

ANALYSE FINANCIÈRE AVEC NLP

Données textuelles pour les prévisions de marché et la gestion des risques

Présenté par : Nazir YAYE YOUSOUUF et Osman SAID ALI

PLAN

01

Introduction

02

Analyse de sentiment

03

Reconnaissance des entités nommées

04

Extraction des relations

05

Difficultés rencontrées

INTRODUCTION

L'Impact des Mots sur les Marchés

Pourquoi ?

En finance, les marchés digèrent en temps réel tout signal textuel.

Quand un tweet fait trembler Wall Street

“ Un matin d'avril 2025, un simple message de Donald Trump évoquant une possible suspension des droits de douane bouleverse les marchés.



En quelques minutes, plusieurs actions grimpent de 3%.



Deux jours plus tard, le président se rétracte, et le marché efface tous ses gains.

OBJECTIFS & CONTEXTE

Explorer le potentiel des techniques modernes de TAL pour comprendre et exploiter les données textuelles en contexte financier



Analyse de sentiments



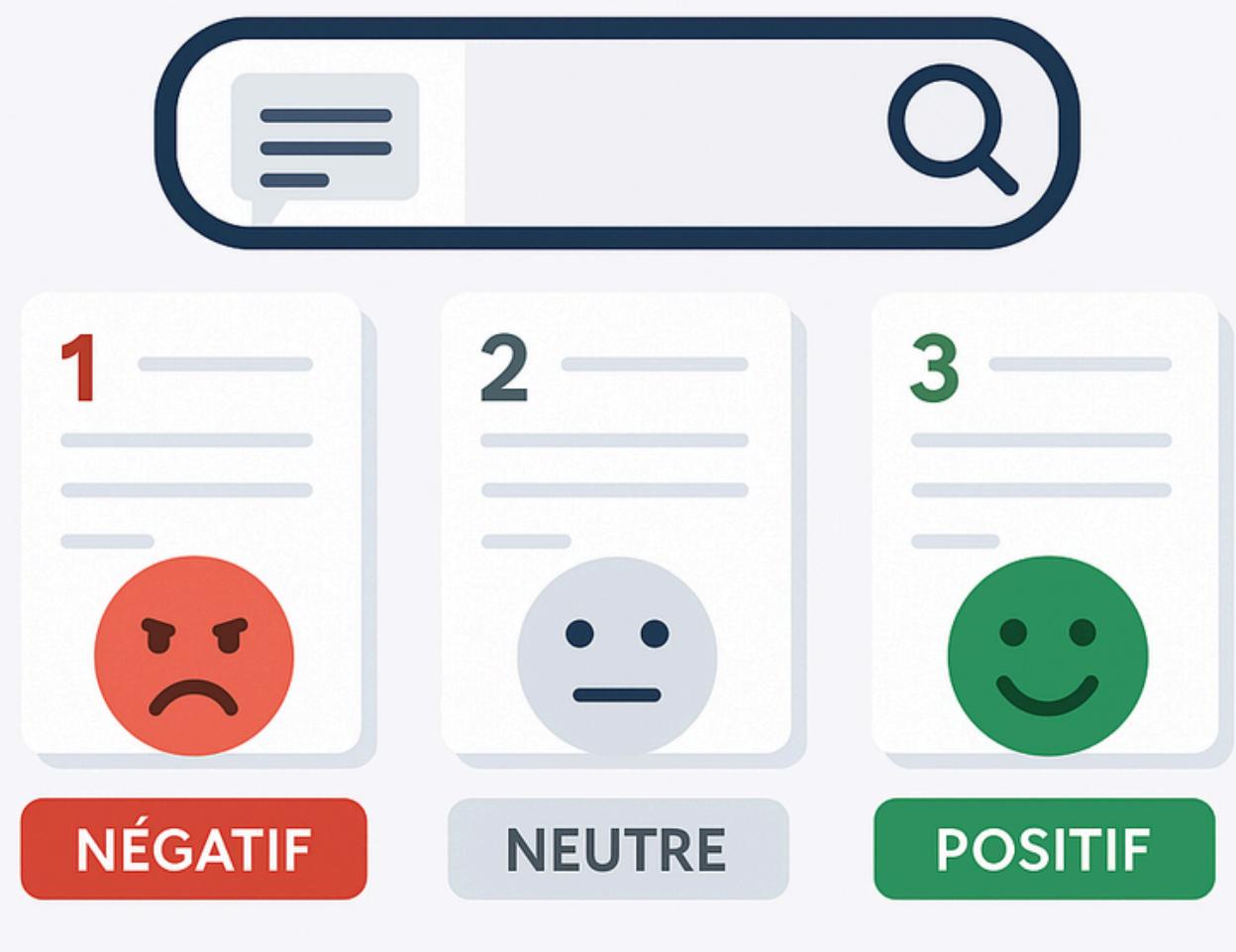
Reconnaisances des entités nommées



Extraction des relations

ANALYSE DE SENTIMENT

L'analyse de sentiment consiste à catégoriser des textes selon la tonalité émotionnelle qu'ils véhiculent.



Analyse de sentiments

Données

Dataset 1

Twitter

+

Dataset 2

FinGPT

Données fusionnées



Prétraitement

Scénario 1

Garder les stop-words
Suppression de ponctuations
Vectorisation TF-IDF pour les modèles classiques
Tokenisation pour les modèles avancés



Scénario 2

Suppression des stop-words
Suppression de ponctuations
Vectorisation TF-IDF pour les modèles classiques
Tokenisation pour les modèles avancés

Modèles testés

- Modèles Classiques

Modèle	Avec stop-words		Sans stop-words	
	Accuracy	F1-score macro	Accuracy	F1-score macro
SVM	0.900	0.894	0.887	0.881
Random Forest	0.911	0.902	0.921	0.913
XGBoost	0.851	0.839	0.837	0.825
LightGBM	0.860	0.851	0.847	0.837
Voting Ensemble	0.912	0.904	0.911	0.904

Modèles testés

- Transformers

- Analyse des résultats

Modèle	Avec stop-words		Sans stop-words	
	Accuracy	F1-score macro	Accuracy	F1-score macro
ProsusAI/finbert	0.948	0.945	0.940	0.936
bert-base-uncased	0.947	0.944	0.942	0.938
distilbert-base-uncased	0.943	0.941	0.939	0.935

LIME : « Qu'est-ce qui a poussé ce tweet à être classé négatif ? »

SHAP : « Quelle est la contribution équitable de chaque mot sur la décision globale et locale ? »

Reconnaissance des entités nommées

Données

Dataset 1

OntoNotes 5

+

Dataset 2

FiNER-ORD

Fusion + remapping BIO → 4 types : PER, ORG, LOC, MISC

Reconnaissance des entités nommées

Modèles testés

- Transformers & Co

Modèle	Token Acc	Precision macro	Recall macro	F1 macro
roberta-base	0.934	0.840	0.818	0.829
bert-base-cased	0.933	0.842	0.814	0.827
ProsusAI/finbert	0.918	0.839	0.798	0.818
distilbert-base-cased	0.914	0.831	0.791	0.810
CRF-baseline	0.727	0.065	0.076	0.053

Exemple d'utilisation

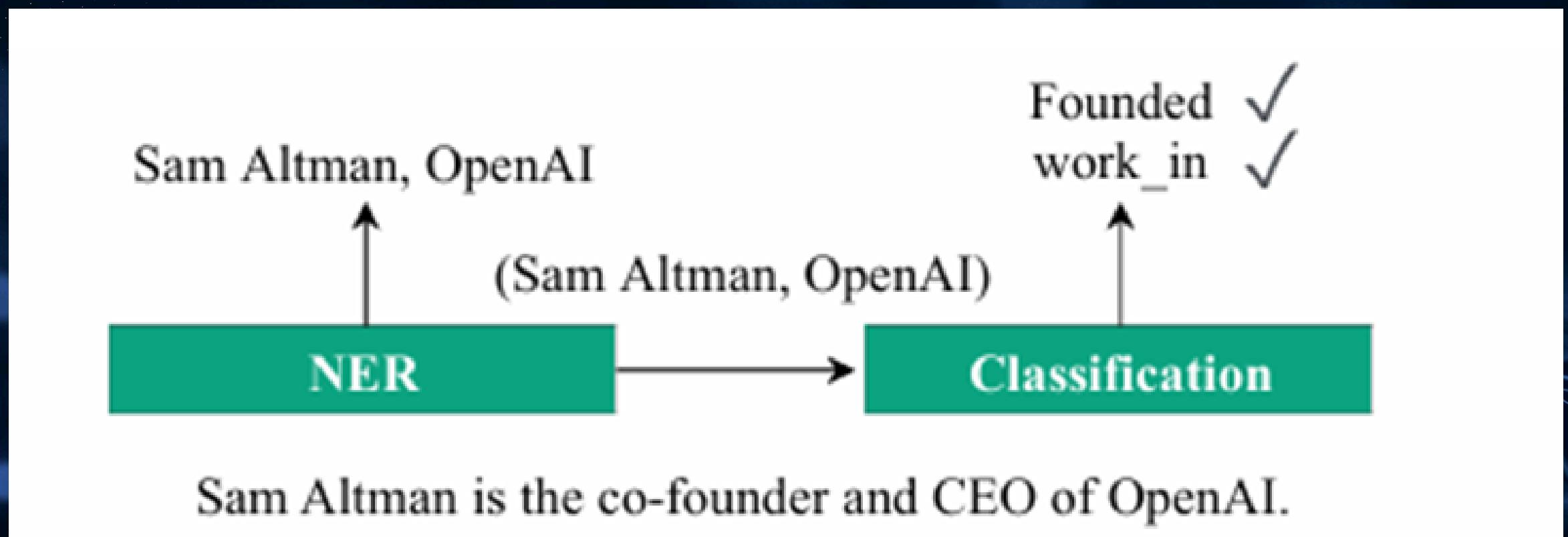
Emmanuel Macron traveled to America to visit Tesla company and Elon Musk.

Réponse :

Entité	Vérité	roberta-base
Emmanuel Macron	PER	PER
America	LOC	LOC
TESLA	ORG	ORG
Elon Musk	PER	PER



Extraction de relations

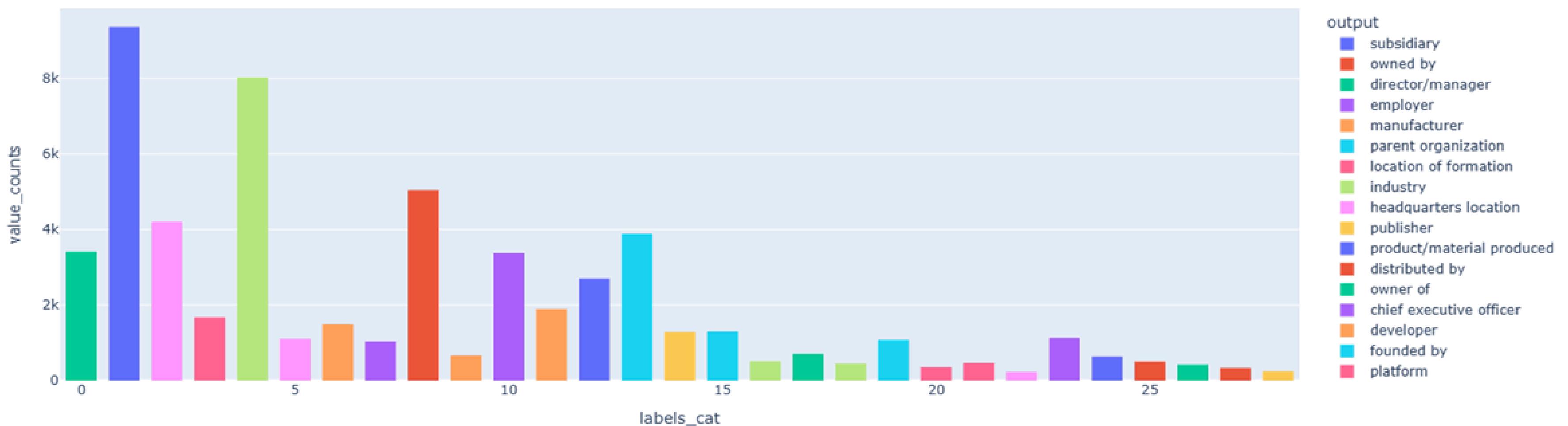


Exemple d'extraction de relation (Zhao et al. 2024)

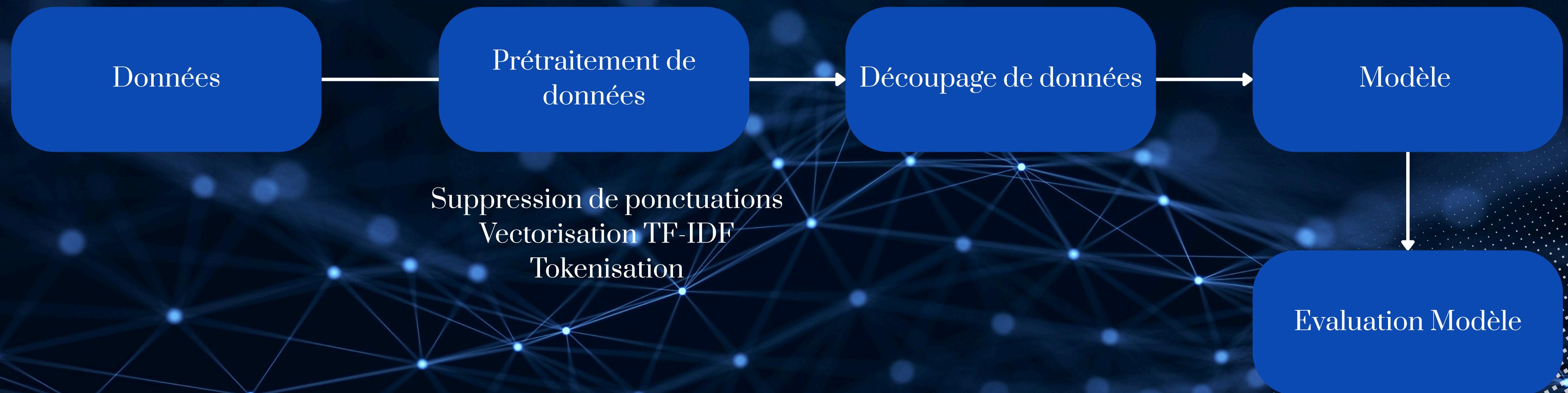
identifier
classifier
extraire
les liens entre
les entités

Données pour l'extraction des relations de FinGPT

Distribution des étiquettes dans les données pour la tâche de l'extraction de relation.



Méthodologie



Modèles testés et Performances

Modèle	Accuracy	F1-score macro	Précision macro	Rappel macro
SVM	0.708	0.685	0.692	0.684
XGBoost	0.704	0.697	0.685	0.717
Random Forest	0.707	0.689	0.689	0.694
Bert	0.929	0.850	0.890	0.829
Llama 3 8B instruct	0.976	0.935	0.937	0.933

Difficultés Rencontrées

- Principales Difficultés : Qualité des données, le traitement de certaines données, manque de ressources de calcul, pas assez de temps pour faire la tâche question-réponse
- Solutions apportées : Fusionner les données, chercher de plateforme colab, paperspace(le calcul)

Démonstration de l'application