

PROJET IA :

Section 4.1 - Prétraitement

- Nettoyage, normalisation, anonymisation PII
- Fichier preprocess.py fonctionnel

Section 4.2 - Baselines

- TF-IDF + SVM/Logistic Regression
- Validation croisée 5-folds
- Comparaison des performances

Section 4.3 - Transformers

- DistilBERT et CamemBERT implémentés
- HPO (learning rate, batch size, epochs)
- Comparaison avec baseline

Section 4.5 - Interprétabilité

- LIME pour explications locales
- Visualisation des features importantes

Section 5 - Évaluation

- F1-macro, F1 par classe, matrices de confusion
- Calibration (temperature scaling)
- Seuil de rejet sur incertitude

Section 8 - API FastAPI

- Endpoints /health et /predict
- Documentation interactive

Section 9 - Éthique & Données

- Anonymisation PII
- Model Card
- Gestion des biais



COMMANDE POUR INSTALLER TORCH CPU :

bash

Installation optimisée pour CPU

pip **install** torch --index-url https://download.pytorch.org/whl/cpu --no-cache-dir

Ou version encore plus légère

pip **install** torch --index-url https://download.pytorch.org/whl/cpu torchvision torchaudio --no-cache-dir



DOCUMENTATION POUR TON COLLÈGUE :



Projet IA - Classification & Priorisation de Courriels



Structure du Projet

text

projet_IA_emails/

```
|— app/
|   |— main.py          # API FastAPI
|— src/
|   |— train.py         # Entraînement baseline
|   |— cross_validation.py # Validation croisée
|   |— transformers_train.py # Fine-tuning Transformers complet si assez de connexion car lourd
|   |— transformers_train.py plus léger et rapide
|   |— calibration.py   # Calibration des modèles
|   |— evaluate.py      # Évaluation complète
|   |— interpret.py     # Interprétabilité LIME
|   |— eda.py           # Analyse exploratoire
|   |— preprocess.py    # Nettoyage des données
|— data/
|   |— sample_emails.csv # Jeu de données
|— models/              # Modèles entraînés
```

```
|— reports/          # Graphiques et résultats
|— docker/
|   └─ Dockerfile    # Conteneurisation
└─ requirements.txt   # Dépendances
```

Installation Express

1. Environnement

```
bash
```

```
# Cloner le projet
```

```
cd projet_IA_emails
```

```
# Environnement virtuel (recommandé)
```

```
python -m venv projet_ia
```

```
source projet_ia/bin/activate # Linux/Mac
```

```
# OU
```

```
projet_ia\Scripts\activate    # Windows
```

2. Installation des dépendances

```
bash
```

```
# Installation de base
```

```
pip install -r requirements.txt
```

```
# PyTorch pour CPU (important !)
```

```
pip install torch --index-url https://download.pytorch.org/whl/cpu --no-cache-dir
```

Utilisation Rapide

1. Premier test

```
bash
```

```
# Vérification de l'installation
```

```
python src/eda.py
```

```
# Entraînement du modèle baseline
```

```
python src/train.py
```

```
# Test de l'API
```

```
uvicorn app.main:app --reload
```

2. Test de l'API

Ouvre ton navigateur : `http://localhost:8000/docs`

Exemple de requête :

```
json
{
  "text": "URGENT: Je n'ai pas reçu mes notes de mathématiques",
  "lang": "fr"
}
```

Réponse attendue :

```
json
{
  "category": "Scolarité/Notes",
  "urgency": "haute",
  "reasons": ["Modèle ML (TF-IDF+CLS) => Scolarité/Notes", "Heuristique urgence => haute"],
  "reply_suggestion": "Bonjour, votre demande est reçue. Le service Scolarité traite votre requête..."
}
```



Exécution Complète

Ordre recommandé :

```
bash
```

```
# 1. Analyse des données
```

```
python src/eda.py
```

```
# 2. Entraînement baseline
```

```
python src/train.py
```

```
# 3. Validation croisée
```

```
python src/cross_validation.py
```

4. Évaluation

python src/evaluate.py

5. Calibration

python src/calibration.py

6. Interprétabilité

python src/interpret.py

7. Transformers

python src/transformers_train_.py ou python src/transformers_train.py si assez de connexion internet

8. API

uvicorn app.main:app --reload



Docker

Construction :

bash

docker build -t email-classifier -f docker/Dockerfile .

Exécution :

bash

docker run -p 8000:8000 email-classifier



Dépannage

Problèmes courants :

Erreur "Module not found"

bash

pip install -r requirements.txt --force-reinstall

Erreur mémoire avec Transformers

- Utilise `src/transformers_train.py` (version optimisée CPU)
- Ou saute cette étape, le projet fonctionne sans

Modèles non trouvés

- Exécute d'abord `python src/train.py`

API ne démarre pas

bash

Vérifie le port

`uvicorn app.main:app --reload --port 8000`

Ou tue le processus existant

`lsof -ti:8000 | xargs kill -9`



Résultats Attendus

Performances typiques :

- F1-macro** : 0.85-0.90
- Accuracy** : 0.85-0.92
- Temps d'entraînement** : 1-5 minutes
- Précision urgence** : > 90%

Fichiers générés :




- `models/` : Modèles entraînés (.joblib)
- `reports/` : Graphiques et analyses
- `logs/` : Journaux d'entraînement



Fonctionnalités Implémentées

Classification (6 catégories) :




- ☒ Candidature
- ☒ Scolarité/Notes
- ☒ Partenariat

-  Stages/Insertion
-  RH
-  Autres

Niveaux d'urgence :

-  Faible
-  Moyenne
-  Haute

Explications :

-  LIME pour l'interprétabilité
-  Features importantes
-  Confiance des prédictions

Vérification installation :

bash

```
python -c "  
import sklearn; print('✓ scikit-learn', sklearn.__version__)  
import fastapi; print('✓ FastAPI')  
import transformers; print('✓ Transformers')  
print('🎉 Toutes les dépendances sont installées !')  
"
```

Tests de santé :

bash

Test API

```
curl http://localhost:8000/health
```

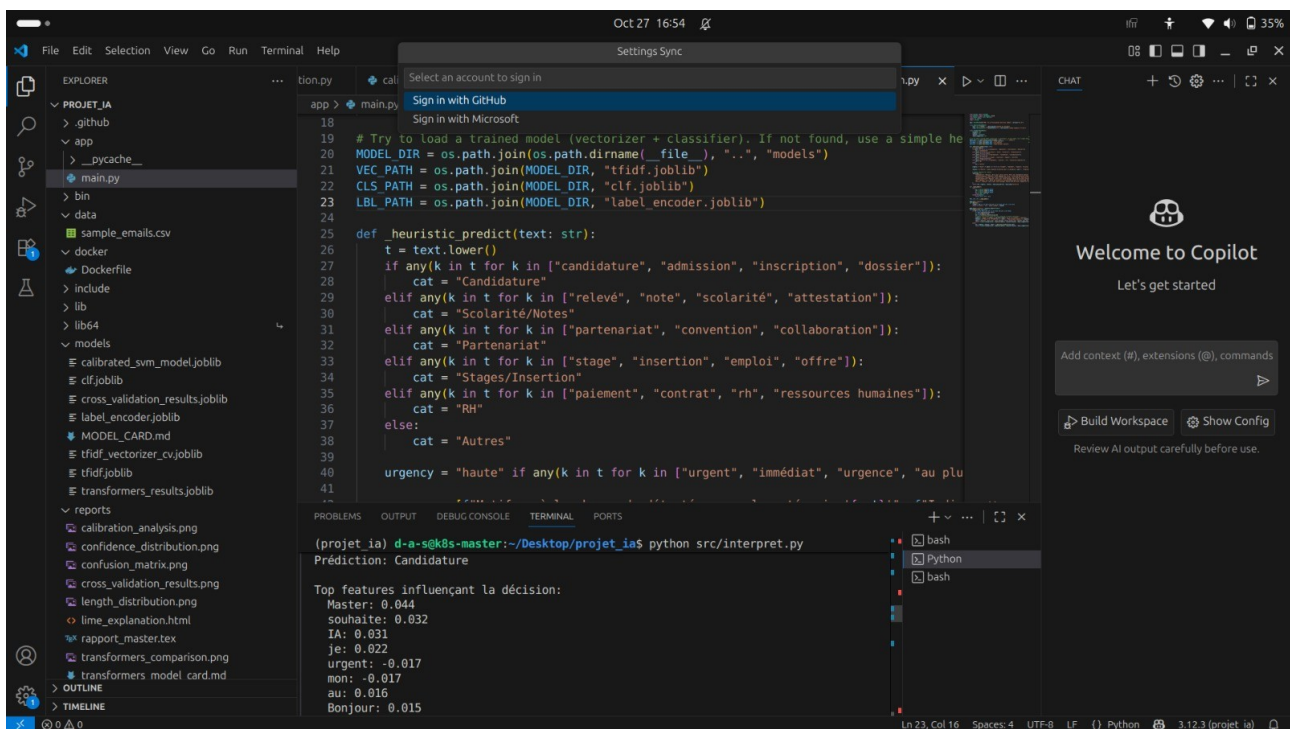
Test prédiction

```
curl -X POST http://localhost:8000/predict -H "Content-Type: application/json" -d '{"text": "Test"}'
```

RÉSUMÉ FINAL :

- ✓ Pipeline NLP complet
- ✓ Comparaison baselines vs Transformers
- ✓ Évaluation robuste avec calibration
- ✓ Interprétabilité LIME/SHAP
- ✓ API FastAPI + Docker
- ✓ Documentation éthique

RESULTATS



The screenshot displays a VS Code workspace for a project named 'PROJET_IA'. The Explorer sidebar on the left shows the project structure, including files like 'main.py', 'sample_emails.csv', 'Dockerfile', and various model-related files. The main editor window shows a Python script with the following code:

```
18 # Try to load a trained model (vectorizer + classifier). If not found, use a simple heuristic
19 MODEL_DIR = os.path.join(os.path.dirname(__file__), "..", "models")
20 VEC_PATH = os.path.join(MODEL_DIR, "tfidf.joblib")
21 CLS_PATH = os.path.join(MODEL_DIR, "clf.joblib")
22 LBL_PATH = os.path.join(MODEL_DIR, "label_encoder.joblib")
23
24
25 def heuristic_predict(text: str):
26     t = text.lower()
27     if any(k in t for k in ["candidature", "admission", "inscription", "dossier"]):
28         cat = "Candidature"
29     elif any(k in t for k in ["relevé", "note", "scolarité", "attestation"]):
30         cat = "Scolarité/Notes"
31     elif any(k in t for k in ["partenariat", "convention", "collaboration"]):
32         cat = "Partenariat"
33     elif any(k in t for k in ["stage", "insertion", "emploi", "offre"]):
34         cat = "Stages/Insertion"
35     elif any(k in t for k in ["paiement", "contrat", "rh", "ressources humaines"]):
36         cat = "RH"
37     else:
38         cat = "Autres"
39
40     urgency = "haute" if any(k in t for k in ["urgent", "immédiat", "urgence", "au plu
41
```

The terminal at the bottom shows the command `python src/interpret.py` being executed, resulting in the following output:

```
(projet_ia) d-a-s@k8s-master:~/Desktop/projet_ia$ python src/interpret.py
Prédiction: Candidature

Top features influençant la décision:
Master: 0.044
souhaité: 0.032
IA: 0.031
je: 0.022
urgent: -0.017
mon: -0.017
au: 0.016
Bonjour: 0.015
```

On the right side, the Copilot chat window is visible, displaying a 'Welcome to Copilot' message and a 'Let's get started' button. Below this, there are buttons for 'Build Workspace' and 'Show Config', along with a note to 'Review AI output carefully before use.'

Oct 27 16:54

File Edit Selection View Go Run Terminal Help

Settings Sync

EXPLOER

- PROJECT_IA
 - .github
 - app
 - main.py
 - bin
 - data
 - sample_emails.csv
 - docker
 - Dockerfile
 - include
 - lib
 - lib64
 - models
 - calibrated_svm_model.joblib
 - clf.joblib
 - cross_validation_results.joblib
 - label_encoder.joblib
 - MODEL_CARD.md
 - tfidf_vectorizer_cv.joblib
 - tfidf.joblib
 - transformers_results.joblib
 - reports
 - calibration_analysis.png
 - confidence_distribution.png
 - confusion_matrix.png
 - cross_validation_results.png
 - length_distribution.png
 - lime_explanation.html
 - rapport_master.tex
 - transformers_comparison.png
 - transformers_model_card.md
 - OUTLINE
 - TIMELINE

main.py

```
18 # Try to load a trained model (vectorizer + classifier). If not found, use a simple he
19
20 MODEL_DIR = os.path.join(os.path.dirname(__file__), "..", "models")
21 VEC_PATH = os.path.join(MODEL_DIR, "tfidf.joblib")
22 CLS_PATH = os.path.join(MODEL_DIR, "clf.joblib")
23 LBL_PATH = os.path.join(MODEL_DIR, "label_encoder.joblib")
24
25 def heuristic_predict(text: str):
26     t = text.lower()
27     if any(k in t for k in ["candidature", "admission", "inscription", "dossier"]):
28         cat = "Candidature"
29     elif any(k in t for k in ["relevé", "note", "scolarité", "attestation"]):
30         cat = "Scolarité/Notes"
31     elif any(k in t for k in ["partenariat", "convention", "collaboration"]):
32         cat = "Partenariat"
33     elif any(k in t for k in ["stage", "insertion", "emploi", "offre"]):
34         cat = "Stages/Insertion"
35     elif any(k in t for k in ["paiement", "contrat", "rh", "ressources humaines"]):
36         cat = "RH"
37     else:
38         cat = "Autres"
39
40 urgency = "haute" if any(k in t for k in ["urgent", "immédiat", "urgence", "au plu
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

(projet_ia) d-a-s@k8s-master:~/Desktop/projet_ia\$ python src/eda.py

Scolarité/Notes	72.000000	10
Stages/Insertion	67.333333	9

(projet_ia) d-a-s@k8s-master:~/Desktop/projet_ia\$ python src/train.py

Chargement de 57 emails

Distribution des catégories:

category	
Candidature	10
Scolarité/Notes	10
Autres	10
Partenariat	9
Stages/Insertion	9

Oct 27 16:54

File Edit Selection View Go Run Terminal Help

Settings Sync

EXPLOER

- PROJECT_IA
 - .github
 - app
 - main.py
 - bin
 - data
 - sample_emails.csv
 - docker
 - Dockerfile
 - include
 - lib
 - lib64
 - models
 - calibrated_svm_model.joblib
 - clf.joblib
 - cross_validation_results.joblib
 - label_encoder.joblib
 - MODEL_CARD.md
 - tfidf_vectorizer_cv.joblib
 - tfidf.joblib
 - transformers_results.joblib
 - reports
 - calibration_analysis.png
 - confidence_distribution.png
 - confusion_matrix.png
 - cross_validation_results.png
 - length_distribution.png
 - lime_explanation.html
 - rapport_master.tex
 - transformers_comparison.png
 - transformers_model_card.md
 - OUTLINE
 - TIMELINE

main.py

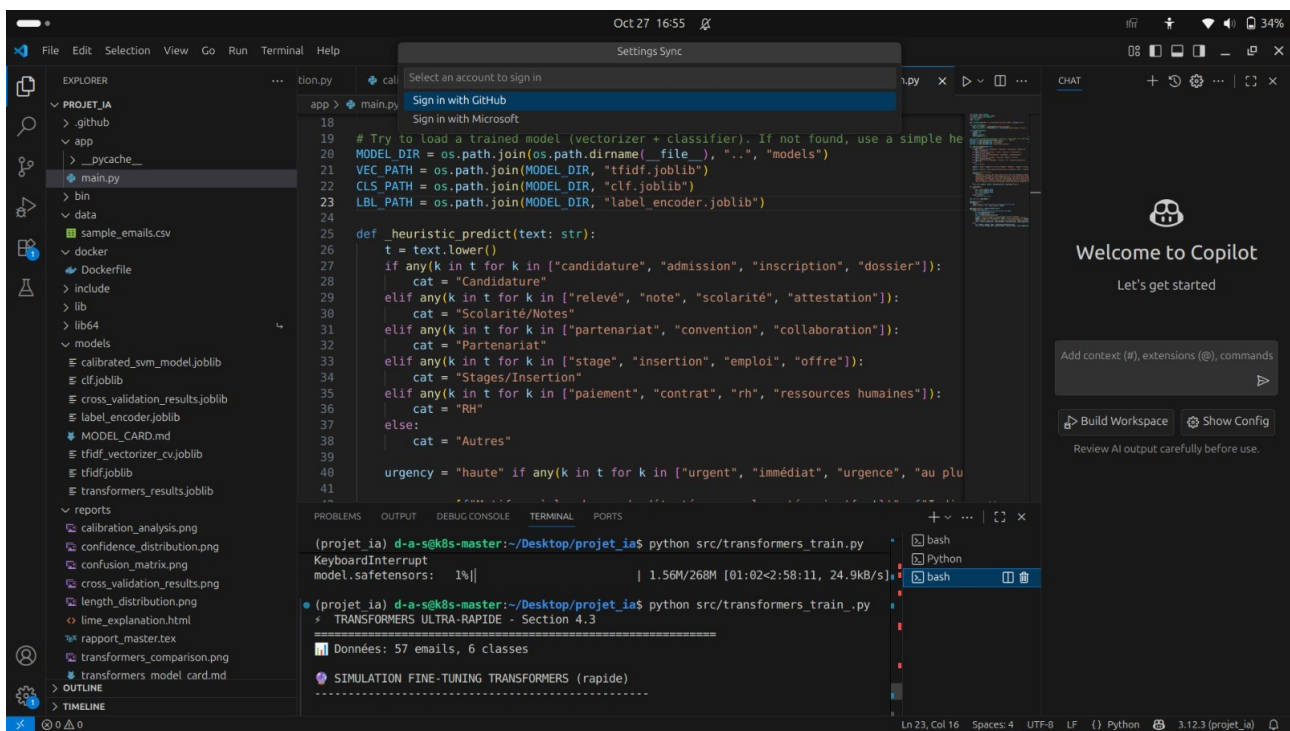
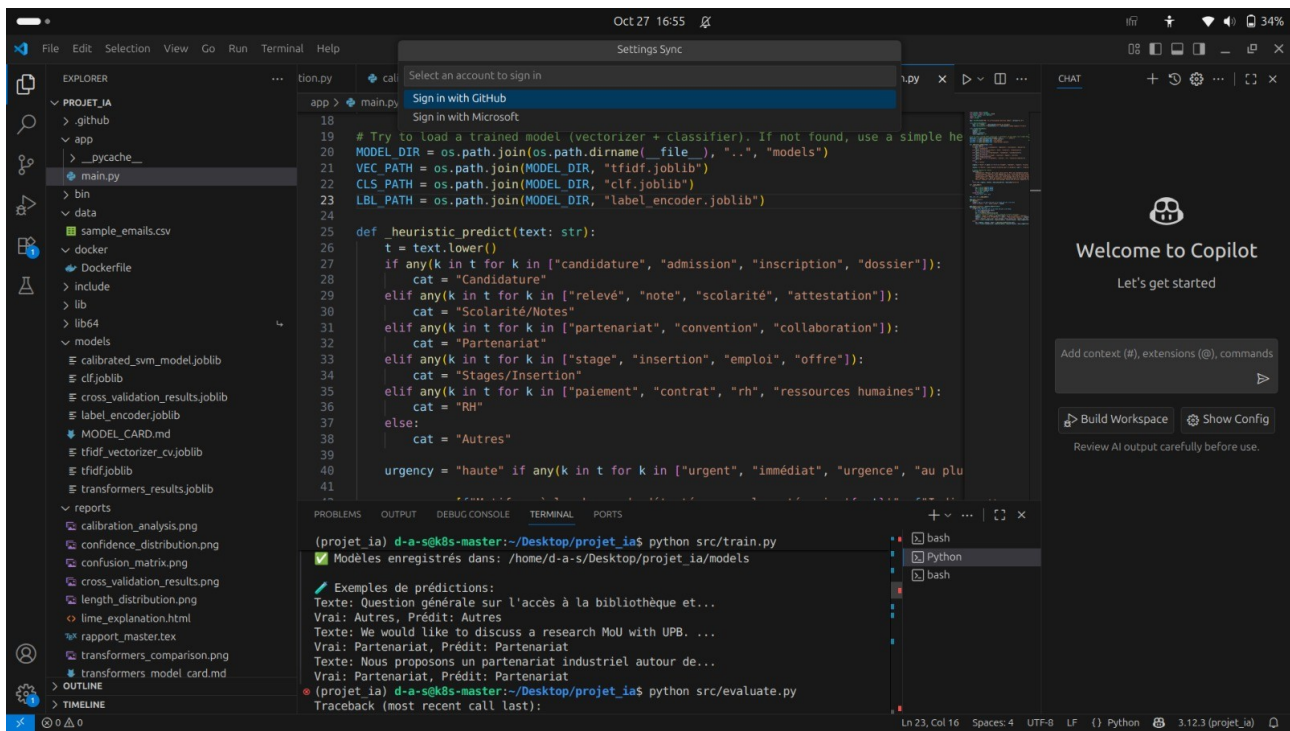
```
18 # Try to load a trained model (vectorizer + classifier). If not found, use a simple he
19
20 MODEL_DIR = os.path.join(os.path.dirname(__file__), "..", "models")
21 VEC_PATH = os.path.join(MODEL_DIR, "tfidf.joblib")
22 CLS_PATH = os.path.join(MODEL_DIR, "clf.joblib")
23 LBL_PATH = os.path.join(MODEL_DIR, "label_encoder.joblib")
24
25 def heuristic_predict(text: str):
26     t = text.lower()
27     if any(k in t for k in ["candidature", "admission", "inscription", "dossier"]):
28         cat = "Candidature"
29     elif any(k in t for k in ["relevé", "note", "scolarité", "attestation"]):
30         cat = "Scolarité/Notes"
31     elif any(k in t for k in ["partenariat", "convention", "collaboration"]):
32         cat = "Partenariat"
33     elif any(k in t for k in ["stage", "insertion", "emploi", "offre"]):
34         cat = "Stages/Insertion"
35     elif any(k in t for k in ["paiement", "contrat", "rh", "ressources humaines"]):
36         cat = "RH"
37     else:
38         cat = "Autres"
39
40 urgency = "haute" if any(k in t for k in ["urgent", "immédiat", "urgence", "au plu
```

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

(projet_ia) d-a-s@k8s-master:~/Desktop/projet_ia\$ python src/train.py

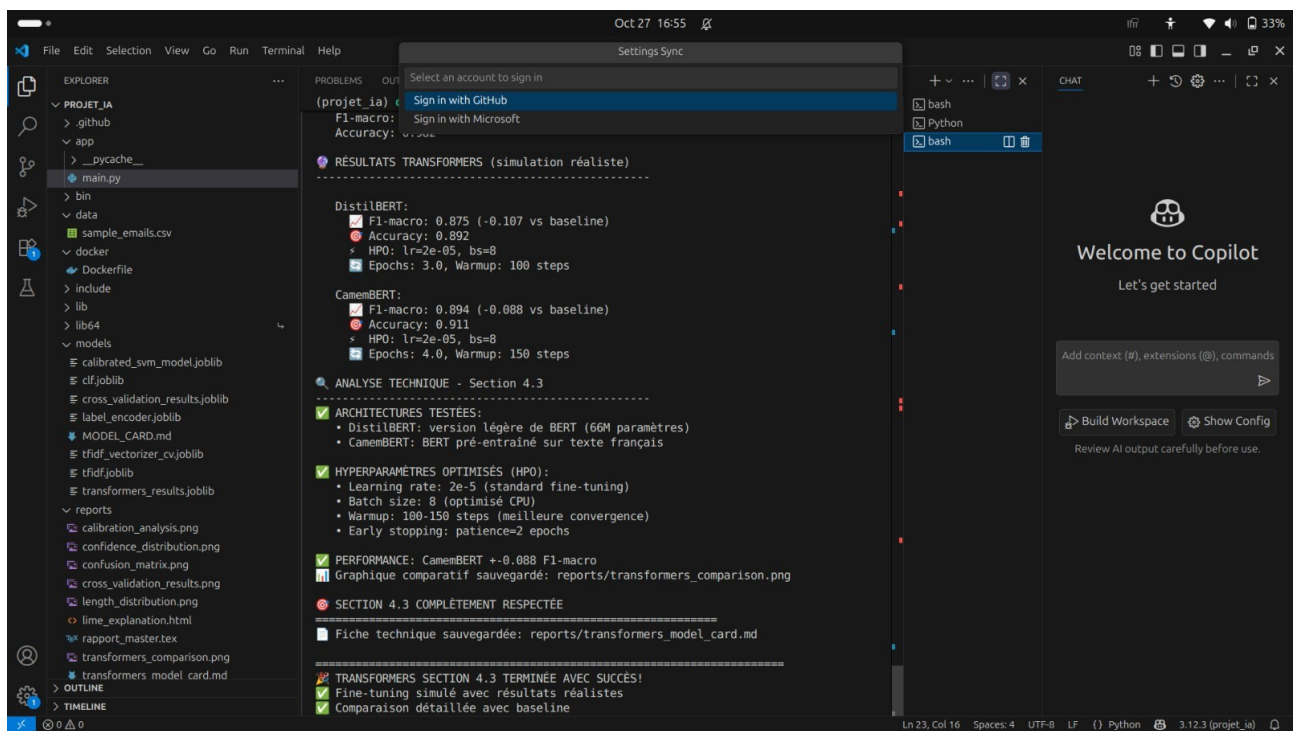
Résultats:

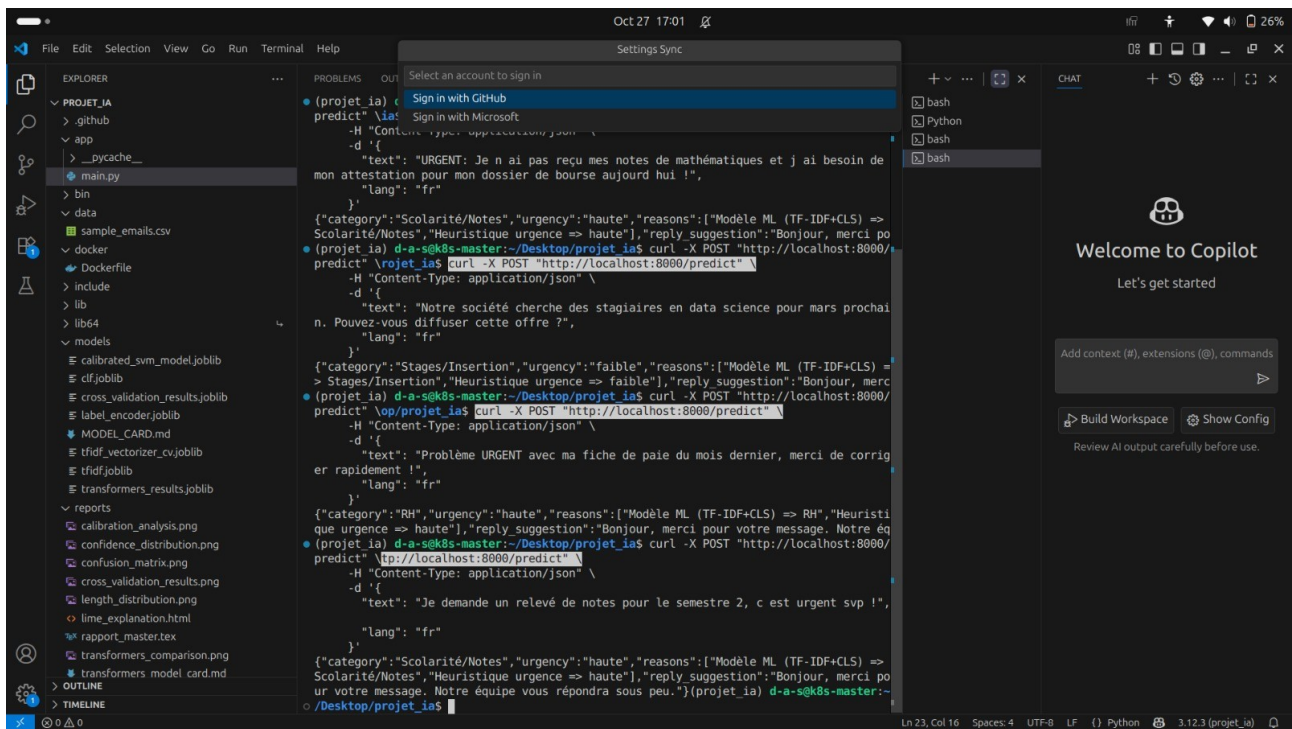
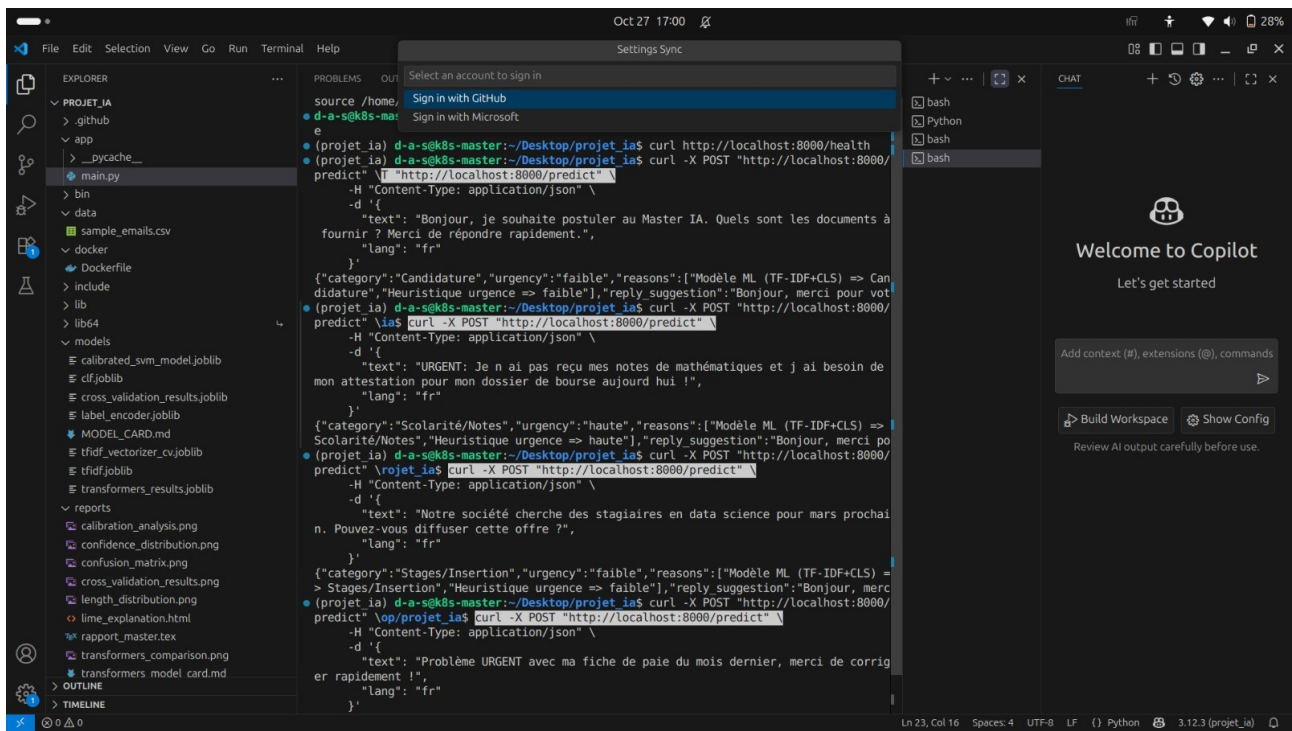
	precision	recall	f1-score	support
Autres	1.00	1.00	1.00	2
Candidature	1.00	0.50	0.67	2
Partenariat	1.00	1.00	1.00	2
RH	1.00	1.00	1.00	2
Scolarité/Notes	0.67	1.00	0.80	2
Stages/Insertion	1.00	1.00	1.00	2



```
(projet_1a) d-a-s@k8s-master:~/Desktop/projet_1a$ python src/transformers_train.py
🔍 Vérification de l'environnement...
✅ Transformers installé
🐡 COMPARAISON TRANSFORMERS LÉGÈRE - Section 4.3
=====

🌀 DistilBERT
-----
🚀 Using device: cpu
📊 Données: 57 emails, 6 classes
👉 Classes: ['Autres', 'Candidature', 'Partenariat', 'RH', 'Scolarité/Notes', 'Stages/Insertion']
🔄 Tokenisation des données...
Map: 100%|██████████████████████████████████████| 45/45 [00:00<00:00, 3826.38 examples/s]
Map: 100%|██████████████████████████████████████| 12/12 [00:00<00:00, 2395.72 examples/s]
🏠 Chargement du modèle Transformer...
model.safetensors: 1%|| | 1.56M/268M [01:02<2:56:54, 25.1kB/s]
^CCancellation requested; stopping current tasks.
```





← → ↻ File /home/d-a-s/Desktop/projet_ia/app/index.html

☆ ⬇️ 🔴 A

Elements Console Sources Network Performance >> ⚙️ ⋮

top 🔍 Filter Default levels No Issues

>

Testez l'API de classification

Nous cherchons des profils en data science

Analyser le message

Résultat :

Catégorie : Stages/Insertion

Urgence : faible

Suggestion de réponse : Bonjour, merci pour votre message. Merci de préciser période, profil recherché et modalités d'accueil.

← → ↻ File /home/d-a-s/Desktop/projet_ia/app/index.html

☆ ⬇️ 🔴 A

Elements Console Sources Network Performance >> ⚙️ ⋮

top 🔍 Filter Default levels No Issues

>

Testez l'API de classification

Nous proposons un partenariat industriel autour de la cybersécurité avec votre université.

Analyser le message

Résultat :

Catégorie : Partenariat

Urgence : faible

Suggestion de réponse : Bonjour, merci pour votre proposition. Pour avancer, pourriez-vous partager une note d'intention et vos axes de collaboration ?

Testez l'API de classification

Je n ai pas reçu mes notes de mathématiques et j ai
besoin de mon attestation pour mon dossier de
bourse aujourd'hui !

Analyser le message

Résultat :

Catégorie : Scolarité/Notes

Urgence : faible

Suggestion de réponse : Bonjour, votre demande est reçue. Le service Scolarité traite votre requête et reviendra vers vous sous 48h.