

Classification du Texte AI vs Humain

I. Introduction

Le développement rapide des modèles d'intelligence artificielle (IA) a entraîné une prolifération de textes générés automatiquement. Ces textes peuvent imiter des écrits humains avec une grande précision, ce qui soulève des préoccupations concernant leur utilisation dans des contextes sensibles, tels que la désinformation. Ce projet vise à résoudre le problème de classification de texte, où l'objectif est de déterminer si un texte a été généré par une IA ou par un humain. Pour cela, un modèle d'apprentissage profond a été développé, utilisant un sous-ensemble équilibré d'un jeu de données contenant des textes générés par IA et par des humains.

II. Objectif du Projet

L'objectif principal de ce projet est de développer un modèle capable de classer automatiquement des textes en deux catégories :

- Texte généré par IA
- Texte généré par un humain

III. Dataset

Le dataset utilisé pour ce projet est intitulé "AI vs Human Text". Il est disponible publiquement sur Kaggle sous le nom AI_Human.csv

Ce dataset contient des échantillons de textes générés soit par des modèles d'Intelligence Artificielle (comme ChatGPT ou d'autres modèles de génération de texte), soit écrits par des êtres humains.

Chaque ligne du dataset est composée de deux colonnes principales :

- **text** : le contenu textuel lui-même,
generated : une étiquette binaire indiquant l'origine du texte :
 - 1.0 → texte généré par une IA,
 - 0.0 → texte rédigé par un humain.
- **Taille initiale** : 487235
- **Classes** : Deux classes équilibrées après traitement (IA vs Humain).
- **Langue** : Principalement l'anglais.
- **Format** : CSV, encodé en UTF-8.

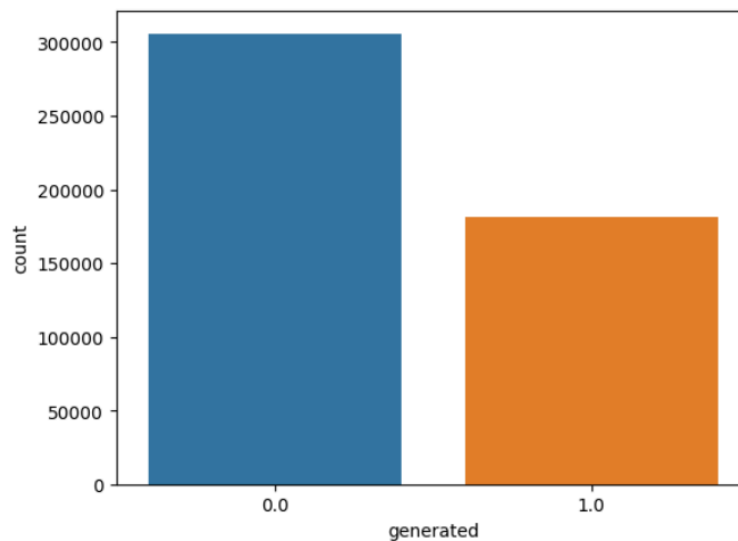
IV. Prétraitement des Données (Data Preprocessing)

1. Réduction de la Taille du Dataset

Étant donné l'énorme taille du dataset initial, il n'était pas envisageable d'utiliser l'ensemble des données dans les délais impartis et avec les ressources disponibles.

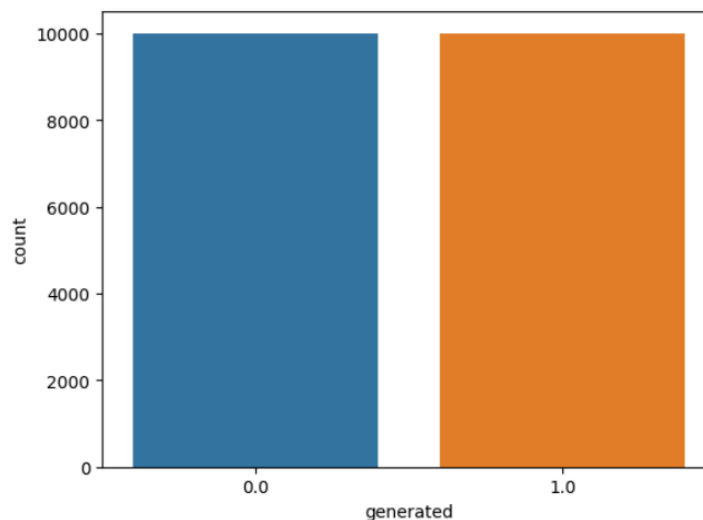
Ainsi, un **sous-échantillonnage équilibré** a été réalisé, sélectionnant **20 000 textes** au total (10 000 IA + 10 000 Humains) pour assurer un compromis entre **qualité d'apprentissage** et **temps de traitement raisonnable**.

1. dataset originale : “ai-vs-human-text”



Total : 487235
Total text by AI : 181438
Total text by Human : 305797

2. échantillon de dataset utilisé dans le projet :



Total data points after balancing: 20000
AI-generated text: 10000
Human-generated text: 10000

2. Nettoyage et Traitement Textuel

Pour garantir une qualité optimale des données textuelles avant l'entraînement, plusieurs étapes de **prétraitement linguistique** ont été appliquées :

- **Tags Removal**
Spell Check (Correction orthographique)
- **Stop Words Removal**
- **Tokenisation**
- **Troncation et Padding**
- **Encodage**

V. Modèle Utilisé

1. Choix du Modèle

Pour résoudre ce problème de classification binaire de textes, nous avons opté pour un modèle basé sur un Transformer pré-entraîné. nous avons choisi d'utiliser un modèle pré-entraîné de type **BERT Multilingue**, spécifiquement :

bert-base-multilingual-cased

2. Description du Modèle

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) est un modèle de type Transformer développé par Google. Il est basé sur un mécanisme d'attention qui traite les séquences de texte de manière bidirectionnelle, permettant ainsi une meilleure compréhension du contexte global.

Architecture :

- **12 couches (layers)**
- **12 têtes d'attention**
- **768 dimensions d'embedding**
- **Environ 110 millions de paramètres**

Le modèle est adapté à notre tâche de classification binaire par :

- L'ajout d'une couche de classification dense (fully connected) avec une sortie unique (sigmoïde) pour prédire la probabilité que le texte soit généré par une IA.
- Un fine-tuning du modèle : seules les couches finales de BERT et la couche de classification ont été entraînées afin de :
 - Réduire les besoins computationnels
 - Minimiser le risque d'overfitting
 - Accélérer le processus d'apprentissage compte tenu des ressources et du temps disponibles

VI. Entraînement et validation du modèle

1. Notebook :

<https://www.kaggle.com/code/yasminebabdelkader/human-vs-ai-text-detecy-tradrly>

VII. test du modèle

```
test_acc, _ = eval_model(
    model,
    test_data_loader,
    loss_fn,
    device,
    len(df_test)
)

test_acc.item()
```

Evaluating: 100% |██████████| 375/375 [00:45<00:00, 8.26batch/s, loss=0.0688]
0.9906666666666666

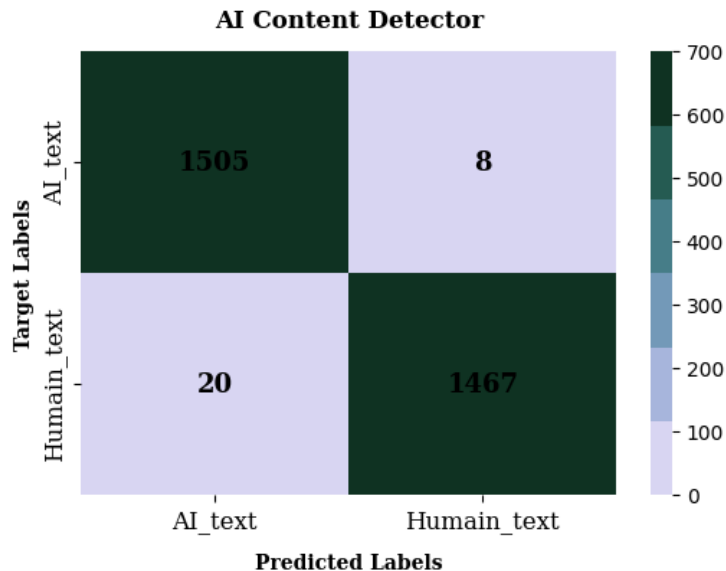
→ l'évaluation du modèle sur les données de test donne un excellent résultat de 99,07% de précision.

```
print(classification_report(y_test, y_pred, target_names=class_names, digits=4))
```

	precision	recall	f1-score	support
AI_text	0.9869	0.9947	0.9908	1513
Humain_text	0.9946	0.9866	0.9905	1487
accuracy			0.9907	3000
macro avg	0.9907	0.9906	0.9907	3000
weighted avg	0.9907	0.9907	0.9907	3000

→ Le texte AI est détecté avec une précision de 98,69% et un rappel de 99,47%, tandis que le texte humain est identifié avec une précision de 99,46% et un rappel de 98,66%.

VIII. matrice de confusion



→ Sur 1513 textes créés par IA, 1505 ont été correctement identifiés et 8 ont été incorrectement classés comme textes humains. Sur 1487 textes humains, 1467 ont été correctement identifiés et 20 ont été incorrectement classés comme textes IA. Cette visualisation confirme la haute précision du modèle dans la distinction entre contenu généré par IA et contenu écrit par des humains.

IX. Déploiement du Modèle

Dans le but de rendre le modèle accessible de manière pratique, nous avons initié le développement d'une plateforme web pour le déployer.

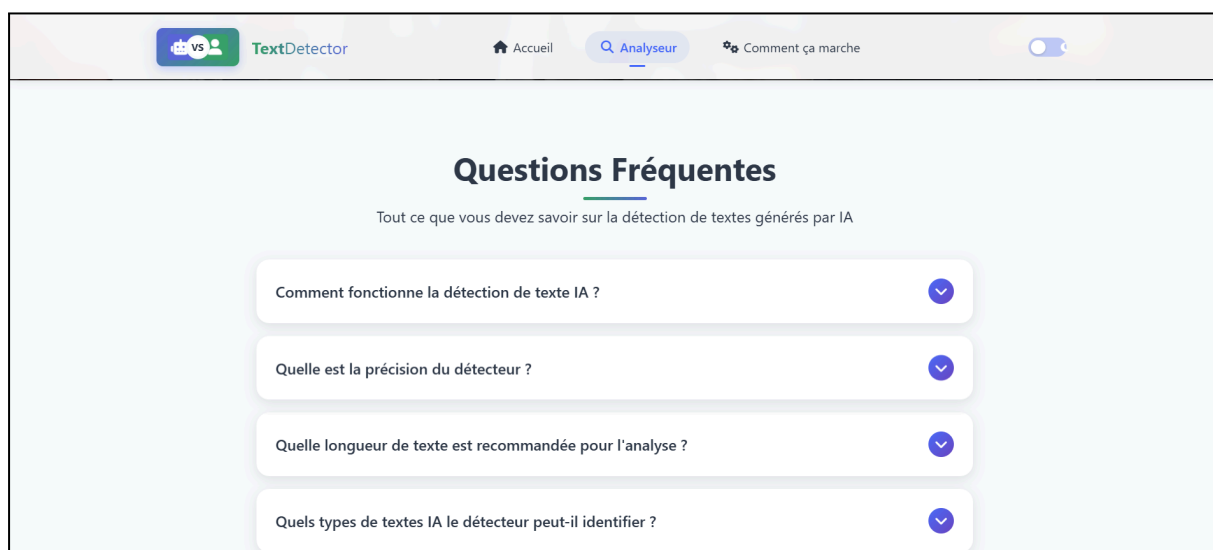
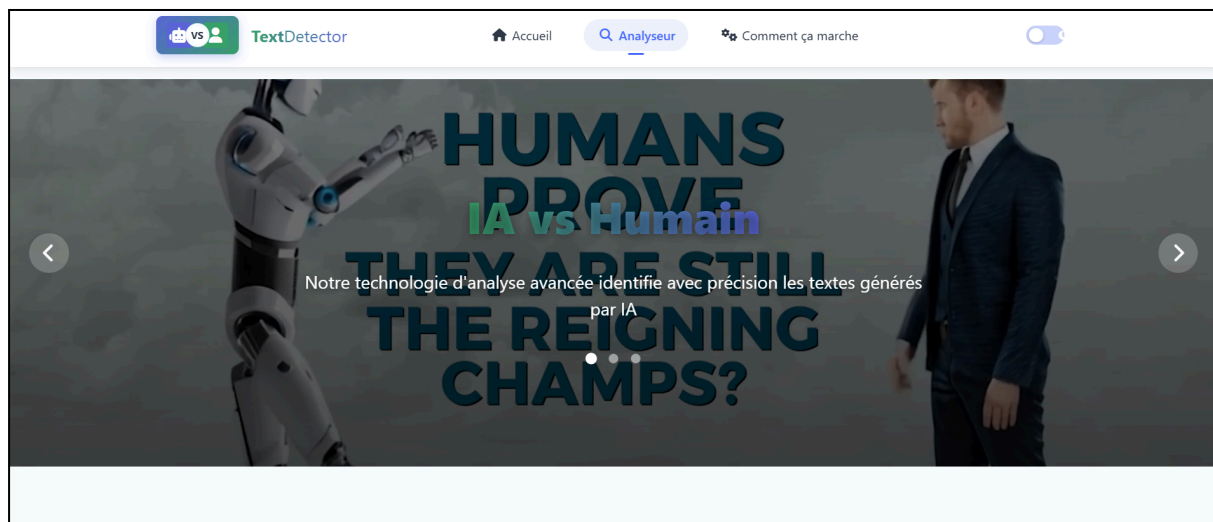
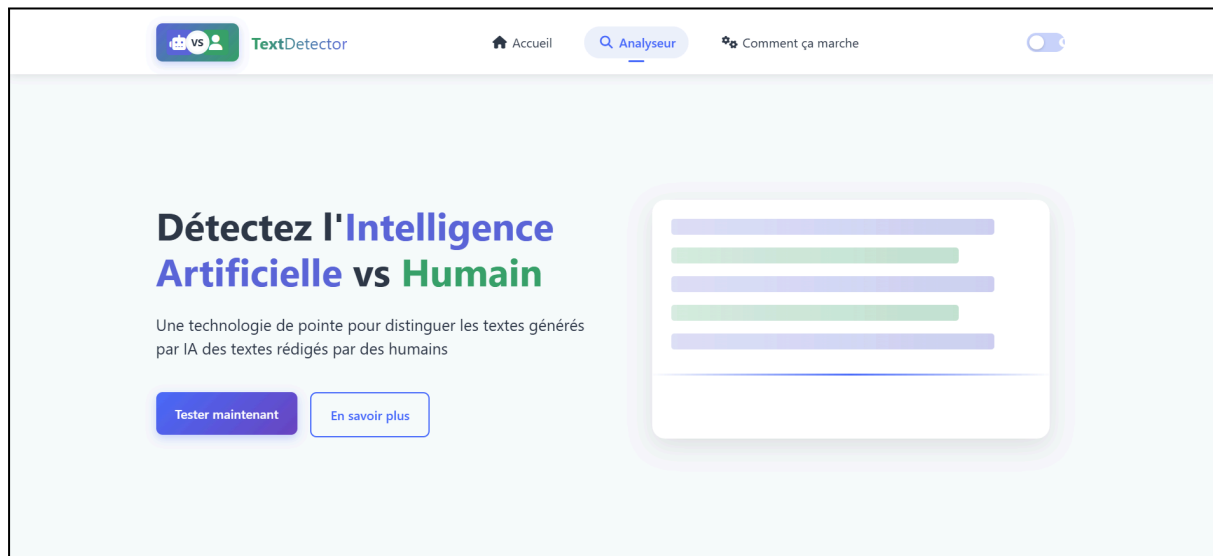
Cette plateforme est en cours de création en utilisant les technologies suivantes :


- HTML et CSS pour la partie front-end (interface utilisateur).
- Flask pour le back-end (serveur web léger en Python).

Intégration du modèle BERT pour faire des prédictions directement à partir de l'interface.


L'objectif est de permettre aux utilisateurs d'entrer un texte et de recevoir instantanément la prédiction du modèle indiquant si le texte est généré par une intelligence artificielle ou par un humain.

La plateforme est actuellement en cours de finalisation, et représente une étape importante vers la mise en production de notre solution.



 **TextDetector**

[Accueil](#) [Analyseur](#) [Comment ça marche](#)



Entrez votre texte

Collez ou rédigez le texte que vous souhaitez analyser (300 mots minimum recommandés)

Collez ici le texte que vous souhaitez analyser...

Coller du presse-papier

Effacer

Texte d'exemple

Analyser le texte

Résultats de l'analyse

Consultez les résultats détaillés de l'analyse de votre texte

Aucune analyse effectuée

Entrez un texte dans le panneau de gauche et cliquez sur "Analyser" pour commencer