



# ANALYSE DE SENTIMENTS GRÂCE AU DEEP LEARNING

Grégoire MUREAU





# INTRODUCTION

Problématique: **Air Paradis** souhaite maîtriser sa communication sur Twitter et se prémunir autant que possible de *bad buzzs*.

---

## Solution souhaitée

Il s'agit d'entraîner un modèle prédictif permettant d'évaluer le caractère positif ou négatif d'un tweet qui serait émis par **Air Paradis** sur son compte twitter.



# OBJECTIFS DÉTAILLÉS



Entraînement d'un modèle de prédiction :

- modèles simples, avancés, BERT
- dans un environnement MLOps



Développement d'une API de prédiction des tweets:

- conteneurisée (Docker)
- Github CI / CD



Déploiement de l'API



Challenge personnel : utilisation au maximum d'un serveur personnel:

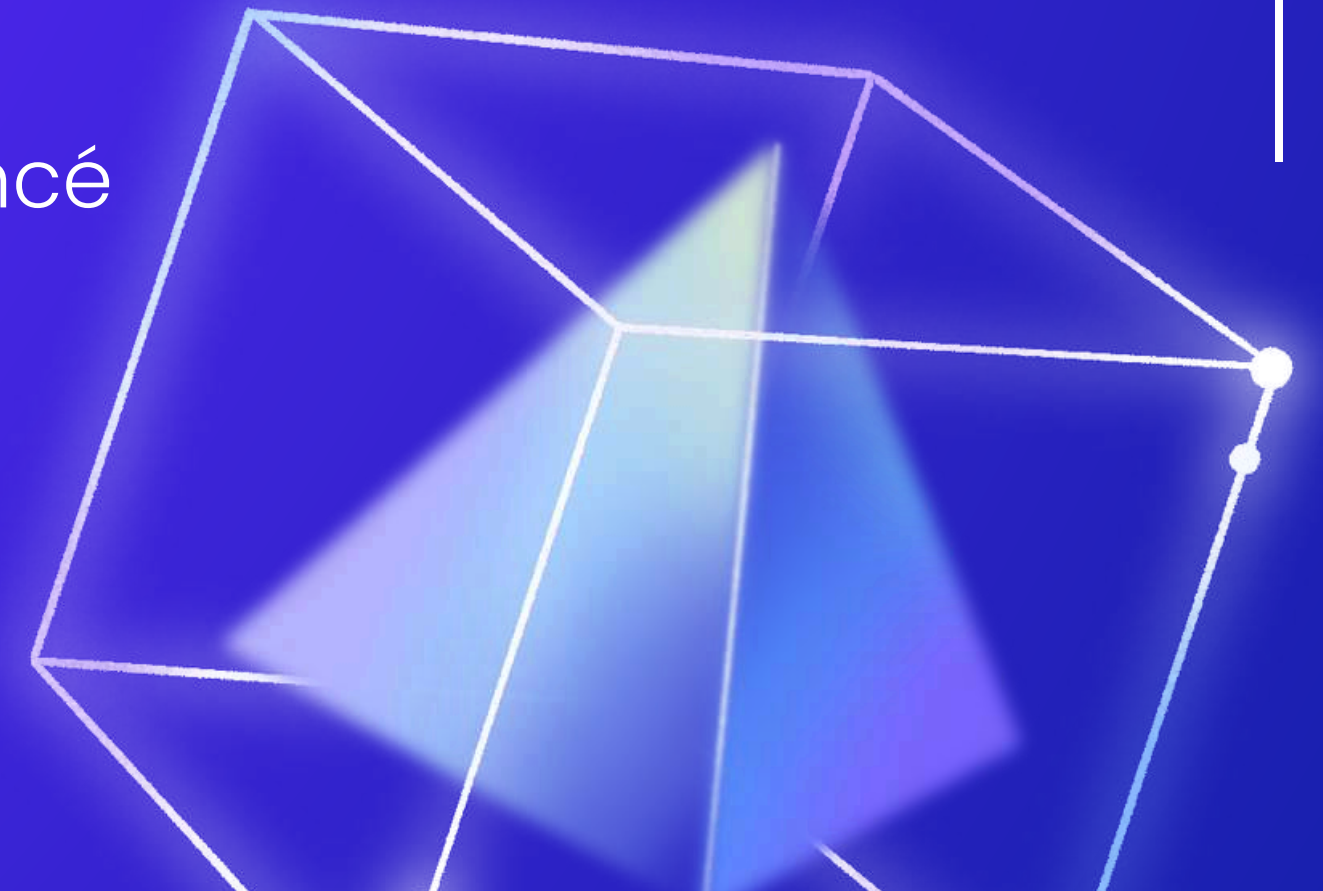
- Minimiser la dépendance aux GAFAM.
- Assurer la souveraineté des données.
- Garantir l'indépendance vis-à-vis des infrastructures des grandes entreprises.





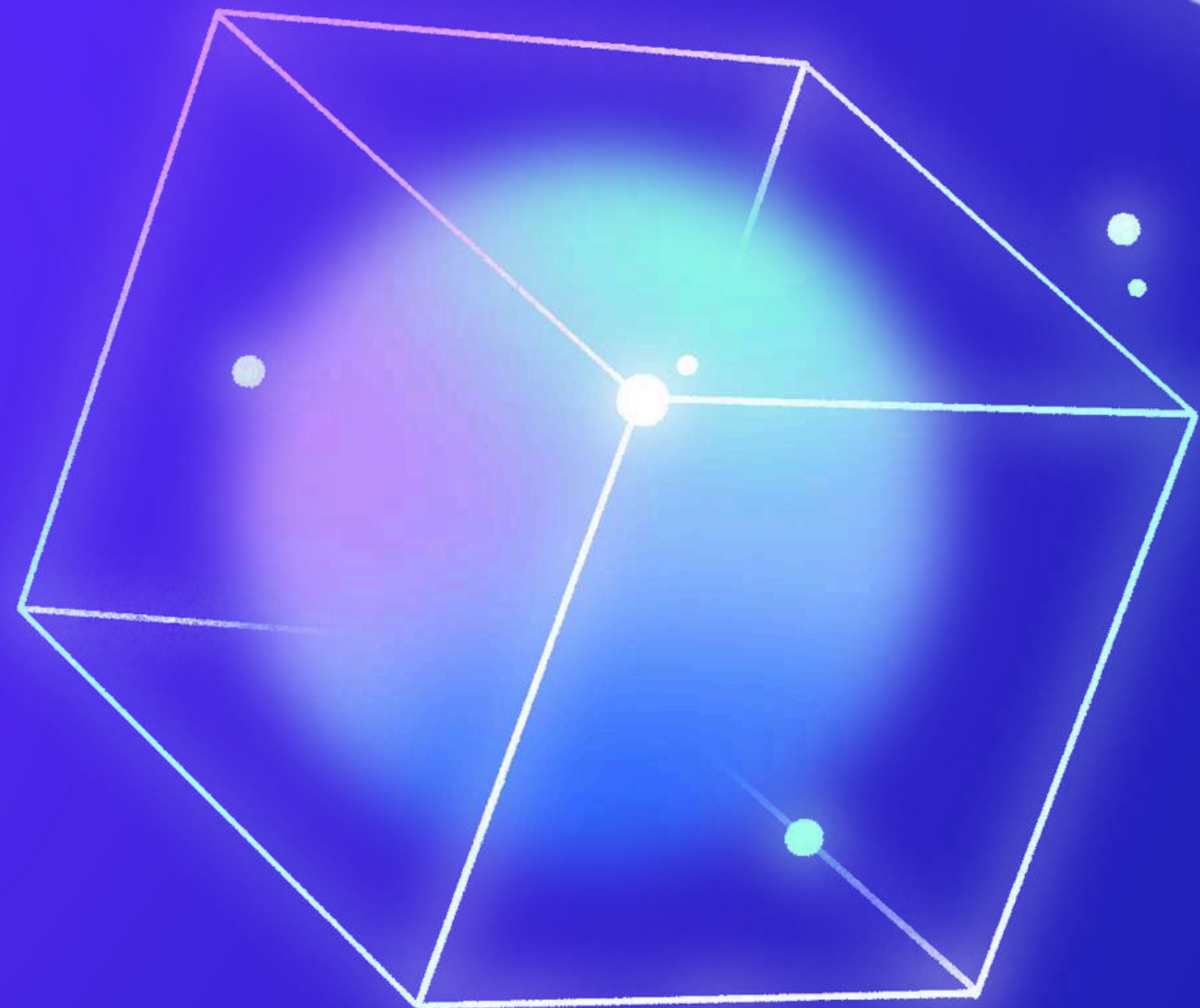
# SOMMAIRE

1. Entraînement de modèles dans un environnement MLOps
  - Méthodologie
  - Travail en environnement MLOps
  - Résultats
2. Mise en production d'un modèle sur mesure avancé
  - Développement API
  - CI/CD
  - Démonstration (prédiction, performances)
3. Règles RGPD
4. Conclusion et Limites





# ENTRAINEMENT DE MODÈLES



# MÉTHODOLOGIE

## JEU DE DONNÉES, KPI, MODÈLES

- BDD : Sentiment140 (1.6M de tweets)
- Entraînement sur Google Collab (GPU), échantillon 100 000-200 000 (downsampling)
- **Objectif** : Éviter le maximum possible un bad buzz → KPI privilégié : **precision** :  $TP/(TP+FP) \neq$  accuracy :  $(TP+TN)/total$

Approche	Vectorisation / Embedding	Modèle	Forces	Faiblesses
Classique	BoW ou TF-IDF	Naive Bayes	Simple, rapide, interprétable	Ignore le contexte, peu flexible
		SVM	Bon séparateur, robuste	Sensible aux réglages, lent sur petits jeux
		Random Forest	Robuste, peu de surapprentissage	Moins lisible, lourd si gros
		Régression Logistique	Rapide, claire, efficace en linéaire	Limité aux relations simples
Avancée (Deep Learning)	Word2Vec ou GloVe (300d)	LSTM	Capte le sens, suit l'ordre des mots	Lent, exigeant, peu interprétable
Contextuelle (Transformers)	DistilBERT embeddings contextuels	DistilBERT	Très bon contexte, généralisable	Coûteux, difficile à déployer



# MÉTHODOLOGIE

## PRÉTRAITEMENT

encodage html	Just flew w/ @AirParadis &#128640; &amp; I&#8217;m in LOVE! Epic service, comfy seats, good prices at https://airparadis.com #bestflight #AirParadis
casse	Just flew w/ @AirParadis 🚀 & I'm in LOVE! Epic service, comfy seats, good prices at https://airparadis.com #bestflight #AirParadis
tokens <url>	just flew w/ @airparadis 🚀 & i'm in love! epic service, comfy seats, good prices at https://airparadis.com #bestflight #airparadis
token <MENTION>	just flew w/ @airparadis 🚀 & i'm in love! epic service, comfy seats, good prices at <URL> #bestflight #airparadis
hashtag split	just flew w/ <MENTION> 🚀 & i'm in love! epic service, comfy seats, good prices at <URL> #bestflight #airparadis
punctuation + special chars	just flew w/ <MENTION> 🚀 & i'm in love! epic service, comfy seats, good prices at <URL> # bestflight # airparadis
non-printable / control chars	just flew w <MENTION> im in love! epic service comfy seats good prices at <URL> # bestflight # airparadis
tokenization	just flew w <MENTION> im in love! epic service comfy seats good prices at <URL># bestflight # airparadis
lemmatization	['just', 'flew', 'w', '<', 'MENTION', '>', 'im', 'in', 'love', '!', 'epic', 'service', 'comfy', 'seats', 'good', 'prices', 'at', '<', 'URL', '>', '#', 'bestflight', '#', 'airparadis']
stop words (except important)	['just', 'flew', 'w', '<', 'MENTION', '>', 'im', 'in', 'love', '!', 'epic', 'service', 'comfy', 'seat', 'good', 'price', 'at', '<', 'URL', '>', '#', 'bestflight', '#', 'airparadis']
	['flew', 'w', '<', 'MENTION', '>', 'im', 'love', '!', 'epic', 'service', 'comfy', 'seat', 'good', 'price', '<', 'URL', '>', '#', 'bestflight', '#', 'airparadis']

# MÉTHODOLOGIE

## ENTRAÎNEMENT : MODÈLES SIMPLES

Pourquoi ?

- Établir une baseline
- Modèles à faible coût:
  - rapides à entraîner
  - rapides à utiliser
  - interprétabilité

Modèle	Hyperparamètre	Rôle
Régression Logistique	C	Inverse de la régularisation ( $\downarrow C = +$ régularisation)
	max_iter	Nb max d'itérations pour convergence
	solver	Algorithme d'optimisation
SVM Linéaire	C	Même rôle que pour la régression logistique
	max_iter	Limite d'itérations
	dual	formulation duale (utile si $n\_samples > n\_features$ )
Random Forest	n_estimators	Nb d'arbres
	max_depth	Profondeur max des arbres
	min_samples_split	Min d'échantillons pour un split
Naive Bayes	alpha	Lissage (évite proba nulles)

Pipeline : Downsampling, prétraitement, split, vectorisation, classification supervisée en *cross-validation* (*refit : precision*)



# MÉTHODOLOGIE

## ENTRAINEMENT : MODÈLES SIMPLES

Pourquoi ?

- Établir une baseline
- Modèles à faible coût:
  - rapides à entraîner
  - rapides à utiliser
  - interprétabilité

Modèle	Hyperparamètre	Rôle
Régression Logistique	C	Inverse de la régularisation ( $\downarrow C = +$ régularisation)
	max_iter	Nb max d'itérations pour convergence
	solver	Algorithme d'optimisation
SVM Linéaire	C	Même rôle que pour la régression logistique
	max_iter	Limite d'itérations
	dual	formulation duale (utile si $n\_samples > n\_features$ )
Random Forest	n_estimators	Nb d'arbres
	max_depth	Profondeur max des arbres
	min_samples_split	Min d'échantillons pour un split
Naive Bayes	alpha	Lissage (évite proba nulles)

Pipeline : Downsampling, prétraitement, split, vectorisation, classification supervisée en *cross-validation* (*refit : precision*)

# MÉTHODOLOGIE

## ENTRAÎNEMENT : MODÈLES AVANCÉS

Pourquoi ?

- Capture plus fine sémantique
- Meilleure prise en compte du contexte (bidirectionnel local)
- Modèles à coût modéré
- Meilleure précision attendue

Modèle	Composant	Mots-clés / Points clés
Embeddings	Embedding	Word2Vec/GloVe pré-entraînés, non ajustés
LSTM	SpatialDropout1D	Régularisation, vecteurs entiers mis à zéro
	LSTM couche 1	Bidirectionnel, contexte local, séquence complète
	LSTM couche 2	Bidirectionnel, résumé global, vecteur unique
	Dense intermédiaire	Couche dense, extraction, affinage caractéristiques
	Dropout	Régularisation, évite overfitting
	Dense sortie	Sortie binaire, probabilité sentiment
	Compilation	Loss binaire, optimiseur adaptatif, métriques métier
	Entraînement & callbacks	Early stopping, ajustement learning rate, sauvegarde meilleur modèle (precision)

Pipeline : downsampling, prétraitement, split (train/val/test), embeddings (Word2Vec/GloVe), classification deep learning (min val\_loss, max precision)



# MÉTHODOLOGIE

## ENTRAINEMENT : BERT

Pourquoi ?

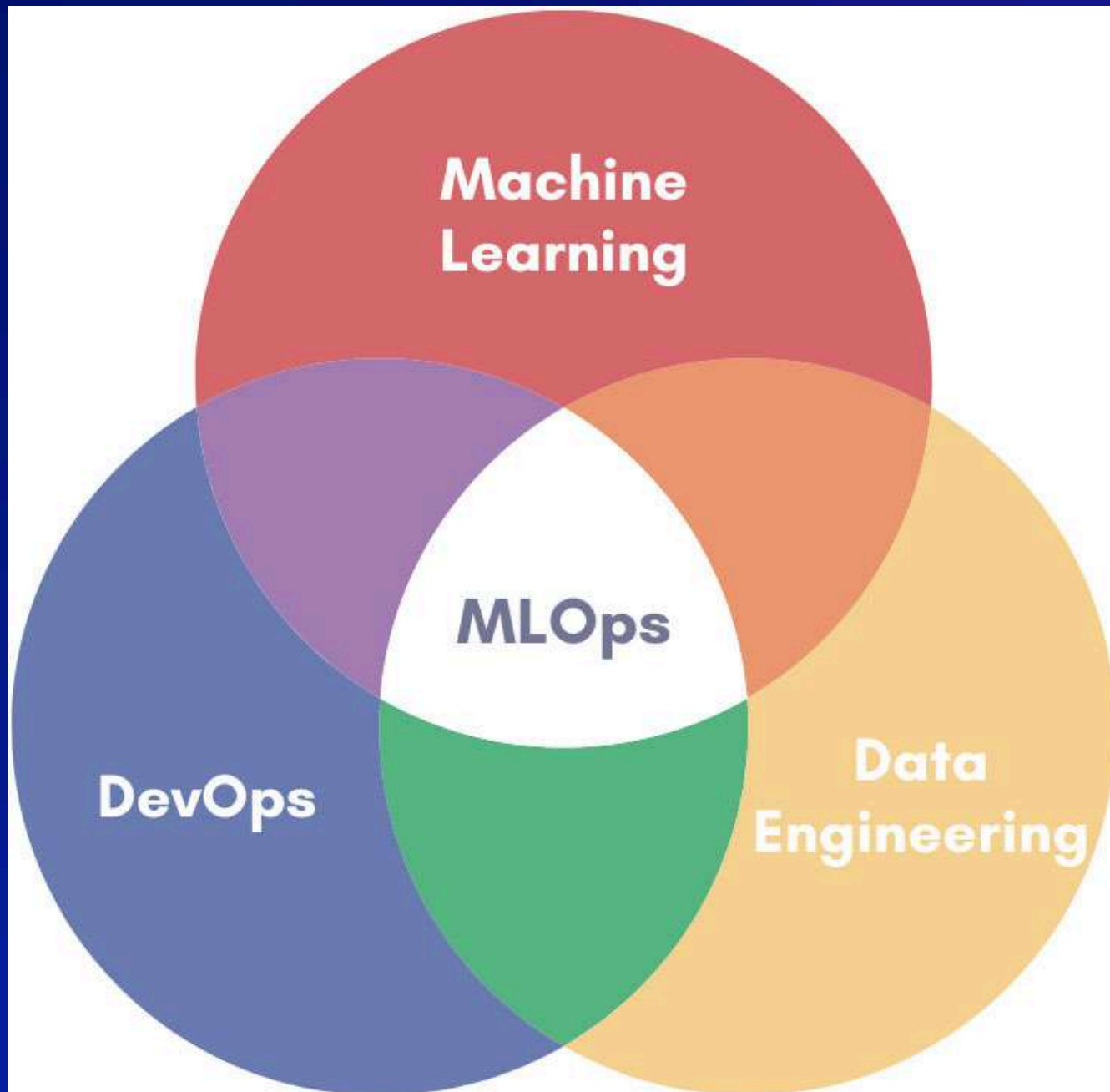
- Compréhension fine du contexte (bidirectionnel global)
- Exploite l'attention pour capturer nuances et dépendances
- Meilleure généralisation

Composant	Paramètre / Valeur	Rôle (mots-clés)
Modèle	distilbert-base-uncased	Modèle BERT allégé, pré-entraîné
Batch	8 (réel), 16 (effectif)	Limite mémoire, accumulation gradients
Prétraitement	max_length = 128	Troncature/padding des tweets
Optimisation	lr = 2e-5	Taux d'apprentissage initial
Scheduler	ReduceLROnPlateau	Baisse lr si val_loss stagne
	patience = 2, facteur = 0.2	Fréquence et intensité de la réduction
	min_lr = 1e-6	Taux plancher
EarlyStopping	patience = 7	Stoppe si val_loss stagne
	restore_best_weights=True	Recharge les meilleurs poids
Critère final	val_precision	Sauvegarde modèle avec meilleure précision

Pipeline : downsampling, prétraitement (léger), split (train, val, test), classification deep learning (DistilBERT), min val\_loss / max precision

# TRAVAIL EN ENVIRONNEMENT MLOPS

## PRINCIPES



Aspect	Machine Learning	Data Engineering	DevOps	MLOps
But principal	Construire et entraîner un modèle	Préparer et transformer les données	Automatiser et fiabiliser les déploiements	Automatiser tout le cycle ML (de la donnée à la production)
Activités clés	Modélisation, entraînement, évaluation	Collecte, nettoyage, feature engineering, pipelines	Gestion du code, CI/CD, monitoring, infrastructure	CI/CD ML, gestion des modèles, surveillance, reproductibilité
Outils courants	TensorFlow, PyTorch, scikit-learn, MLFlow	DVC, Apache Airflow, Kafka, Spark, MinIO	Docker, Kubernetes, Jenkins, Git, Prometheus	MLFlow, Kubeflow, Seldon Core, Airflow, Prometheus
Livrables	Modèle entraîné, métriques de performance	Données prêtes à l'emploi, pipelines automatisés	Code source, containers, scripts de déploiement	Modèles déployés, pipelines d'automatisation, monitoring et alertes
Compétences requises	Mathématiques, statistiques, programmation	Python/SQL, ingénierie des données, architecture cloud	DevOps, scripting, cloud, sécurité	Connaissances en ML, DevOps et data engineering
Cycle de vie	Expérimentation et optimisation	Automatisation des flux de données	Déploiement et maintenance	Orchestration complète du pipeline ML



# TRAVAIL EN ENVIRONNEMENT MLOPS

## UTILISATION DE MLFLOW

### Grands principes :

- Suivi des expériences (params, métriques, artefacts)
- Comparaison & reproductibilité des runs
- Stockage artefacts via buckets S3
- Packaging & déploiement modèles
- Interface web simple & efficace

### Déploiement:

- Docker sur NAS (OpenMediaVault)
- Backend artefacts : MinIO local (S3-like)
- Reverse proxy via SWAG (Nginx + SSL Let's Encrypt)
- Accès réseau local + public sécurisé
- Suivi & gestion centralisée des modèles

# RÉSULTATS

## UTILISATION DE MLFLOW

mlflow 3.0.0

Experiments

Models

Prompts

OC Projet 7

Demo PC

OC Projet 7

Provide Feedback

Add Description

Share

Runs

Models

Experimental

Evaluation

Traces

metrics.rmse < 1 and params.model = "tree"

Time created

State: Active

Datasets

Sort: test\_precision

Columns

Group by

+ New run

	Run Name	Created	Dataset	Duration	Source	Models	Metrics	Parameters
							test_precision	sample_size
	Model_Bert-distilbert-base-uncased	13 days ago	-	1.3min	colab_ke...	model	0.882345554...	100000
	Model_Bert-distilbert-base-uncased	4 days ago	-	49.3s	colab_ke...	model	0.861872273...	100000
	Modele_Avance_LSTM_Word2Vec-Fige	10 days ago	-	57.7s	colab_ke...	model_LSTM_Word2Vec-Fi...	0.797027927...	200000
	Modele_Avance_LSTM_GloVe-300d-Fige	10 days ago	-	46.6s	colab_ke...	model_LSTM_GloVe-300d-...	0.791522966...	200000
	Modele_Avance_LSTM_Word2Vec-Fige	12 days ago	-	2.0min	colab_ke...	model_LSTM_Word2Vec-Fi...	0.783031023...	100000
	Modele_Avance_LSTM_GloVe-300d-Fige	12 days ago	-	1.3min	colab_ke...	model_LSTM_GloVe-300d-...	0.780762393...	100000
	Modele_Simple_Random_Forest_BoW	12 days ago	-	4.9h	colab_ke...	model	0.779801409...	200000
	Modele_Simple_Regression_Logistique_TF-IDF	12 days ago	-	1.7min	colab_ke...	model	0.779591236...	200000
	Modele_Simple_Naive_Bayes_BoW	11 days ago	-	18.9s	colab_ke...	model	0.778522802...	200000
	Modele_Simple_SVM_Lineaire_TF-IDF	12 days ago	-	2.4min	colab_ke...	model	0.778086329...	200000
	Modele_Simple_Naive_Bayes_TF-IDF	12 days ago	-	17.7s	colab_ke...	model	0.777618069...	100000
	Modele_Simple_Naive_Bayes_TF-IDF	11 days ago	-	20.0s	colab_ke...	model	0.776643234...	200000
	Modele_Simple_Regression_Logistique_TF-IDF	12 days ago	-	56.8s	colab_ke...	model	0.776594705...	100000
	Modele_Simple_Naive_Bayes_BoW	12 days ago	-	15.0s	colab_ke...	model	0.776556776...	100000
	Modele_Simple_SVM_Lineaire_TF-IDF	12 days ago	-	1.1min	colab_ke...	model	0.776512940...	100000
	Modele_Simple_Random_Forest_BoW	12 days ago	-	1.8h	colab_ke...	model	0.774588908...	100000
	Modele_Simple_Regression_Logistique_BoW	12 days ago	-	15.3min	colab_ke...	model	0.772291997...	200000
	Modele_Simple_Random_Forest_TF-IDF	12 days ago	-	4.5h	colab_ke...	model	0.770202020...	200000
	Modele_Simple_SVM_Lineaire_BoW	12 days ago	-	8.0min	colab_ke...	model	0.768696148...	200000
	Modele_Simple_Regression_Logistique_BoW	12 days ago	-	6.5min	colab_ke...	model	0.767199311...	100000
	Modele_Simple_SVM_Lineaire_BoW	12 days ago	-	3.4min	colab_ke...	model	0.765525741...	100000
	Modele_Simple_Random_Forest_TF-IDF	12 days ago	-	1.7h	colab_ke...	model	0.763587497...	100000

22 matching runs

Show more columns (43 total)

Meilleur modèle (BERT)

Meilleurs modèles avancés

Meilleurs modèles simples



# RÉSULTATS

## UTILISATION DE MLFLOW : MEILLEUR MODÈLE AVANCÉ

mlflow3.0.0

ExperimentsModelsPrompts

OC Projet 7

Modele\_Avance\_LSTM\_Word2Vec-Fige

Overview

Model metrics

System metrics

Traces

Artifacts

Description

No description

Details

Created at	07/06/2025, 09:30:11 PM
Created by	root
Experiment ID	1
Status	Finished
Run ID	1de54538ee73443687db131fcc86d922
Duration	57.7s
Datasets used	—
Tags	Add tags
Source	colab_kernel_launcher.py
Registered models	SentimentAnalysisLSTM v12
Registered prompts	—

Metrics (8)

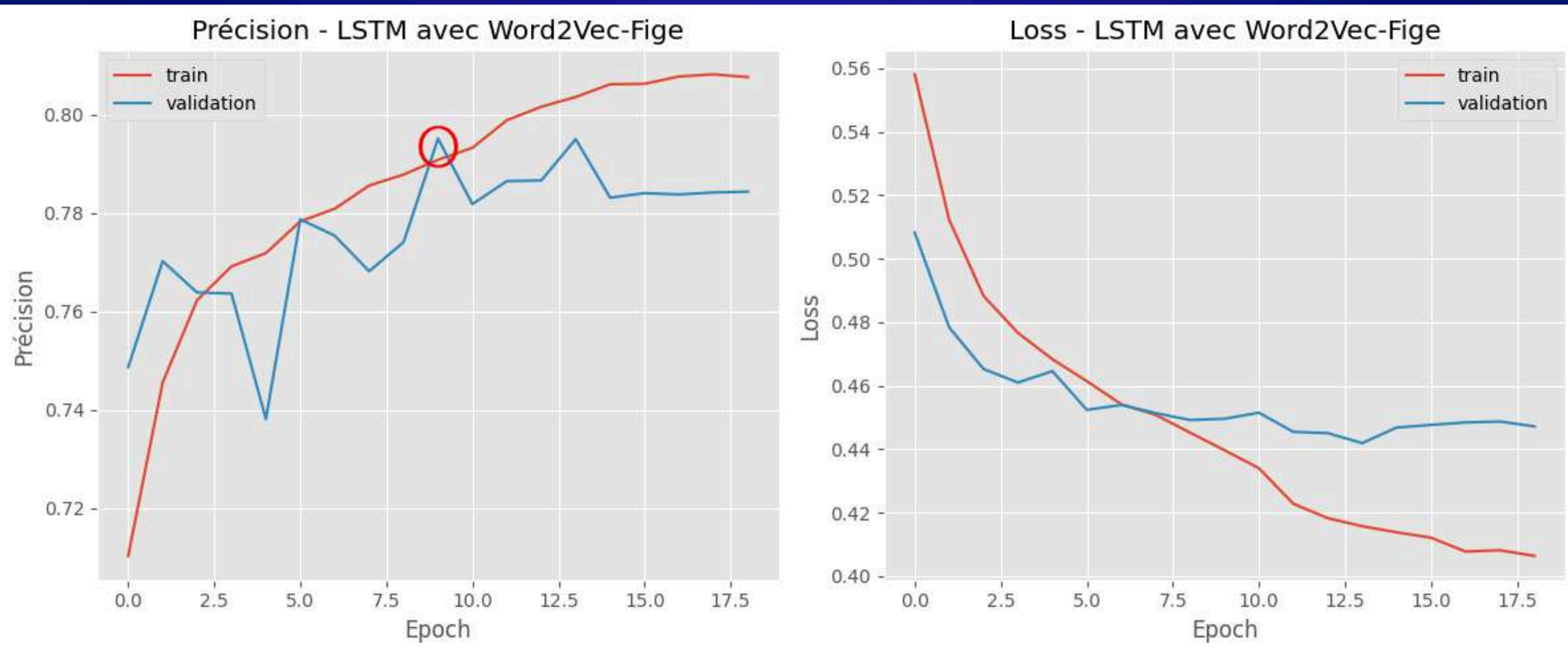
Metric	Value
val_precision	0.7951527833938599
val_loss	0.4496104419231415
val_recall	0.7727386355400085
test_accuracy	0.789825
test_precision	0.7970279272354599
test_recall	0.7777
test_f1	0.787245349867139
test_roc_auc	0.8748396600000001

Parameters (12)

Parameter	Value
model_type	LSTM
embedding_type	Word2Vec-Fige
embedding_dim	300
vocab_size	83731
max_sequence_length	50
trainable_embedding	False
max_epochs	50
real_epochs	19
batch_size	256
precision_policy	mixed_float16
gpu_optimization	XLA,dynamic_memory,prefetch
sample_size	200000

# RÉSULTATS

UTILISATION DE MLFLOW : MEILLEUR MODÈLE AVANCÉ







# MISE EN PRODUCTION



# DÉVELOPPEMENT API

## MLFLOW : MEILLEUR MODÈLE AVANCÉ

mlflow3.0.0

Experiments

Models

Prompts

🌙

GitHub

Docs

Registered Models

Share and manage machine learning models. [Learn more](#)

Filter registered models by name or tags

Name

Latest version

Staging

Production

Created by

Last modified

Tags

SentimentAnalysisLSTM

Version 12

Version 10

Version 12

07/06/2025, 09:38:2...

—

Create Model

mlflow3.0.0

Experiments

Models

Prompts

🌙

GitHub

Docs

Registered Models

SentimentAnalysisLSTM

Created Time: 07/03/2025, 03:22:52 PM

Last Modified: 07/06/2025, 09:38:26 PM

> Description

Edit

> Tags

▼ Versions

All

Active 3

Compare

New model registry UI

Version	Registered at	Created by	Stage	Description
✔ Version 12	07/06/2025, 09:38:25 PM		Production	
✔ Version 11	07/04/2025, 11:36:06 PM		Archived	
✔ Version 10	07/04/2025, 11:12:06 PM		Staging	
✔ Version 9	07/04/2025, 10:59:08 PM		Staging	
✔ Version 8	07/04/2025, 10:32:53 PM		Archived	
✔ Version 7	07/04/2025, 10:17:48 PM		Archived	
✔ Version 6	07/04/2025, 11:40:13 AM		Archived	
✔ Version 5	07/04/2025, 11:02:56 AM		Archived	
✔ Version 4	07/03/2025, 06:59:52 PM		Archived	

1



# DÉVELOPPEMENT API

## CONSTRUCTION DE SENTIMENT API

- FastAPI :
  - requête du modèle le plus récent (MLFlow) + téléchargement
  - gestion de plusieurs requêtes (prédiction)
  - journalisation des erreurs
  - envoi de rapports d'erreur (mail, 3 erreurs en moins de 5 minutes)
- Streamlit:
  - interface graphique
  - échange de requêtes vers FastAPI
- NGinx reverse proxy : gestion des 2 services sur un seul port

# CI/CD

## DESCRIPTION DE LA PIPELINE

- Dockerisation de l'application
- Github workflow CI/CD :
  - Validation de tests unitaires
  - Conteneurisation automatique et push vers dockerhub

The image shows two screenshots related to CI/CD. The top screenshot displays a GitHub Actions workflow named 'p7\_ci-cd.yml' triggered on 'push'. It consists of two steps: 'test' (1m 41s) and 'build\_and\_push' (3m 6s), both marked with green checkmarks. The bottom screenshot shows the Docker Hub page for the repository 'grgmdmn/sentiment\_api'. It includes the Docker Hub logo, navigation links, and details about the image, such as 'By grgmdmn · Updated about 10 hours ago' and 'OpenClassrooms Sentiment Prediction API'. It also shows a 'MACHINE LEARNING & AI' tag and statistics: 0 stars and 170 pulls.

p7\_ci-cd.yml  
on: push

test 1m 41s

build\_and\_push 3m 6s

dockerhub

Explore / grgmdmn / sentiment\_api

grgmdmn/sentiment\_api

By grgmdmn · Updated about 10 hours ago

OpenClassrooms Sentiment Prediction API

IMAGE

MACHINE LEARNING & AI

☆0 ↓170



# CI/CD

## DESCRIPTION DE LA PIPELINE

- Déploiement sur NAS:
  - Installation via Docker-Compose (configuration yaml)
  - Surveillance et mise à jour via Watchtower

RépondreRépondre à tousTransférerArchiverIndésirableSupprimerAutres

W

Watchtower

@etik.com

Pour @etik.com

08:10

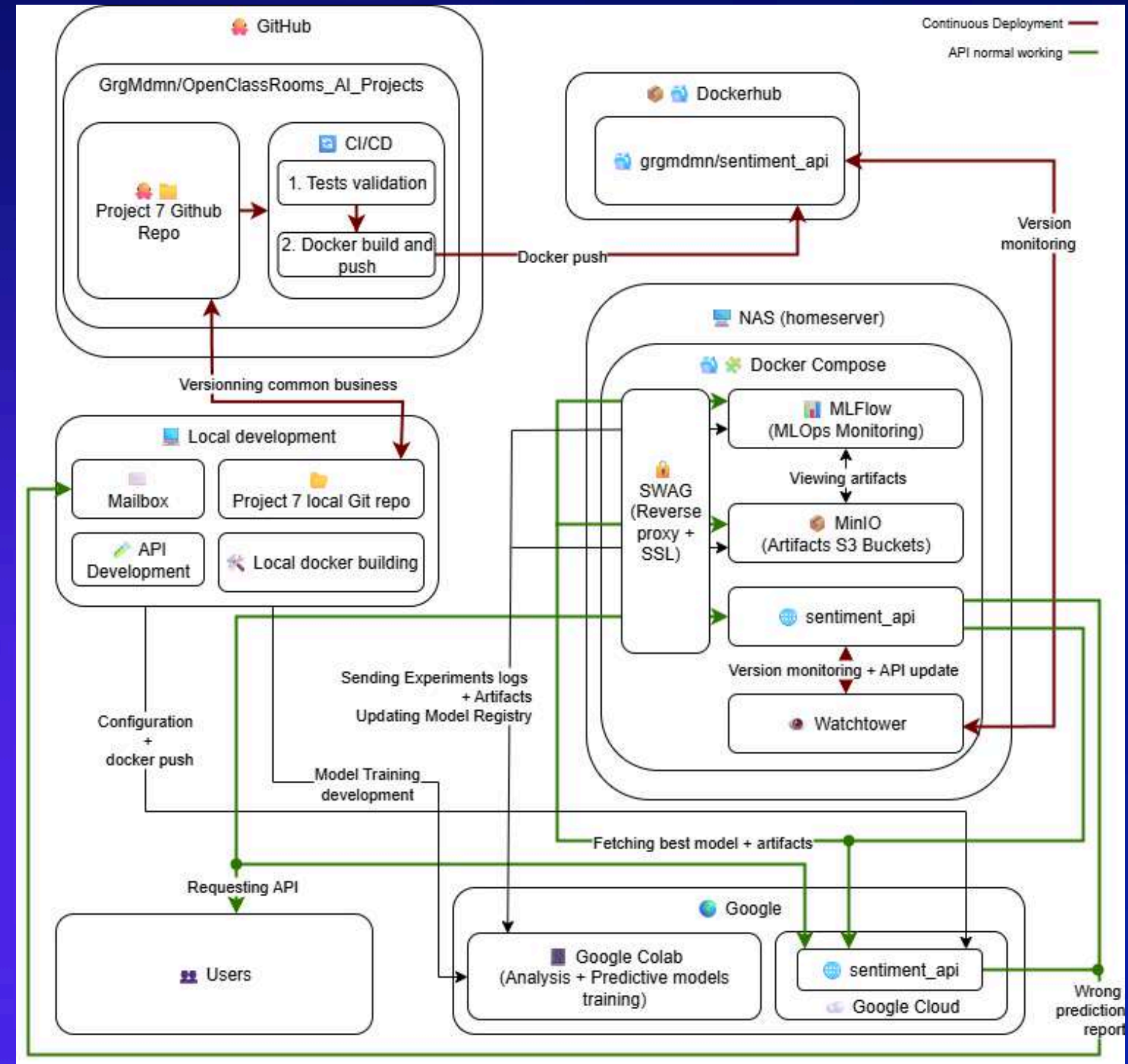
Watchtower updates on omv

Found new grgmdmn/sentiment\_api:latest image (326e0de7c9ad)  
Stopping /sentiment\_api (93144397debc) with SIGTERM  
Creating /sentiment\_api  
Removing image c4476bc52647

Services   Compose   Files		
Nom ^	Description ↕	Etat ↕
MinIO	Creating S3 buckets	Up
MLFlow	MLOps for Machine Learning (OpenClassRooms Project 7)	Up
Nextcloud	Cloud Drive App	Up
Portainer	docker containers catalog	Down
PostgreSQL	Relational Database	Up
Sentiment_API	Tweets Sentiment Prediction : uses Word Embedding + LSTM deep learning	Up
SWAG	Secure Web Application Gateway	Up
watchtower	Dockers Updates Monitoring	Up
1 sélectionné / 8 total		

# CI/CD

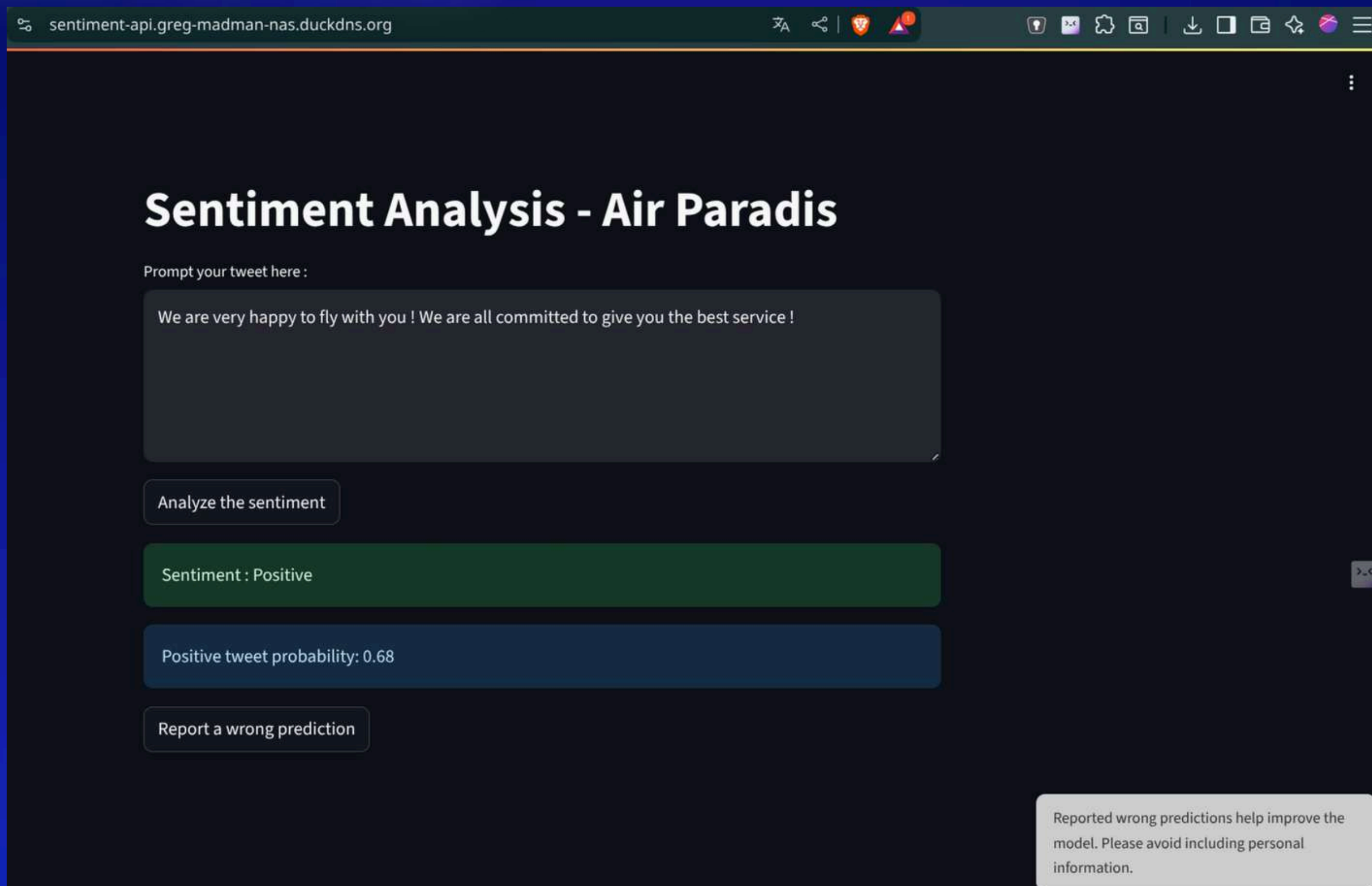
## RÉCAPITULATIF





# DÉMONSTRATION API

ARRIVÉE SUR LE SITE



The screenshot shows a web browser window with the URL `sentiment-api.greg-madman-nas.duckdns.org`. The page has a dark theme and features a title "Sentiment Analysis - Air Paradis". Below the title, there is a prompt "Prompt your tweet here :" followed by a text input area containing the text "We are very happy to fly with you ! We are all committed to give you the best service !". A button labeled "Analyze the sentiment" is positioned below the input. The results are displayed in two stacked boxes: a green box showing "Sentiment : Positive" and a blue box showing "Positive tweet probability: 0.68". At the bottom left, there is a button labeled "Report a wrong prediction". At the bottom right, a light gray box contains the text: "Reported wrong predictions help improve the model. Please avoid including personal information."

sentiment-api.greg-madman-nas.duckdns.org

## Sentiment Analysis - Air Paradis

Prompt your tweet here :

We are very happy to fly with you ! We are all committed to give you the best service !

Analyze the sentiment

Sentiment : Positive

Positive tweet probability: 0.68

Report a wrong prediction

Reported wrong predictions help improve the model. Please avoid including personal information.

# DÉMONSTRATION API

PRÉDICTION :

## Sentiment Analysis - Air Paradis

Prompt your tweet here :

We've changed... our loyalty program! Now you can enjoy a third flight at 50% off after your second booking! The perfect opportunity to treat yourself and enjoy a new trip for new adventures! ✈️

Analyze the sentiment

Sentiment : Positive

Positive tweet probability: 0.64

Report a wrong prediction



# DÉMONSTRATION API

ERREUR 1 :

## Sentiment Analysis - Air Paradis

Prompt your tweet here :

Don't forget to subscribe to the "enhanced life jacket" to ensure your safety! There's a 50% discount for the next two days! Happy Flight! 🚨🚢✈️

Analyze the sentiment

Sentiment : Positive

Positive tweet probability: 0.68

✅ Thanks for your report.

# DÉMONSTRATION API

ERREUR 2:

## Sentiment Analysis - Air Paradis

Prompt your tweet here :

We are very proud to announce that our flight and pricing policy has changed: from now on, you can subscribe to a “skip-the-line” toilets option ! 😁

Analyze the sentiment

Sentiment : Positive

Positive tweet probability: 0.67

✓ Thanks for your report.



# DÉMONSTRATION API

ERREUR 3 + RAPPORT D'ERREUR:

## Sentiment Analysis - Air Paradis

Prompt your tweet here :

I just learned that my important flight today was cancelled, but @Air Paradis gave me \$10 compensation so I could buy a Twix and go to the swimming pool: THANK YOU SO MUCH!!!!

Analyze the sentiment

Sentiment : Positive

Positive tweet probability: 0.53

✓ Thanks for your report.

✉ A complete report has been sent to the website administrator.

# DÉMONSTRATION API

## RÉCEPTION DE RAPPORT D'ERREUR

RépondreRépondre à tousTransférerArchiverIndésirableSupprimerAutres

Air Paradis Monitor

@etik.com

Pour @etik.com

15:28

Rapport d'erreurs - Prédications sentiments (11/07/2025 13:28)

Hello,

A new error report has been generated for the Air Paradis sentiment prediction API.

Total number of reports : 3  
Report date : 11/07/2025 à 13:28

Report summary :

- Tweet: "Don't forget to subscribe to the "enhanced life jacket" to ensure your safety! There's a 50% discoun..."  
Prediction reported as incorrect: POSITIVE (P = 0.68 )
- Tweet: "We are very proud to announce that our flight and pricing policy has changed: from now on, you can s..."  
Prediction reported as incorrect: POSITIVE (P = 0.67 )
- Tweet: "I just learned that my important flight today was cancelled, but @Air Paradis gave me \$10 compensati..."  
Prediction reported as incorrect: POSITIVE (P = 0.53 )

Recommended actions:

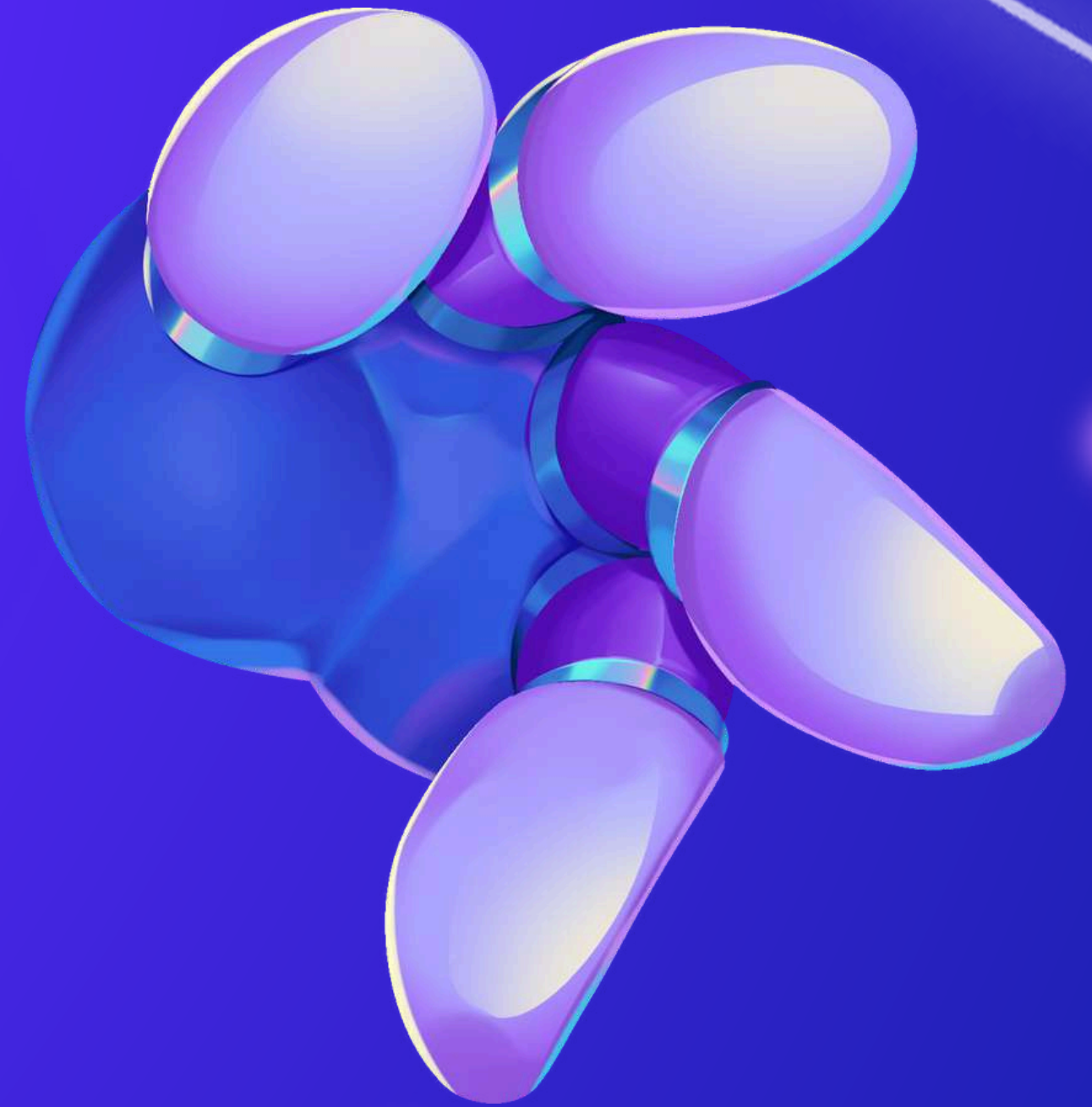
- Analyze reported tweets to identify patterns
- Consider re-training the model if necessary
- Check quality of training data

COMPLETE DATA (JSON) :

```
'''json
{
  "report_generated_on": "2025-07-11T13:28:21.752693",
  "reports_amount": 3,
  "reports": [
    {
      "id": 1,
      "tweet": "Don't forget to subscribe to the "enhanced life jacket" to ensure your safety! There's a 50% discount for the next two days! Happy Flight! 🛫👉👈",
      "incorrect_prediction": "positive (p = 0.68 ) ",
      "tweet_length": 143
    },
    {
      "id": 2,
      "tweet": "We are very proud to announce that our flight and pricing policy has changed: from now on, you can subscribe to a "skip-the-line" toilets option ! 🚽",
      "incorrect_prediction": "positive (p = 0.67 ) ",
      "tweet_length": 148
    },
    {
      "id": 3,
      "tweet": "I just learned that my important flight today was cancelled, but @Air Paradis gave me $10 compensation so I could buy a Twix and go to the swimming pool: THANK YOU SO MUCH!!!!",
      "incorrect_prediction": "positive (p = 0.53 ) ",
      "tweet_length": 175
    }
  ],
  "metadata": {
    "model": "SentimentAnalysisLSTM",
    "api_version": "1.0",
    "reporting_threshold": 3
  }
}
'''
```



RESPECT RGPD



# RESPECT RGPD

Contexte : déploiement sur NAS (serveur personnel / petit serveur d'entreprise)

Principe RGPD	Mesures mises en place
Licéité, loyauté, transparence	API publique, pas d'info perso collectée, usage clair
Limitation des finalités	Prédiction sentiment, rapport erreurs anonyme
Minimisation des données	Logs erreurs (anonymes), pas d'info utilisateur stockée
Exactitude des données	Traitement temps réel, retours utilisateurs en cas d'erreur
Sécurité et confidentialité	SSL (Let's Encrypt), NAS sécurisé, accès privé modèles

Reported wrong predictions help improve the model. Please avoid including personal information.



# CONCLUSION ET PERSPECTIVES



# CONCLUSION

- Entraînement de **modèles prédictifs** dans un environnement MLFlow assurant le suivi des performances et le déploiement des meilleurs modèles.
- Déploiement d'une **API publique** de prédiction de sentiments basée sur le meilleur modèle "Avancé" (Word2Vec + LSTM, sample\_size = 200 000, precision  $\simeq$  80%) avec suivi des performances
- Intégration et déploiement continu : commit + push → tests unitaires + construction docker → dockerhub → NAS (updates: watchtower)
- Travail général dans un environnement MLOps
- (Utilisation maximale de solutions souveraines, respect RGPD)



# PERSPECTIVES

- Prétraitement perfectible :
  - ~~Correction orthographique~~
  - ~~Casse~~ (perte d'infos émotionnelles)
  - **Emojis** (emoji.demojize())
  - ~~features numériques~~ (longueur, ponctuation, emojis, liens, hashtags)
- Modèles explorés :
  - GRU et TextCNN non testés
  - Entraînement précision pure
  - **DistilBERT (88% précision) non déployé**
- Divers :
  - Détection du sarcasme/ironie : limite intrinsèque, nécessite fine-tuning spécifique ou approches avancées
  - Suivi et monitoring : journalisation basique (emails erreurs), absence de base de logs et dashboard (SQL, Grafana)
  - Étude réponses à nos tweets
  - Dérive du modèle ?

THANK YOU!

