DÉVELOPPEZ UNE PREUVE DE CONCEPT

OPENCLASSROOMS - INGÉNIEUR MACHINE LEARNING

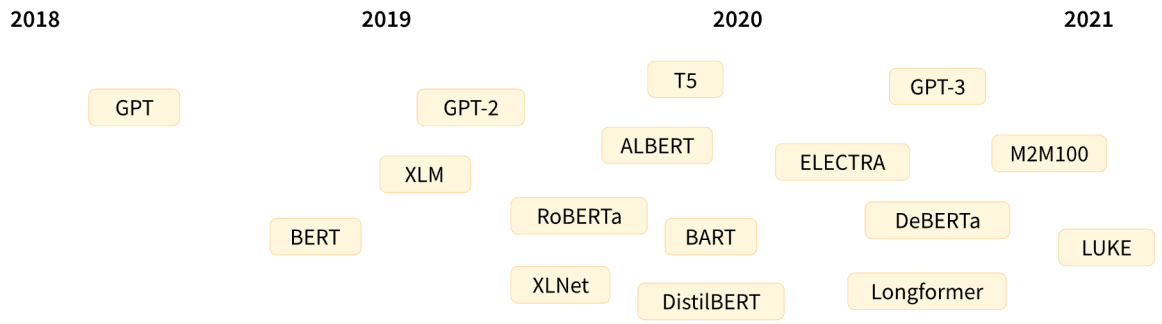
THIBAUD GROSJEAN - MAI 2022

OBJECTIFS

- Implémenter une technique de pointe
- Définir le plan de travail prévisionnel
- Mener le travail de recherche
- Réaliser un état de l'Art
- Comparer les performances à un modèle baseline

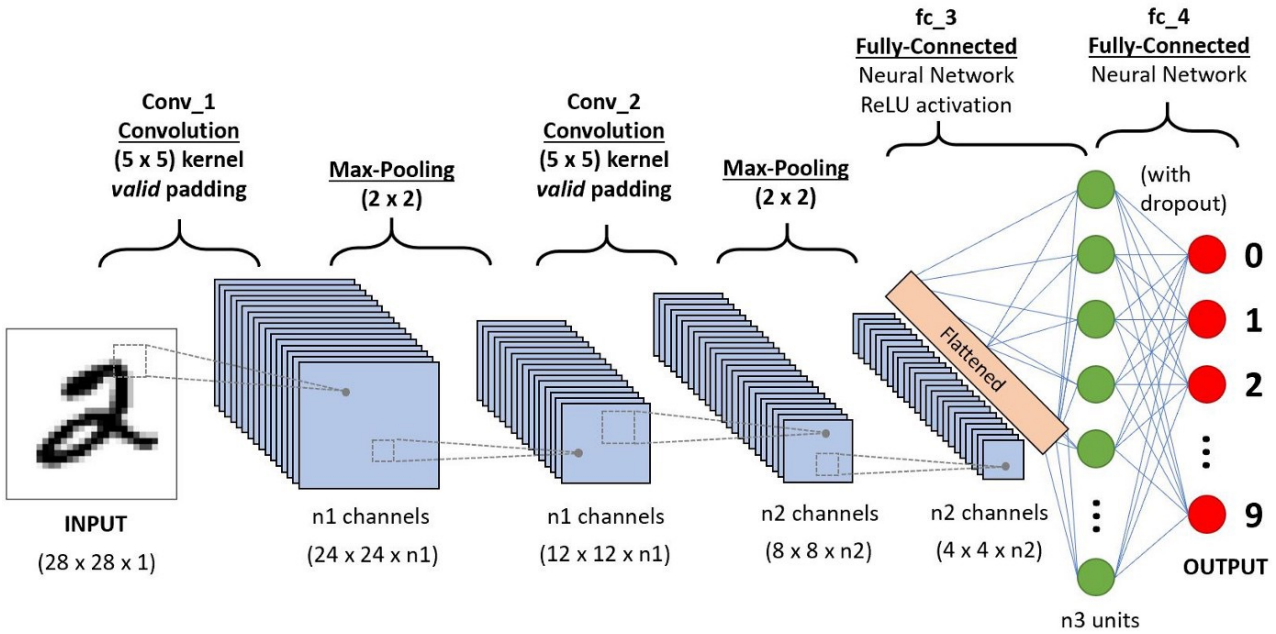
BIBLIOGRAPHIE

- Ashish Vaswani et al, Attention Is All You Need, 2017
- Kevin Clark et al., ELECTRA: Pre-Training Text Encoders Aa Discriminators Rather Than Generators, 2020
- Pengcheng He et al., DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, 2021
- Francesco Barbieri et al., TWEETEVAL: Unified Benchmark and Comparative Evaluation for Tweet Classification, 2020
- Daniel Loureiro et al., TimeLMs: Diachronic Language Models from Twitter, 2022



TRANFORMEURS

- Transductif
- Parallelisable
- Auto-Attention
- Auto-Supervisé
- Compréhension statistique

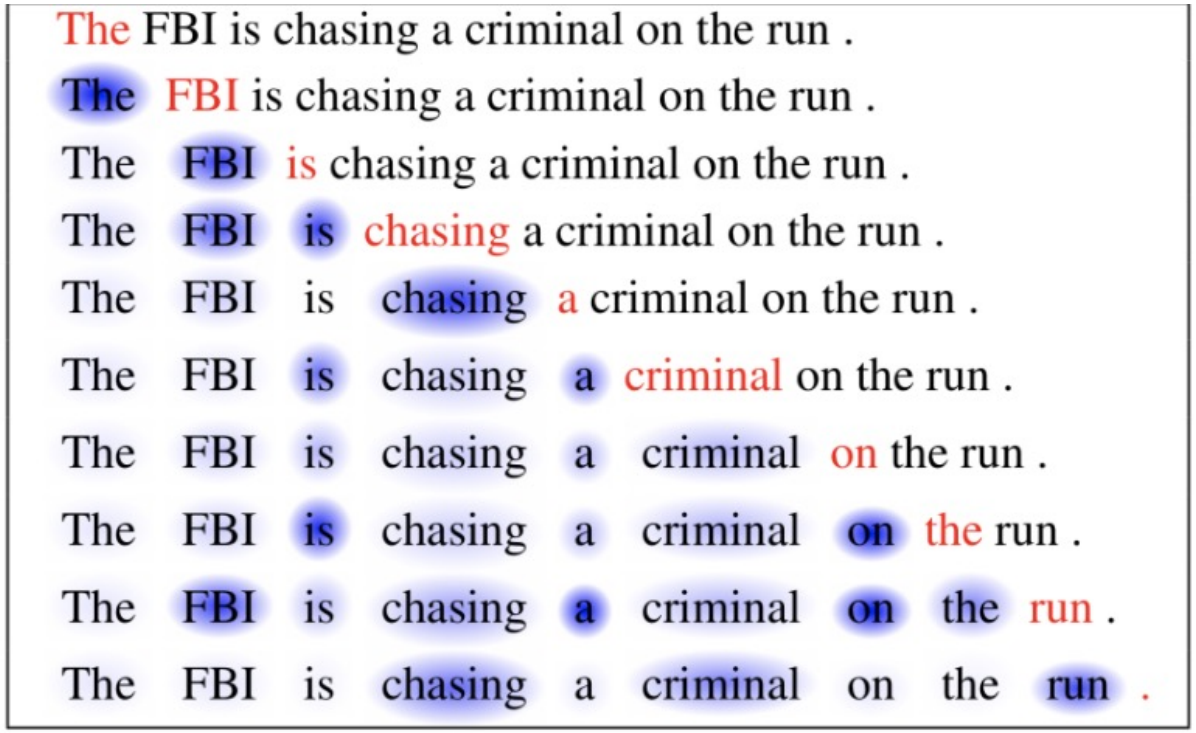


TRANFORMEURS

- CNN
- Séquentialité

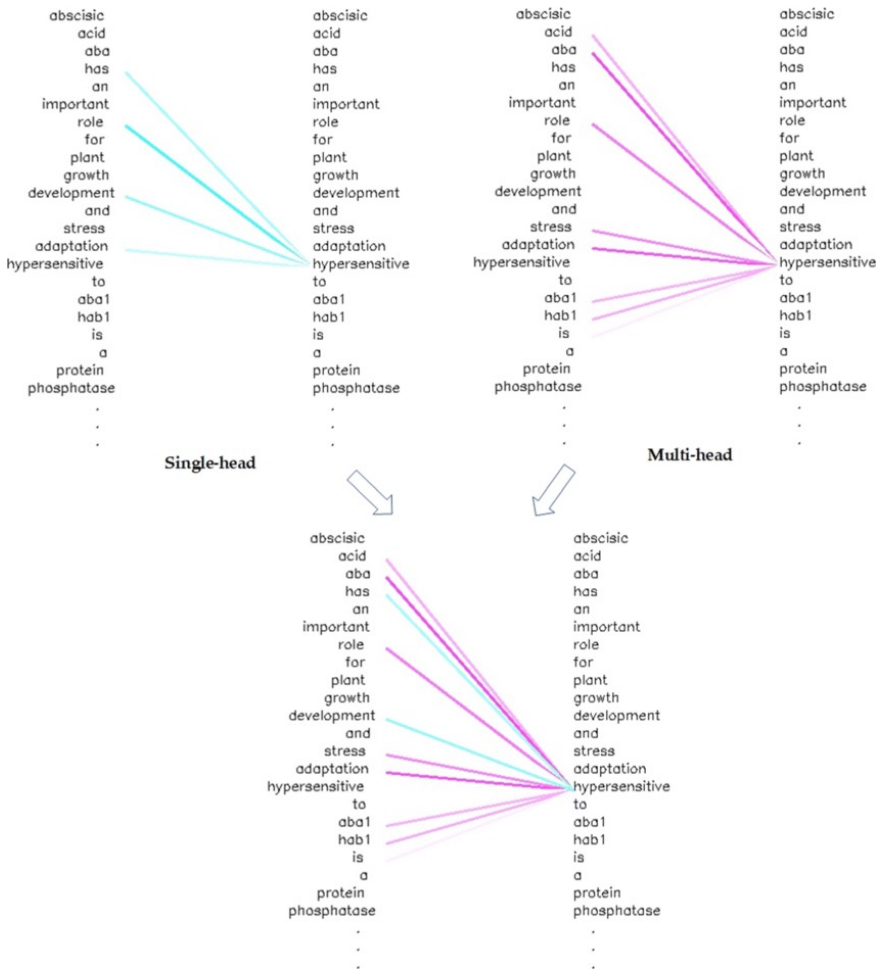
ATTENTION

- Qu'est-ce que l'attention ?
- Attention par produit scalaire
- Optimisation
- Encodeur
- Encodeur-Décodeur



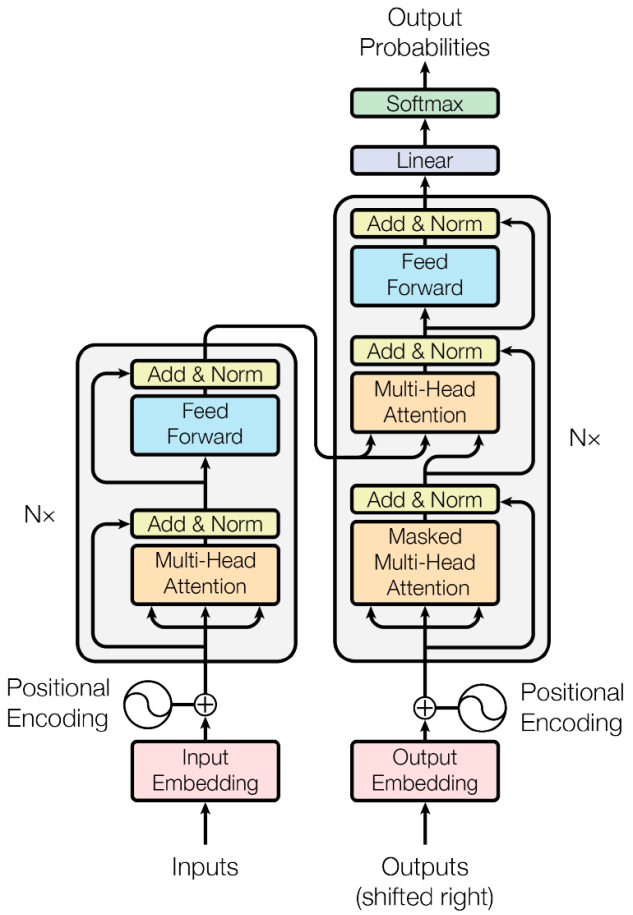
ATTENTION

- Décodeur



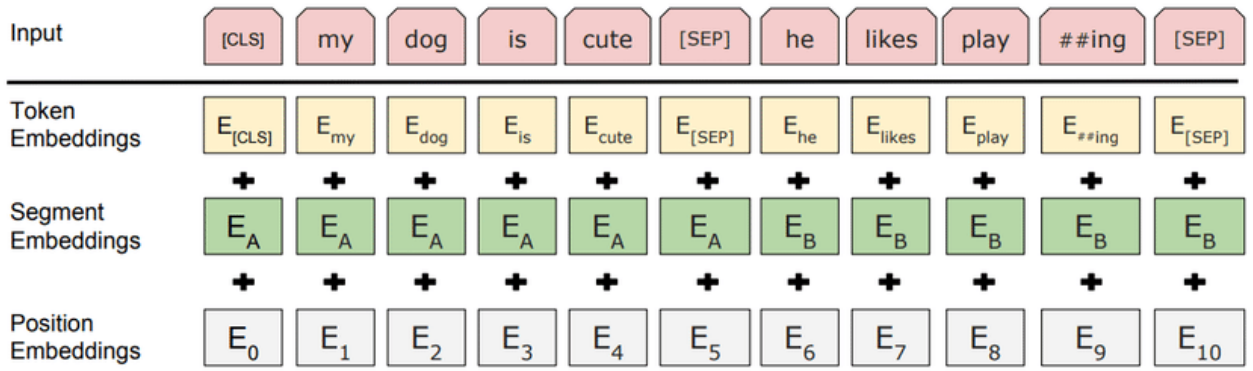
ATTENTION

- Attention à têtes multiples
- Parallélisation
- **Attention Démêlée**



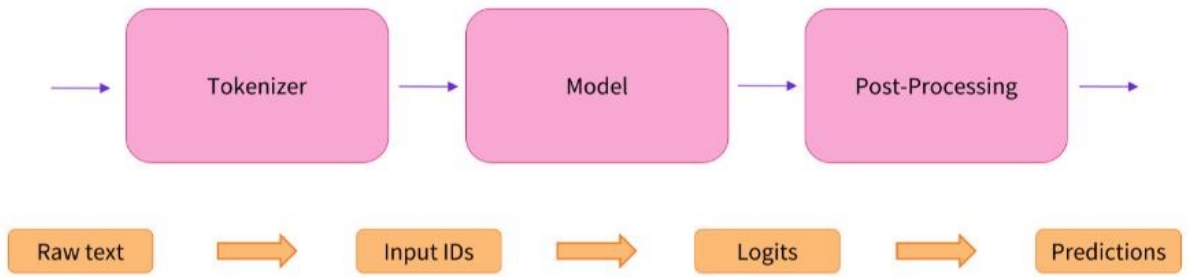
ARCHITECTURE

- 6+6 layers d'attention
- 768 neurones de préclassification
- Drop-out (0.3)
- Neurones de classification



SÉQUENCES

- Vecteurs
  - *Input ids* (tokens)
  - *Input types*
  - *Attention masks*
  - *Labels* (cibles)
- Longueur de séquence: 100



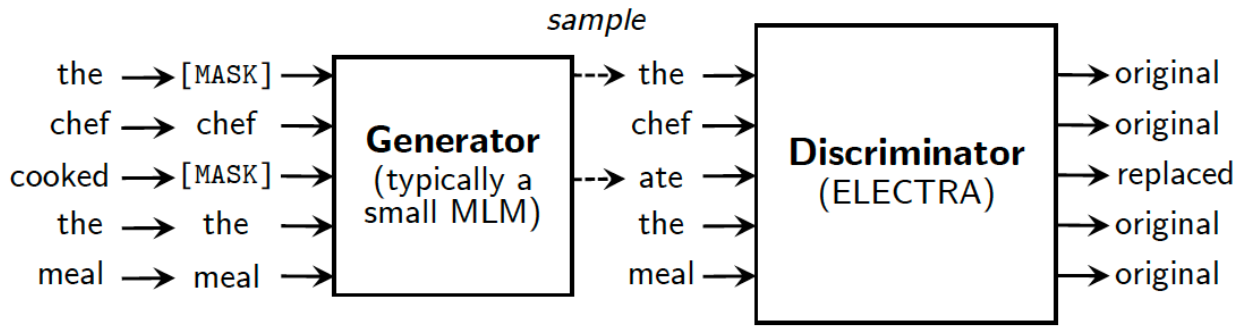
PIPELINE

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

Model	Wiki+Book 16GB	OpenWebText 38GB	Stories 31GB	CC-News 76GB	Giga5 16GB	ClueWeb 19GB	Common Crawl 110GB	CC100 2.5TB
BERT	✓							
XLNet	✓				✓	✓	✓	
RoBERTa	✓	✓	✓	✓				
DeBERTa	✓	✓	✓					
$\text{DeBERTa}_{1.5B}$	✓	✓	✓	✓				
DeBERTaV3	✓	✓	✓	✓				
mDeBERTa_{base}								✓

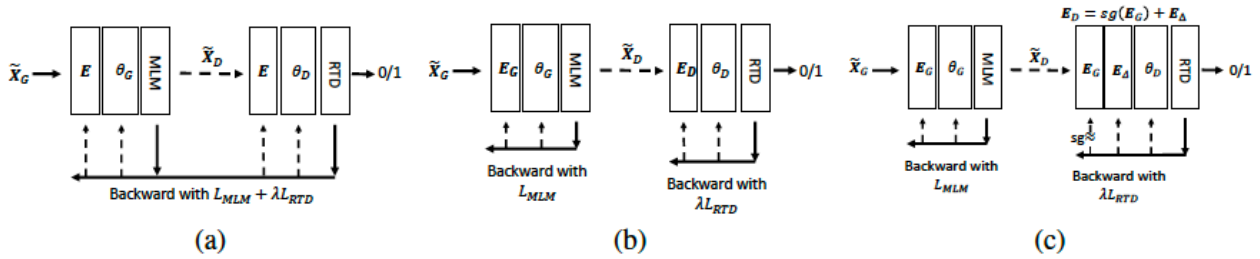
PRÉ-ENTRAINEMENT

- Réprésentations de langage
- **161GB**
- Puissance de calcul massive



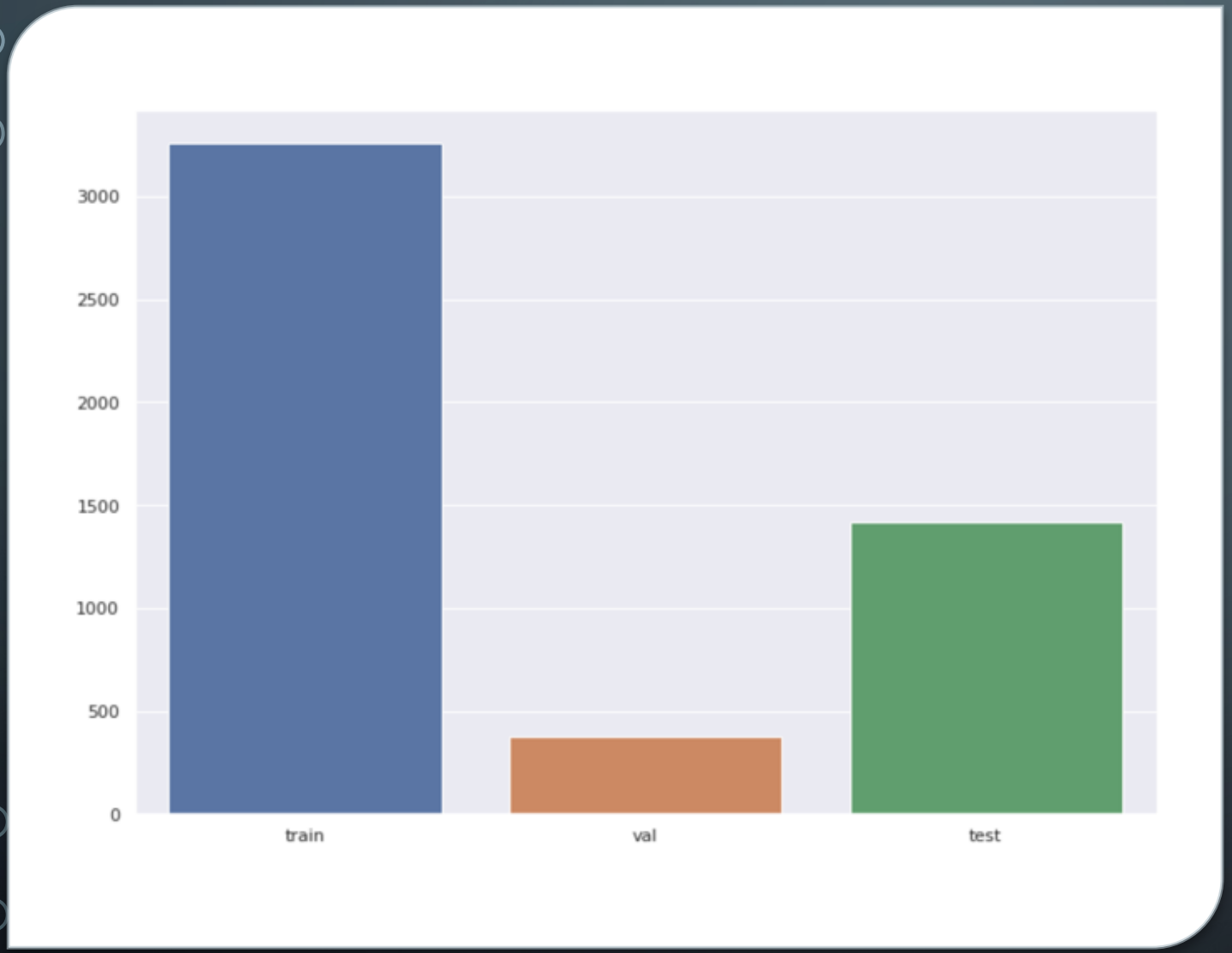
PRÉ-ENTRAINEMENT

- NSP
- MLM
- Tokens corrompus “[‘MASK’]”
- 15%
- **RTD** (ELECTRA)



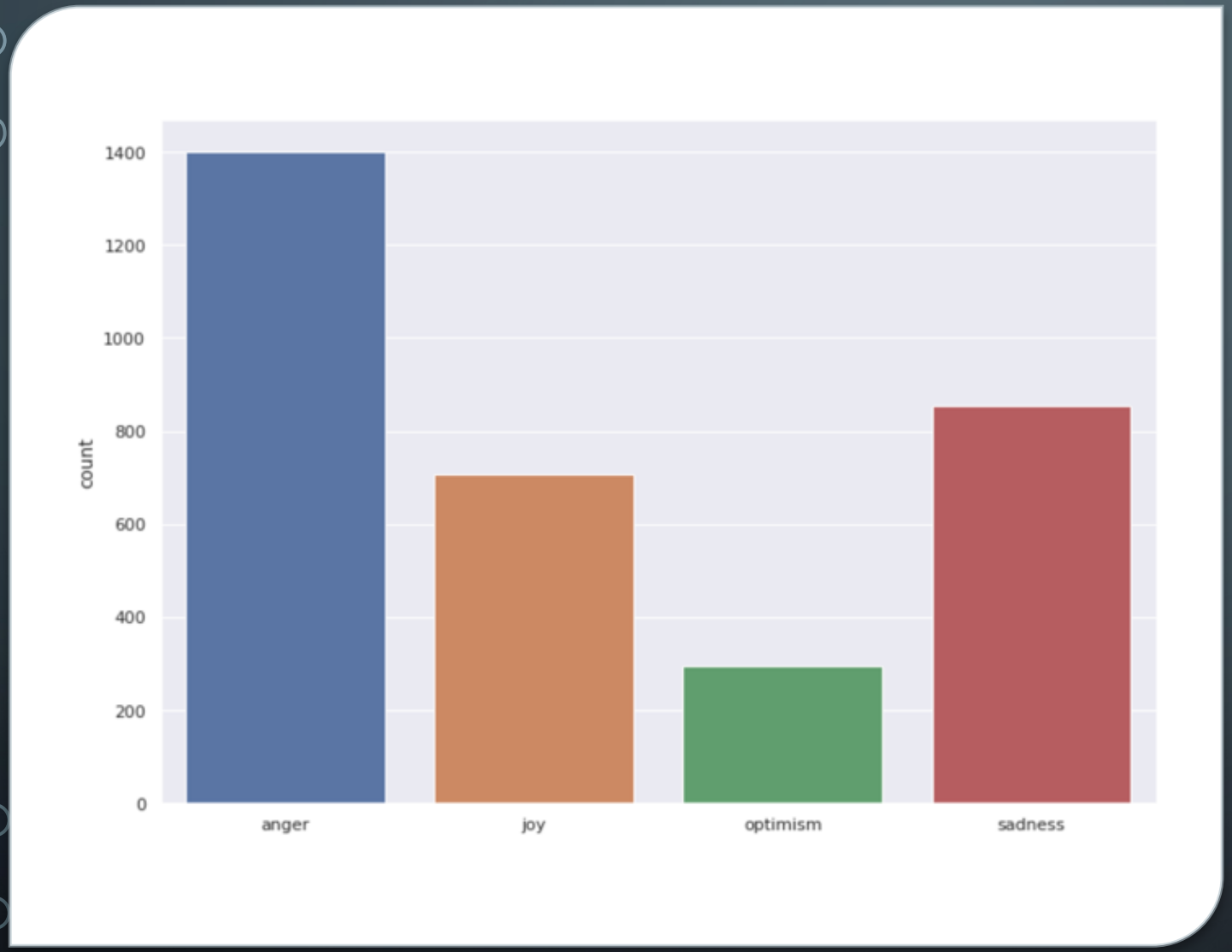
SÉQUENCES & GRADIENTS

- Tug of war
- Partage de séquences (ES)
- Non partage de séquences (NES)
- **GDES**



JEU DE DONNÉES

- TWEETEVAL Emotion
- 5052 tweets
  - Entrainement : 3257
  - Validation : 374
  - Test : 1421
- 4 classes
  - Anger
  - Joy
  - Optimism
  - Sadness



JEU DE DONNÉES

- Distributions déséquilibrée
- La classe modale “anger” représente 43% de jeu d’entrainement
- Score : F1 macro-moyenné

IMPLÉMENTATION

IMPLÉMENTATION

- Sickit Learn
- HuggingFace
- PyTorch
- Google Colab (TPU)
- Weights and Biases
- transformers_sentiment_imdb.ipynb

RÉSULTATS

- Jeu de validation

Validation	CM	LR	USE	BERT (base)	DeBERTaV3 (base)
F1 macro	14.9	30.5	45.1	71.4	**76.8**
Accuracy	42.7	43.8	60.6	77.5	**83.6**

RÉSULTATS

- Jeu de test

Test	CM	LR	USE	BERT (base)	TimeLMS	DeBERTaV3 (base)
F1 macro	14.0	28.2	54.4	77.2	80.2	**82.4**
Accuracy	39.2	41.8	65.1	80.5	-	**84.7**

CONCLUSION

- Exploration technique transversale
- Découverte de Pytorch et Weights & Biases
- Classification multiple déséquilibrée
- 4 algorithmes implémentés
- Score supérieur au benchmark (jeu de test)

ÉCHANGE & QUESTIONS

Merci de votre attention !