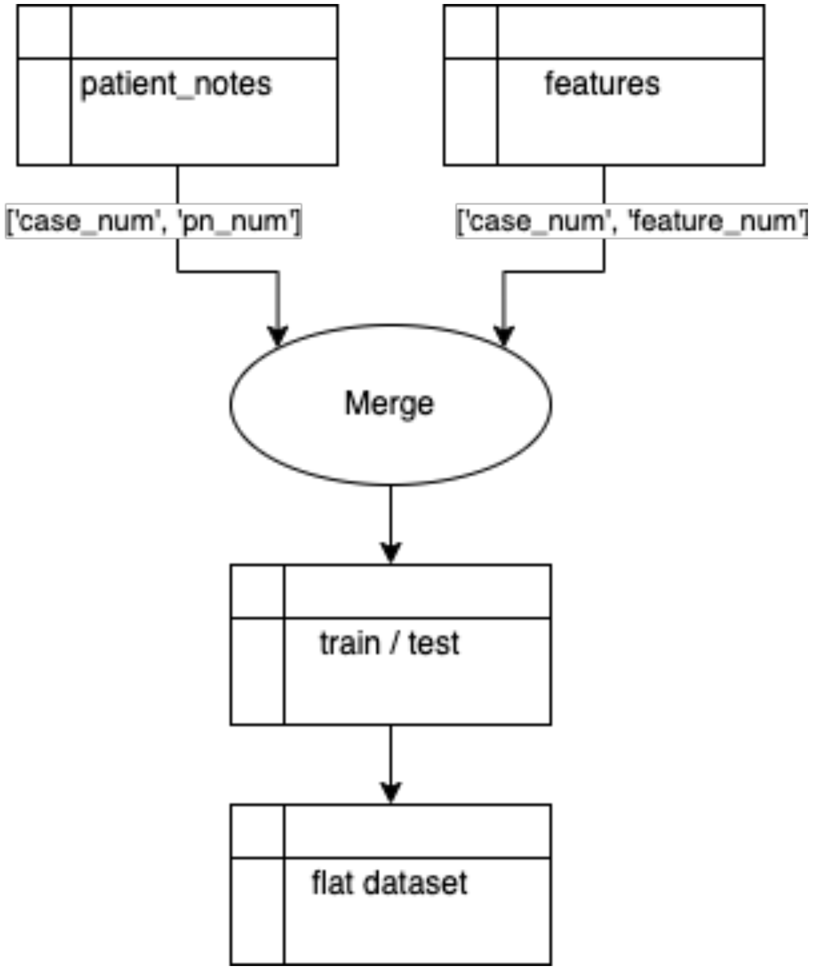


2018
2019
2020
2021

GPT
GPT-2
XLM
BERT
T5
ALBERT
RoBERTa
XLNet
BART
DistilBERT
GPT-3
ELECTRA
DeBERTa
Longformer
M2M100
LUKE

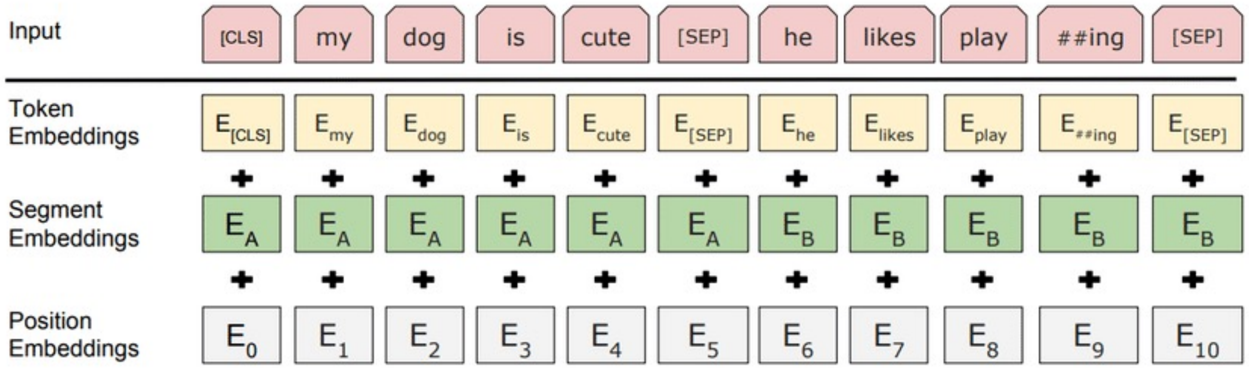
TRANFORMEURS

- *BERT*
- *RoBERTa*
- Modèles *auto-supervisés*



DONNÉES

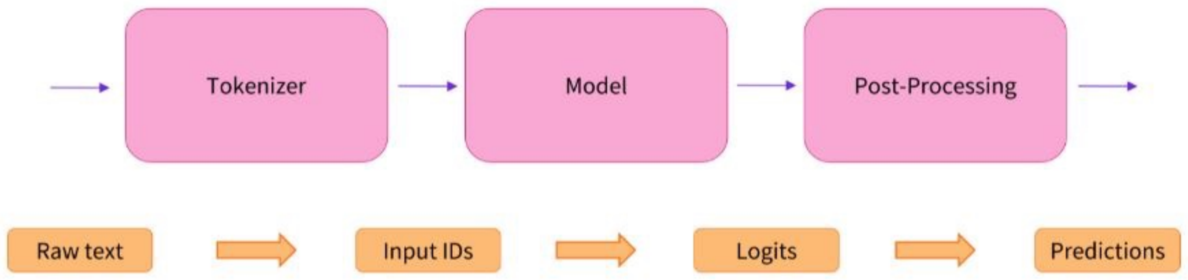
- *Patient Notes* (42146)
- *Features* (143)
- Création d'un fichier plat (14300 samples)


Input
[CLS] my dog is cute [SEP] he likes play ##ing [SEP]
Token Embeddings
$E_{[CLS]}$ E_{my} E_{dog} E_{is} E_{cute} $E_{[SEP]}$ E_{he} E_{likes} E_{play} $E_{\#\#ing}$ $E_{[SEP]}$
Segment Embeddings
E_A E_A E_A E_A E_A E_A E_B E_B E_B E_B E_B
Position Embeddings
E_0 E_1 E_2 E_3 E_4 E_5 E_6 E_7 E_8 E_9 E_{10}

EMBEDDINGS

- Vecteurs representant numériquement les caractéristiques du *document*
- *Input ids* (tokens)
- *Attention masks*
- *Labels* (cibles)

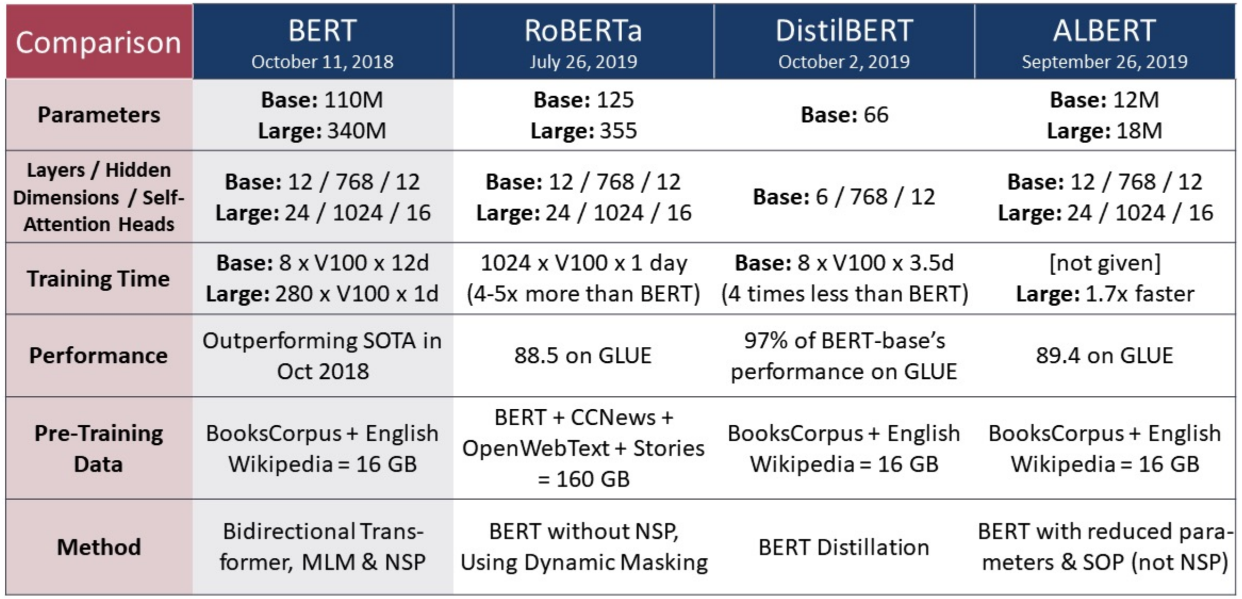
PIPELINE



Tokenizer → Model → Post-Processing

Raw text → Input IDs → Logits → Predictions

- Simple en apparence…
- Mais complexe
- Le *Tokenizer* permet d’inverser les *logits* du modèle



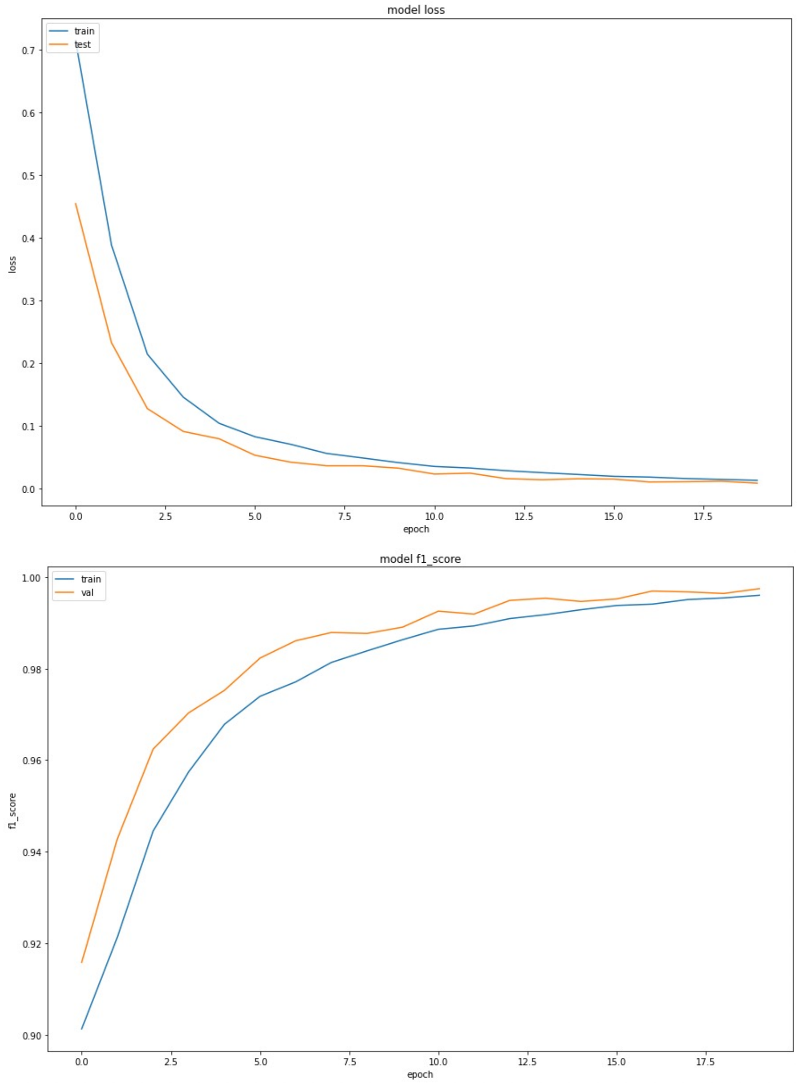
Comparison	BERT October 11, 2018	RoBERTa July 26, 2019	DistilBERT October 2, 2019	ALBERT September 26, 2019
Parameters	**Base:** 110M **Large:** 340M	**Base:** 125 **Large:** 355	**Base:** 66	**Base:** 12M **Large:** 18M
Layers / Hidden Dimensions / Self-Attention Heads	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16	**Base:** 6 / 768 / 12	**Base:** 12 / 768 / 12 **Large:** 24 / 1024 / 16
Training Time	**Base:** 8 x V100 x 12d **Large:** 280 x V100 x 1d	1024 x V100 x 1 day (4-5x more than BERT)	**Base:** 8 x V100 x 3.5d (4 times less than BERT)	[not given] **Large:** 1.7x faster
Performance	Outperforming SOTA in Oct 2018	88.5 on GLUE	97% of BERT-base's performance on GLUE	89.4 on GLUE
Pre-Training Data	BooksCorpus + English Wikipedia = 16 GB	BERT + CCNews + OpenWebText + Stories = 160 GB	BooksCorpus + English Wikipedia = 16 GB	BooksCorpus + English Wikipedia = 16 GB
Method	Bidirectional Transformer, MLM & NSP	BERT without NSP, Using Dynamic Masking	BERT Distillation	BERT with reduced parameters & SOP (not NSP)

AMÉLIORATION DU SCORE

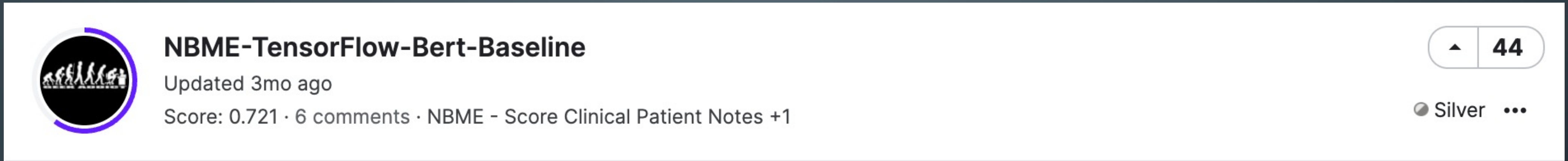
- Parallélisation
- Split des données
- Visualisation
- Implémentation d'un modèle plus performant *(RoBERTa)*
- Optimisation du nombre d'*epochs*
- *Model Checkpoint*

ENTRAINEMENT & OPTIMISATION

- Le modèle *merge* correctement



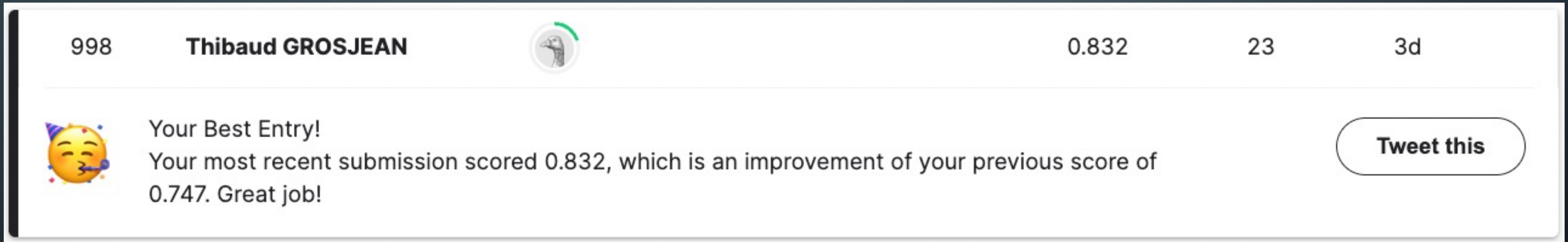
AMÉLIORATION DU SCORE



NBME-TensorFlow-Bert-Baseline

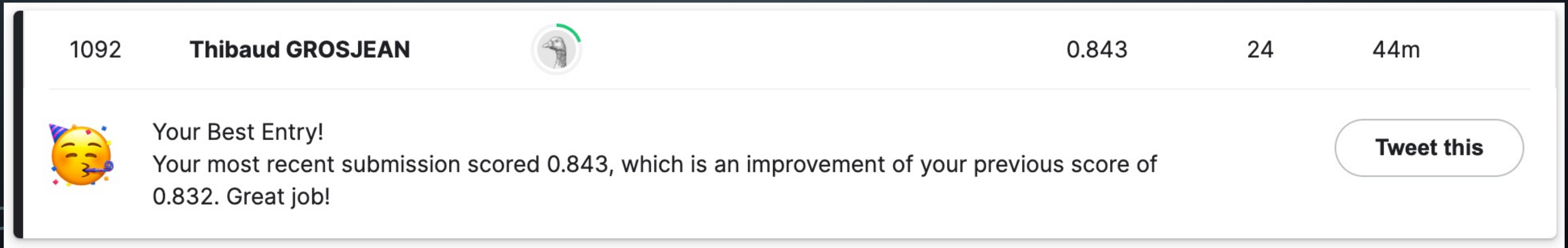
Updated 3mo ago

Score: 0.721 · 6 comments · NBME - Score Clinical Patient Notes +1



998	**Thibaud GROSJEAN**	0.832	23	3d

Your Best Entry!
Your most recent submission scored 0.832, which is an improvement of your previous score of 0.747. Great job!



1092	**Thibaud GROSJEAN**	0.843	24	44m

Your Best Entry!
Your most recent submission scored 0.843, which is an improvement of your previous score of 0.832. Great job!

AMÉLIORATION DE LA RESSOURCE EXISTANTE

- Création d’une ressource dictatisée
- Ajout de commentaires & de ressources
- Simplification du process *Kaggle notebooks*
- Amélioration du style
- Amélioration de la méthodologie en *Data Science*

CONCLUSION & PISTES D’AMÉLIORATION

- Prise de main de *Kaggle Notebooks* & de la *Parallelisation*
- Découverte des *Transformers* & de *Hugging Face*

- Utilisation des ressources communautaires
- Amélioration du score
- Partage de la ressource créée

ÉCHANGE & QUESTIONS

Merci de votre attention !