
Indexation disciplinaire automatisée de publications : de l'entraînement local d'un modèle de Machine Learning aux Modèles de langage pré-entraînés

Géraldine Geoffroy

Le contexte (2018-2019)

Un baromètre Open Access local développé avant l'ouverture du BSO national

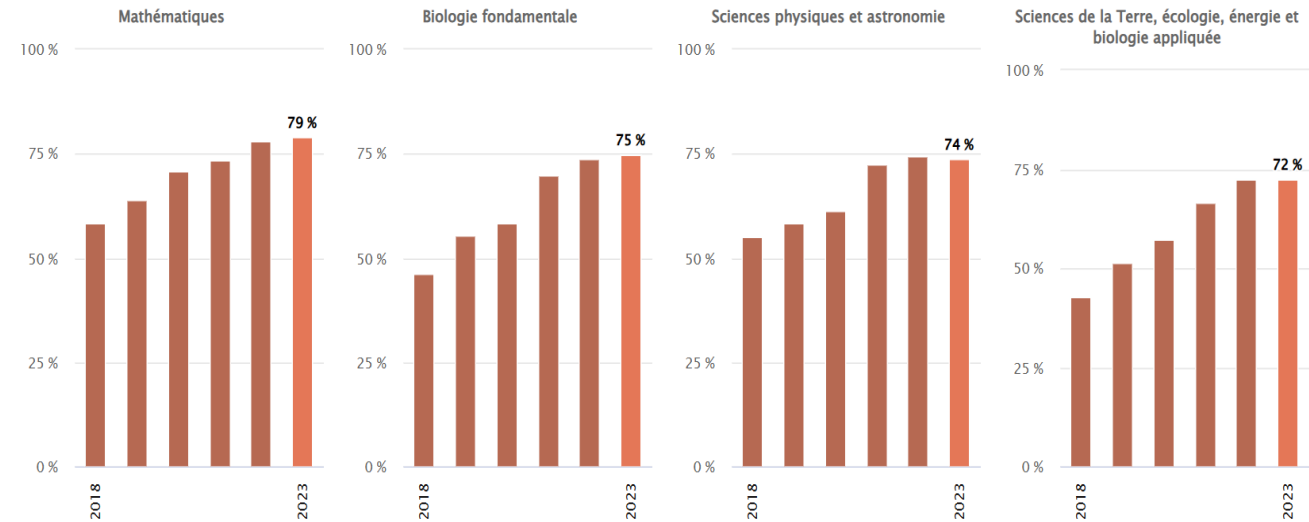
La problématique

Adapter notre baromètre local pour se conformer (et se comparer) aux indicateurs nationaux

L'enjeu

Produire des indicateurs d'ouverture des publications par discipline -> ajouter une indexation par discipline à notre corpus de publications

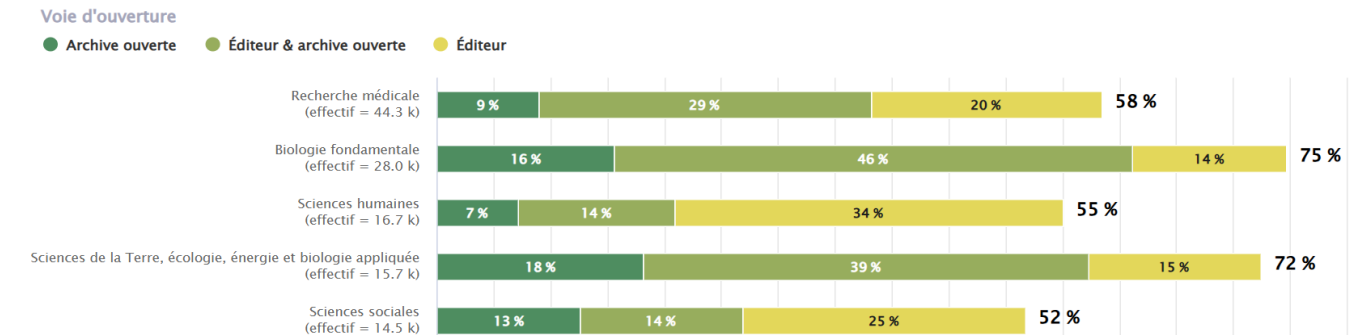
Taux d'accès ouvert par discipline et par année d'observation, pour les publications françaises, avec un DOI Crossref, parues durant l'année précédente (disciplines présentées dans l'ordre du taux d'accès décroissant)



Répartition des publications françaises, avec un DOI Crossref, par voie d'ouverture pour chaque discipline (publications de 2022)

Trier par :

Plus grand effectif Plus fort taux d'accès ouvert



Méthodologie : développement d'un modèle de Machine Learning de multi-classification pour automatiser la classification des publications non présentes dans le jeu de données du BSO

- datasets d'entraînement et de test
- Equilibrage du jeu de données d'entraînement
- Tâches NLP de nettoyage des données textuelles
 - Conversion en matrice pondérée en TF-IDF (approche Bags of words)

Baromètre de la science ouverte (général)

Ce jeu de données recense les données de publications sous-jacentes au baromètre français de la Science Ouverte.

[open science](#) [open access](#) [HAL](#) [Unpaywall](#) [archive ouverte](#) [archives ouvertes](#) [BSO](#) [French Open Science Monitor](#) [recherche](#) [science ouverte](#)
[publications scientifiques](#) [publication](#) [publications](#) [Baromètre de la Science Ouverte](#)

Producteur : **Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation**
Licence :  **Licence Ouverte v2.0 (Etalab)**
Statistiques : **902 888 enregistrements, 3 792 téléchargements**
Actualisation : **il y a 2 ans**
Accès directs : [Informations](#) | [Tableau](#) | [Export de fichiers](#) | [API](#)

| doi | title | year | journal_name | publisher | bso_classification | bso_classe_encoded |
|------------------------------------|---|------|---|------------------------------------|--------------------|--------------------|
| 10.1001/jama.2016.0278 | Association of Admission to Veterans Affairs H... | 2016 | JAMA | American Medical Association (AMA) | Medical research | 7 |
| 10.1016/j.diagmicrobio.2016.04.014 | Prospective multicenter surveillance identifie... | 2016 | Diagnostic Microbiology and Infectious Disease | Elsevier BV | Medical research | 7 |
| 10.1177/0018720816651536 | Effects of Epinephrine Auto-Injector Shape and... | 2016 | Human Factors: The Journal of the Human Factor... | SAGE Publications | Biology (fond.) | 0 |
| 10.1016/j.jacc.2016.03.507 | Association of Guideline-Based Admission Treat... | 2016 | Journal of the American College of Cardiology | Elsevier BV | Medical research | 7 |
| 10.1111/1475-6773.12455 | A Randomized, Controlled Trial of a Shared Pan... | 2016 | Health Services Research | Wiley | Medical research | 7 |

```
df_bso_filtered['cleaned_feature'].head(5)
```

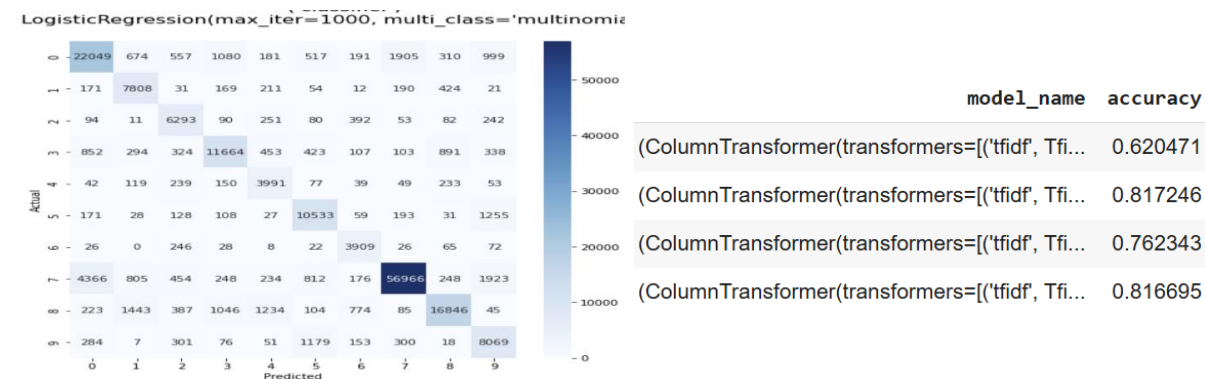
```
0    association admission veteran affair hospital ...
1    prospective multicenter surveillance identifie...
2    effect epinephrine auto injector shape size hu...
3    association guideline based admission treatmen...
4    randomized controlled trial shared panel manag...
Name: cleaned_feature, dtype: object
```

- Entraînement de plusieurs algorithmes de classification supervisée : forêts aléatoires de type classifieur, classificateur Naive Bayes Multinomial, régression logistique multinomiale...

- Choix de l'algorithme le plus performant sur le dataset de test

```
df_result = pd.DataFrame(columns=['model_name', 'accuracy'])
classifiers = [
    RandomForestClassifier(n_estimators=100, max_depth=5, random_state=0),
    LinearSVC(),
    MultinomialNB(),
    LogisticRegression(multi_class='multinomial', solver='lbfgs',max_iter=1000),
]

for classifier in classifiers:
    preprocessor = ColumnTransformer(
        transformers=[
            ('tfidf', TfidfVectorizer(), 'cleaned_feature'), #TfidfVectorizer accepts c
        ],
    )
    pipe = pipeline.Pipeline(
        steps=[
            ('preprocessor', preprocessor),
            #('pca', TruncatedSVD(n_components=5, random_state=42)),
            ('classifier', classifier),
        ],
    )
    model = pipe.fit(X_train_miss, y_train_miss.values.ravel())
    y_pred = model.predict(X_test)
    accuracy = metrics.accuracy_score(y_test,y_pred)
    print(print_confusion_matrix(model,y_pred))
    df_result = df_result.append([{"model_name" : model, "accuracy" : accuracy}])
```



- Tuning des hyperparamètres du modèle choisi pour en optimiser la précision

- Prédiction sur de nouvelles données

```
logmodel_v2.predict(pd.DataFrame(data={'cleaned_feature': ["law norm experimental evidence liability rule"]})))
array(['9'], dtype=object)
```

- Réconciliation des jeux de données étiquetés

- Dashboard mis à jour

| | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0 | 0.78 | 0.77 | 0.78 | 28463 |
| 1 | 0.70 | 0.86 | 0.77 | 9091 |
| 2 | 0.70 | 0.83 | 0.76 | 7588 |
| 3 | 0.80 | 0.76 | 0.77 | 15449 |
| 4 | 0.60 | 0.80 | 0.69 | 4992 |
| 5 | 0.76 | 0.84 | 0.80 | 12533 |
| 6 | 0.67 | 0.89 | 0.77 | 4402 |
| 7 | 0.95 | 0.86 | 0.90 | 66232 |
| 8 | 0.88 | 0.76 | 0.82 | 22187 |
| 9 | 0.62 | 0.77 | 0.69 | 10438 |
| accuracy | | | 0.82 | 181375 |
| macro avg | 0.75 | 0.81 | 0.77 | 181375 |
| weighted avg | 0.83 | 0.82 | 0.82 | 181375 |

| | source | local | mesri |
|--|--------|-------|-------|
| bsc_classification_fr | | | |
| Biologie fondamentale | | 397 | 2476 |
| Chimie | | 39 | 533 |
| Informatique et sciences de l'information | | 1184 | 1107 |
| Ingénierie | | 169 | 413 |
| Mathématiques | | 317 | 792 |
| Recherche médicale | | 532 | 2783 |
| Sciences de la Terre, écologie, énergie et biologie appliquée | | 564 | 2013 |
| Sciences humaines | | 193 | 501 |
| Sciences physiques et astronomie | | 1146 | 1899 |
| Sciences sociales | | 163 | 727 |

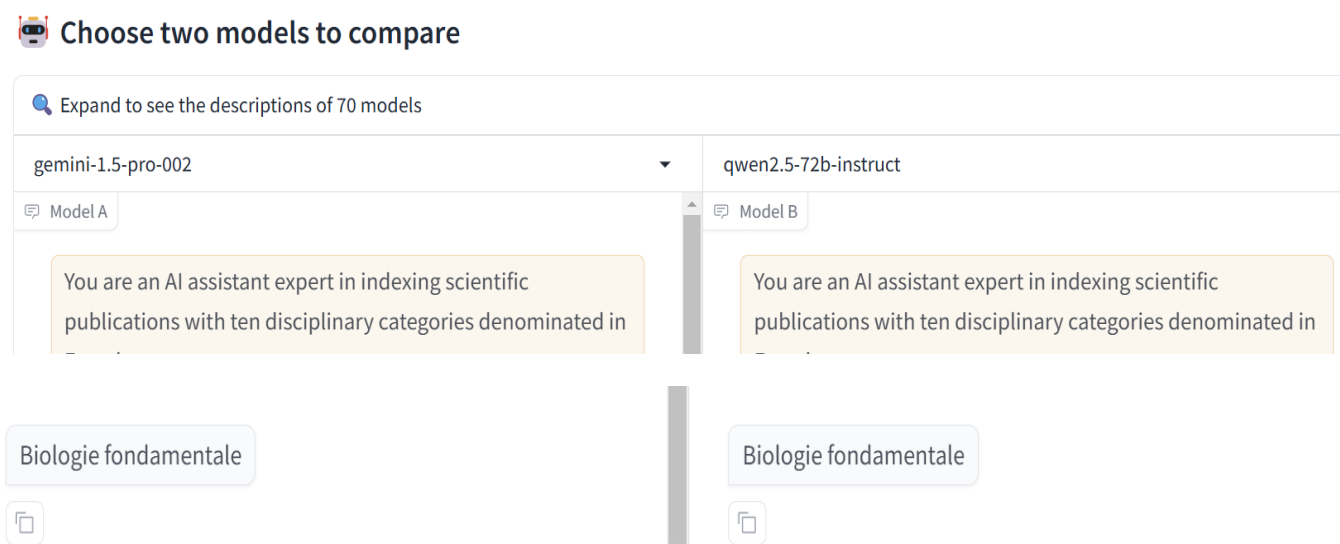


Et si c'était à refaire en 2024 ?

- modèle de langage pré-entraîné au lieu d'un apprentissage from scratch
- LLM : réseau de neurones Transformers capable de faire de l'encodage positionnel sur des séquences de texte au lieu d'un encodage mot à mot
- embeddings sémantiques

Option 1 : quelques tests manuels d'inférence avec plusieurs LLMs

- Chatbot Arena : <https://lmarena.ai/>



You are an AI assistant expert in indexing scientific publications with ten disciplinary categories denominated in French :

- Biologie fondamentale
- Chimie
- Informatique et sciences de l'information
- Ingénierie
- Mathématiques
- Recherche médicale
- Sciences de la Terre, écologie, énergie et biologie appliquée
- Sciences humaines
- Sciences physiques et astronomie
- Sciences sociales

The user gives as input some bibliographic metadata of a publication and you answer contains as output the correct category to classify the article.

****CONSTRAINT:****

Please only suggest the category from the user's input. Do not include the user's input in your response or additional text, comments, or literals.

Option 2 : tests semi-automatisés sur plusieurs modèles

- sur un subset d'une centaine de publications
- en zero-shot prompting et en few-shot prompting
- avec différentes stratégies de conversion record-to-text
- avec un client Python comme point d'entrée unique sur plusieurs serveurs d'API

```
You are an AI assistant expert in indexing scientific publications with ten disciplinary categories denominated in French :  
- Biologie fondamentale  
- Chimie  
- Informatique et sciences de l'information  
- Ingénierie  
- Mathématiques  
- Recherche médicale  
- Sciences de la Terre, écologie, énergie et biologie appliquée  
- Sciences humaine  
- Sciences physiques et astronomie  
- Sciences sociales
```

The user gives as input some bibliographic metadata of a publication and you answer contains as output the correct category to classify the article.

```
**CONSTRAINT:**  
Please only suggest the category from the user's input. Do not include the user's input in your response or additional text, comments, or literals.
```

```
**EXAMPLES**  
Here are a few examples of classified publications:
```

```
**User**:  
Publication title: Retour d'expérience sur la réingénierie de blocs opératoires en partie neuve et en rénovation.  
Journal name: IRBM News.  
Publisher: Elsevier BV.  
Publication in Open Access: False  
**Assistant**: Recherche médicale  
-----
```

```
**User**:  
Publication title: Fatigability in Patients With Multiple Sclerosis During Maximal Concentric Contractions.  
Journal name: Archives of Physical Medicine and Rehabilitation.  
Publisher: Elsevier BV.  
Publication in Open Access: False  
**Assistant**: Recherche médicale  
-----
```

```
You are an AI assistant expert in indexing scientific publications with the ten following disciplinary categories :  
- Biology (fond.)  
- Chemistry  
- Computer and information sciences  
- Engineering  
- Mathematic  
- Medical research  
- Earth, Ecology, Energy and applied biology  
- Humanities  
- Physical sciences and Astronomy  
- Social sciences
```

The user gives as input some bibliographic metadata of a publication and you answer contains as output the correct category to classify the article.

```
**CONSTRAINT:**  
Please only suggest the category from the user's input. Do not include the user's input in your response or additional text, comments, or literals.
```

```
**EXAMPLES**  
Here are a few examples of classified publications:
```

```
**User**:  
The title of the publication is: Retour d'expérience sur la réingénierie de blocs opératoires en partie neuve et en rénovation.  
The publication has been published in the academic journal: IRBM News.  
The publisher is: Elsevier BV.  
The publication is not in open access  
**Assistant**: Medical research  
-----
```

```
**User**:  
The title of the publication is: Fatigability in Patients With Multiple Sclerosis During Maximal Concentric Contractions.  
The publication has been published in the academic journal: Archives of Physical Medicine and Rehabilitation.  
The publisher is: Elsevier BV.  
The publication is not in open access  
**Assistant**: Medical research  
-----
```


Option 2 : résultats

| | openai_gpt-4o | openai_gpt-4o-mini | groq_llama3-70b-8192 | groq_llama3-8b-8192 | groq_mixtral-8x7b-32768 | groq_gemma2-9b-it | text_metadata | bso_classification_fr |
|-----|----------------------------------|---|---|---|---|--|---|---|
| 0 | Recherche médicale | Informatique et sciences de l'information | Informatique et sciences de l'information | Informatique et sciences de l'information | Informatique et sciences de l'information | Informatique et sciences de l'information \n | Publication title: Metastable Resting State Br... | Biologie fondamentale |
| 1 | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale \n | Publication title: Management of Accessory Ren... | Recherche médicale |
| 2 | Sciences humaines | Sciences humaines | Sciences humaines | Sciences humaines | Sciences sociales | Recherche médicale \n | Publication title: Can the Stereotype Threat a... | Sciences humaines |
| 3 | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale \n | Publication title: Targeting the TREK-1 potass... | Biologie fondamentale |
| 4 | Sciences sociales | Sciences sociales | Sciences sociales | Ingénierie | Sciences sociales | Ingénierie \n | Publication title: The rise and fall of R&D ne... | Sciences sociales |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 95 | Sciences physiques et astronomie | Sciences physiques et astronomie | Sciences physiques et astronomie | Sciences physiques et astronomie | Sciences physiques et astronomie | Sciences physiques et astronomie \n | Publication title: Formation of compact system... | Sciences de la Terre, écologie, énergie et bio... |
| 96 | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale | Recherche médicale \n | Publication title: Investigation of Plasma Inf... | Recherche médicale |

```

Model: openai_gpt-4o
Precision: 0.6340
Recall: 0.6132
F1-score: 0.5854
---
Model: openai_gpt-4o-mini
Precision: 0.5292
Recall: 0.5466
F1-score: 0.5164
---
Model: groq_llama3-70b-8192
Precision: 0.5506
Recall: 0.4845
F1-score: 0.4767
---
Model: groq_llama3-8b-8192
Precision: 0.5550
Recall: 0.5160
F1-score: 0.4864
---
Model: groq_mixtral-8x7b-32768
Precision: 0.3042
Recall: 0.1984
F1-score: 0.1970
---
Model: groq_gemma2-9b-it
Precision: 0.5004
Recall: 0.4600
F1-score: 0.4555
---

```

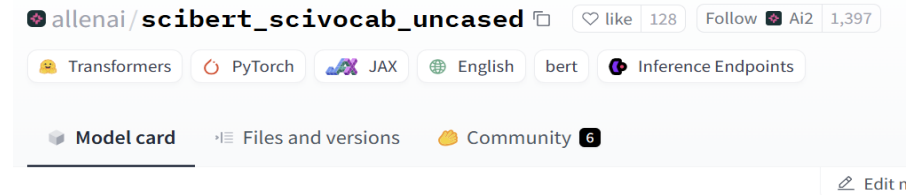
```

Model: openai_gpt-4o
Precision: 0.4943
Recall: 0.4945
F1-score: 0.4645
---
Model: openai_gpt-4o-mini
Precision: 0.6464
Recall: 0.5620
F1-score: 0.5733
---
Model: groq_llama3-70b-8192
Precision: 0.5690
Recall: 0.4368
F1-score: 0.4477
---
Model: groq_llama3-8b-8192
Precision: 0.3488
Recall: 0.3562
F1-score: 0.3370
---
Model: groq_mixtral-8x7b-32768
Precision: 0.1088
Recall: 0.0321
F1-score: 0.0417
---
Model: groq_gemma2-9b-it
Precision: 0.5972
Recall: 0.5018
F1-score: 0.5091
---

```

Option 3 : fine-tuning de SLM

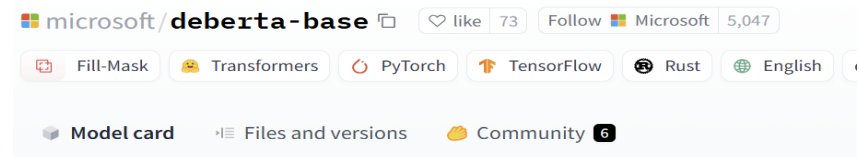
- Small Language Models : modèles de langage légers souvent pré-entraînés sur des tâches spécifiques
- Approche zero-shot : [joeddav/xlm-roberta-large-xnli](#)
- (Full) fine-tuning : [allenai/scibert_scivocab_uncased](#) et [microsoft/deberta-base](#)
- Dataset : 50k publications (source : BSO)



SciBERT

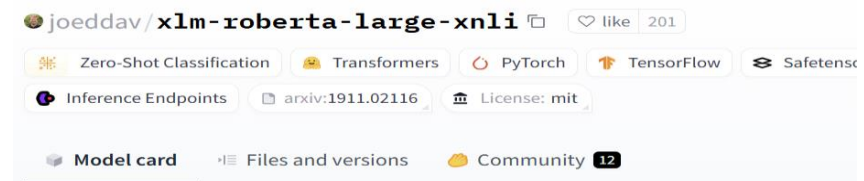
This is the pretrained model presented in [SciBERT: A Pretrained Language Model for Scientific Text](#), which is a BERT model trained on scientific text.

The training corpus was papers taken from [Semantic Scholar](#). Corpus size is 1.14M papers, 3.1B tokens. We use the full text of the papers in training, not just abstracts.



DeBERTa: Decoding-enhanced BERT with Disentangled Attention

[DeBERTa](#) improves the BERT and RoBERTa models using disentangled attention and enhanced mask decoder. It outperforms BERT and RoBERTa on majority of NLU tasks with 80GB training data.



xlm-roberta-large-xnli

Model Description

This model takes [xlm-roberta-large](#) and fine-tunes it on a combination of NLI data in 15 languages. It is intended to be used for zero-shot text classification, such as with the Hugging Face [ZeroShotClassificationPipeline](#).

Option 3 : résultats

- xml-roberta-large

```
Precision: 0.6225  
Recall: 0.4456  
F1-score: 0.4359  
---
```

- DeBERTA finetuning

```
🔄 Precision: 0.6282  
Recall: 0.5965  
F1-score: 0.6004  
---
```

- SciBERT (test sur dataset 5k avec abstract)

```
trainer.train()
```

```
[1760/1760 25:22, Epoch 4/4]
```

| Epoch | Training Loss | Validation Loss | Accuracy | F1 | Precision | Recall |
|-------|---------------|-----------------|----------|----------|-----------|----------|
| 1 | 0.884700 | 0.959680 | 0.652750 | 0.642428 | 0.652864 | 0.652750 |
| 2 | 0.905000 | 0.943745 | 0.655401 | 0.644442 | 0.662422 | 0.655401 |
| 3 | 0.864500 | 0.925826 | 0.660040 | 0.653625 | 0.657308 | 0.660040 |
| 4 | 0.660100 | 0.943427 | 0.669980 | 0.663700 | 0.669931 | 0.669980 |

```
from sklearn.metrics import precision_score, recall_score, f1_score  
  
y_true = eval_df["bso_classification_en"]  
y_pred = eval_df["predicted_label"]  
  
# Handle None values in y_pred before calculating metrics  
y_pred = y_pred.fillna('unknown') # Replace None with a string like 'unknown'  
  
# Calculate precision, recall, and F1-score (macro-averaged)  
precision = precision_score(y_true, y_pred, average='macro', zero_division=0)  
recall = recall_score(y_true, y_pred, average='macro', zero_division=0)  
f1 = f1_score(y_true, y_pred, average='macro', zero_division=0)  
  
print(f"Precision: {precision:.4f}")  
print(f"Recall: {recall:.4f}")  
print(f"F1-score: {f1:.4f}")  
print("----")
```

```
Precision: 0.7209  
Recall: 0.7269  
F1-score: 0.7114  
---
```

Option 3 : avantages

- Publication du dataset d'entraînement et du modèle finetuné sur HuggingFace
- Documentation du code sur Kaggle

```
scibert-finetuned-publications-classificati... Draft saved
File Edit View Run Settings Add-ons Help
+ - ✂ 📄 ▶ ▶▶ Run All Markdown - ● Draft Session (26m) H D C P U R A M :
Full fine-tuning : SciBERT
[ ]: import pandas as pd
import torch
from transformers import AutoModelForSequenceClassification, AutoTokenizer, Tr
from datasets import Dataset, load_dataset, DatasetDict
from sklearn.metrics import accuracy_score, precision_recall_fscore_support

Load model
[ ]: NUM_CLASSES = 10
MODEL_NAME = "allenai/scibert_scivocab_uncased"
tokenizer = AutoTokenizer.from_pretrained(MODEL_NAME)
model = AutoModelForSequenceClassification.from_pretrained(MODEL_NAME, num_label
```

Models 3



Geraldine/scibert-finetuned-bso-publications...

Text Classification • Updated about 18 hours ago • ↓ 10

Geraldine/msmarco-distilbert-base-v4-ead

Feature Extraction • Updated 11 days ago • ↓ 33

Datasets 4



Geraldine/bso-publications-indexation-50k

Viewer • Updated about 21 hours ago • 📄 50k • ↓ 10

Geraldine/Ead-Instruct-10k

Viewer • Updated 1 day ago • 📄 10k • ↓ 11

Option 4 : approche hybride

- Combinaison d'un algorithme de ML appliqué sur des données textuelles encodées avec le modèle d'embeddings sémantiques de DeBERTa

Hybrid Approach Example: DeBERTa Embeddings + Logistic Regression

1. Extract Contextual Embeddings from DeBERTa

DeBERTa can be used to generate sentence embeddings (representing the entire input text). These embeddings are then used as input features for a Logistic Regression classifier.

2. Train a Logistic Regression Model

Use the extracted embeddings as features to train a simple Logistic Regression model for classification.

```
Precision: 0.6225
```

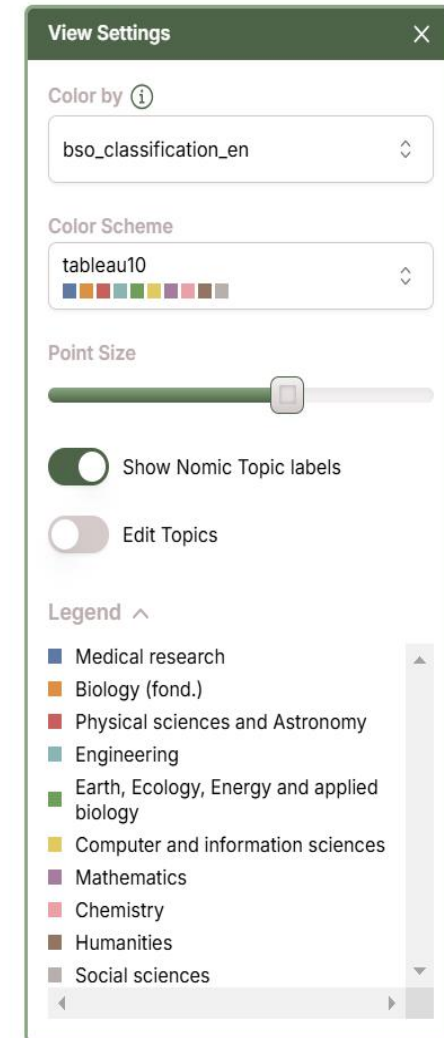
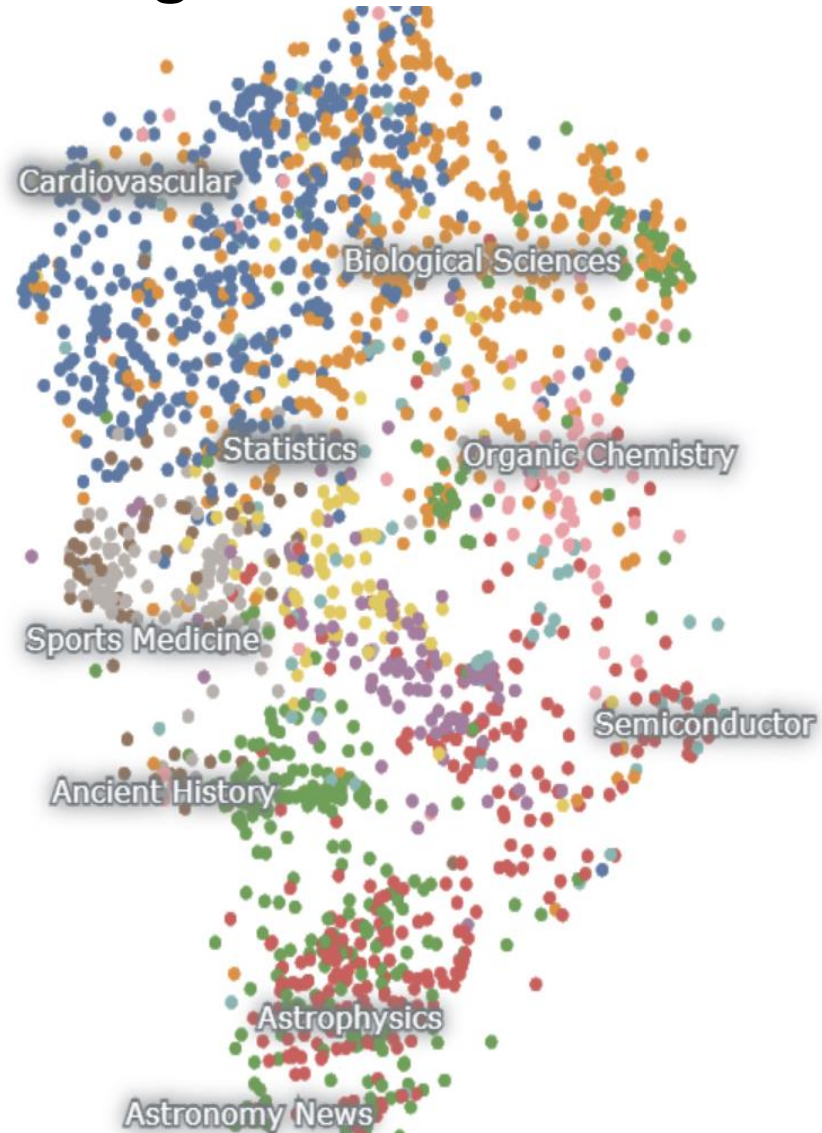
```
Recall: 0.6055
```

```
F1-score: 0.5794
```

```
---
```

Bonus : Dataviz des embeddings

- Nomic Atlas



Conclusion : usage raisonné de l'IA générative

Les modèles d'IA générative vont être très utiles pour :

- faire du RAG (chatbots) sur nos procédures et wikis internes
- faire du RAG à destination des utilisateurs sur nos documentations publiques (sites web, documentation technique...)
- faire du GraphRAG sur nos métadonnées pour créer des moteurs de recherche sémantique
- faire de la cartographie sémantique de collections
- extraire des données structurées à partir de données non structurées (LLMs et VLMs)
- ...

MAIS (selon les tâches) : plus les données sont structurées, plus le ML classique ou les approches hybrides sont efficaces.