# BERT dans la classification multi-classes

Classification et extraction de données dans des documents administratifs

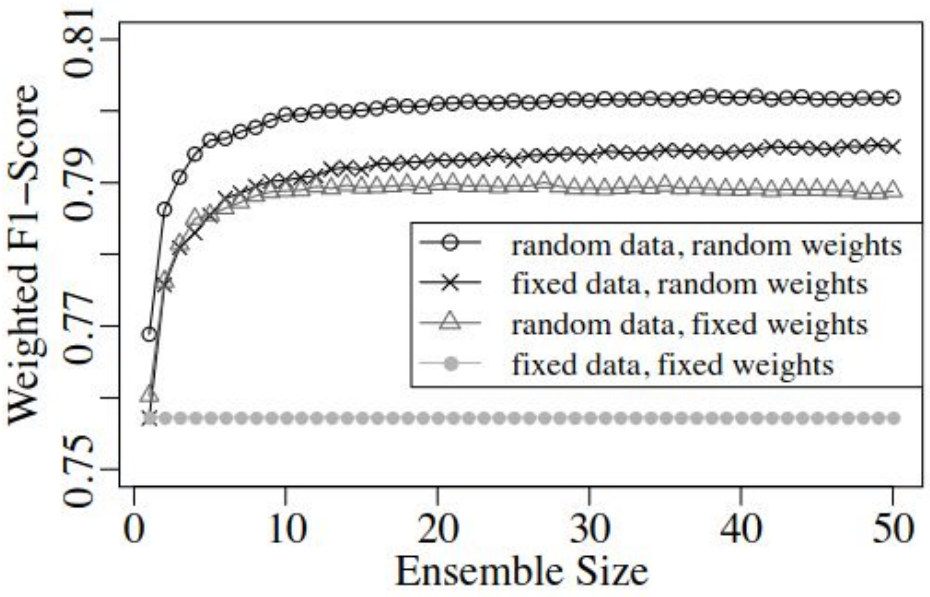
Gindro Quentin - Lagouge Mathis
ING3 IA - 2023/2024

Professeurs référents:
Jourdan Astrid - Le Nir Yannick - Loubiere Peio

# Objectifs :

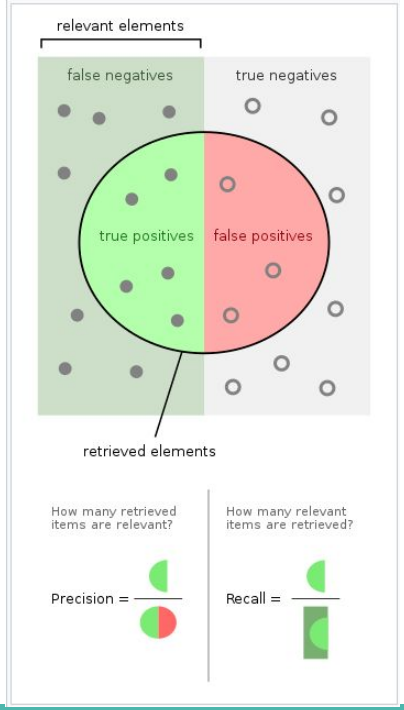
- **Reproduire un BERT base**
- **Reproduire un Bagging BERT**
- **Comparer les 2 modèles aux résultats de la thèse**

# Rappel : Boostrap aggregation (BaggingBERT)




**F1-score en fonction du nombre de modèle BERT**

$$F1 - score = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$



# Construction d'un BERT base

**Utilisation librairie FARM (basé sur les modèles BERT et la librairie Transformers de HuggingFaces)**

- **Tokenizer :**

**Utilise un BERT pré-entraîné adapté à la langue choisie**

**"bert-based-uncased" – anglais**

- **Processor : gère le dataset**

**train/test/dev Metrics : f1-score Prend le Tokenizer**

# Construction d’un BERT base

- **Model : Adaptive Model**

**Basé sur un BERT pré-entraîné sur l’anglais**

- **Optimizer : Ajuster les poids et les biais pendant l’entraînement**

**Prend le modèle learning_rate**

- **Trainer :**

**Regroupe Processor + Optimizer**

# Construction d’un BERT base

- **Paramètres utilisé :**

**batch_size : 32**

**max_sequence_lenght : 64**

**learning_rate : 2e-5**

**embedings_dropout_probability : 0.5**

# Construction d’un BaggingBERT

- Création d’une dizaine de BERT fine-tune pour les regrouper en 1 seul modèle.
- Boucle sur tous les éléments tokenizer du dataset[“test”]
- Comparaison entre la probabilité maximale déjà calculée et la probabilité maximale d’un modèle. Récupération de la classe associée.
- Calcul de l’accuracy, de la précision et du f1_score pour chaque classe.

# Construction d'un BaggingBERT

```
accuracy du modèle: 0.6486956521739130

Classe 0 :
precision : 0.4876325088339223
F1_score : 0.6456140350877193

Classe 1 :
precision : 0.9561551433389545
F1_score : 0.6517241379310345
```

# Comparaison des résultats

**MRPC dataset de GLUE**

**Dataset pour tester la classification de Texte**

- **BERT - base :**

**accuracy 0.82**

- **Bagging BERT :**

**accuracy 0.65**

| Model | BERTBase | BagBERT | B2BERT |
|---|---|---|---|
| MRPC | 85.29 | 86.52 | 87.87 |
| $MNLI_m$ | 84.39 | 84.61 | 85.16 |
| RTE | 67.15 | 71.12 | 72.92 |
| ECDT | 94.54 | 96.10 | 96.88 |
| ChnSent | 93.00 | 93.5 | 94.33 |
| XNLI | 77.51 | 77.63 | 79.08 |

# Conclusion

**Améliorer le modèle Bagging BERT**

**Utiliser le cloud pour booster les performances (Google Collab)**

**Prochaine étape : Construire un modèle Boosting BERT**

# Sources:

**Explication de fine-tuning de BERT:**

- https://huggingface.co/docs/transformers/main_classes/trainer

**Librairie FARM:**

- https://farm.deepset.ai/

# Sources:

**[1]** Performance en classification de données textuelles des passages aux urgences des modèles BERT pour le français, décembre 2021

**[2]** Attention Mechanism, Transformers BERT and GPT: Tutorial and Survey, 2020

**[3]** Bagging BERT Models for Robust Aggression Identification, mai 2020

**[4]** BoostingBERT Integration Multi-Class Boosting into BERT for NLP Tasks, septembre 2020