



# BLOC 5 : Concevoir et déployer des modèles d'apprentissage automatique

YOUREVIEW : l'analyse automatisée de vos commentaires

FURTADO LEAL CARLA

# Sommaire

## Analyse du besoin et résolution de problème

Analyse du besoin et du contexte  
Stratégie de résolution  
Technologie et outils

## Développement de modèles d'apprentissage automatique

Construction des variables  
Sélection des variables  
Entraînement des modèles

## Déploiement et automatisation des modèles

Méthode de sauvegarde  
Processus CI/CD  
Système de monitoring  
Système de collecte de données

Merci pour votre attention

Avez-vous des questions ?



# Analyse du besoin et résolution de problème

Analyse du besoin et du contexte

Stratégie de résolution

Technologie et outils



# MISE EN SITUATION

# MISE EN SITUATION



- Léna situation influenceuse française
- 3,14 millions d'abonnées sur You Tube
- 9<sup>ème</sup> saison des vlog d'aoûts.

Qu'ont-ils pensé de cette saison ?

# Besoins et enjeux

- Connaître les avis des ses abonnés sur ce concept
- Savoir ce qui plait ou non, à améliorer
- Découvrir des idées

→ Enjeux stratégiques et économique



# Problèmes

Volume important

Contrainte de temps

Biais cognitifs

Analyse qualitative  
complexe

Impact  
psychologique



# Problèmes



Comment valoriser les retours client  
via l'analyse automatisé des  
commentaires You Tube ?





# Les solutions existantes

**BRAND24**

**Youlyze**



# BRAND24

Outil d'écoute des médias sociaux  
Se base sur les mentions en ligne

Veille globale pas adaptée dans notre situation  
Déjà largement utilisé

Analyse de sentiments  
Détection de sentiments  
Analyse de sujets  
Utilisation de l'IA

**Youlyze**



**Youlyze**

Adapté à l'analyse des commentaires You Tube  
Adapté pour les influenceurs

Ne semble pas largement utilisé  
Pas de visibilité sur les performances  
Pas de mention d'affinité de langue

Analyse de sentiments  
Détection de sujets  
Identification des commentaires importants  
auxquels il faut répondre

**BRAND24**

**Youlyze**



Outil spécialisé dans l'analyse des commentaires  
You tube

Adapté pour différents profils  
Pas de mention d'affinité de langue

Analyse de sentiments  
Détection d'émotions  
Détection de sujets  
Analyse temporelle de l'engagement

**BRAND24**

**Youlyze**



Ma solution :  
**YOU REVIEW**

Ciblée et adaptée pour les influenceurs  
Spécialisée en langue française

- Analyse de sentiments
- Détections de sujets

# Contraintes et points de vigilance

Disponible pour septembre



Le vlogs d'août c'est **une vidéo tous les soirs pendant tout le mois**. Donc un **rythme soutenu** qui ne laisse **pas vraiment le temps pour l'analyse des retours**.

La parution quotidienne permettra de **tester l'outil en temps réel** pour l'équipe de développement.

**Répondre au besoin d'analyse post-diffusion, dès septembre.**

# Contraintes et points de vigilance

## Disponibilité des données

Les commentaires doivent être extraits de You tube, ici grâce à l'API.

Les **données d'entraînement** doivent être adaptées à l'analyse, elles doivent être **similaires à nos données réelles**.





# Contraintes et points de vigilance

Analyse automatisée

Avoir des modèles avec de **bonnes performances avec un minimum d'ajustements spécifiques**

L'automatisation implique de ne pas pouvoir personnaliser les traitements.



# Contraintes et points de vigilance

## Réglementation

Vigilance au niveau **des conditions d'utilisation de l'api de You tube et de la réglementation RGPD.**

**Les commentaires sont des données à caractères sensibles et personnelles.**

Les traitements sont donc soumis à certaines règles et doivent s'inscrire dans un **cadre légale adapté, ici l'intérêt légitime.**



# Contraintes et points de vigilance

## Coûts et délais



- **Version 1 local opérationnelle:**
    - 2-3 mois
    - Cout minimum car faible utilisation
  - **Version 2 de production :**
    - 2-3 mois
    - Cout à l'utilisation pour l'hébergement des données et la collecte de données via l'api YouTube
- Soit un coût d'infrastructure estimé à : 8-30\$/mois**

# Contraintes et points de vigilance

## Conclusion

Projet réalisable dans une version basique avec les fonctionnalités de base.

Des coûts faibles mais à revaloriser en fonction de l'utilisation.

Mais des délais courts pour une mise en production stable



# Stratégie de résolution



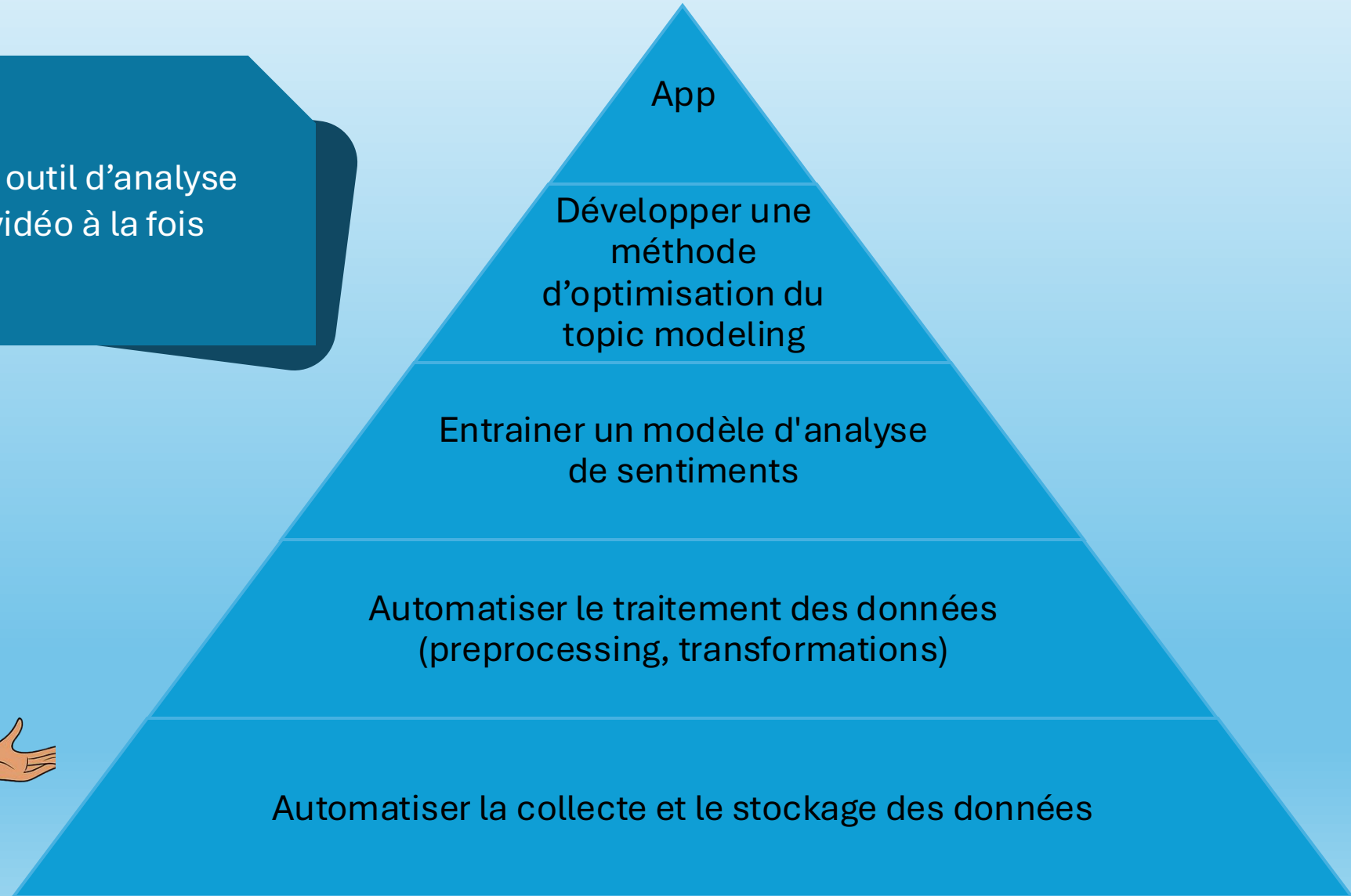
# Stratégie de résolution

Problèmes	Traduction
Est-ce que le contenu proposé plait ?	Analyse de sentiment
Qu'est ce qui a plut et déplu	Topic modeling ciblé sur les sentiments
Quels sont les sujets dominants ?	Topic modeling général



# Stratégie de résolution

Développer un outil d'analyse  
Qui traite une vidéo à la fois





# Technologies et outils

## Infrastructure

- MongoDB/Mongo atlas
- Streamlit : application et déploiement
- GitHub : CI/CD, monitoring et versionning

## Traitement de texte

- Spacy : rapidité et adapté au français
- Gensim : librairie spécialisée

## Analyse de texte

- Hugging face
- Sklearn

→ Accessibilité aux modèles sans les télécharger

## Coûts et compatibilité

Gratuits : dans la version actuelle  
Compatibilité : environnement python compatible pour toutes les tâches et disponible sur GitHub





# Développement de modèles d'apprentissage automatique

Construction des variables

Sélection des variables

Entraînement des modèles

# Données pour l'apprentissage & construction de variables

Jeux de données

Source : Kaggle

Origine : API You Tube v3

Langue : Internationale

Transformations :

- Traduction
- Preprocessing
- Vectorisation



# Données pour l'apprentissage & construction de variables

Jeux de données

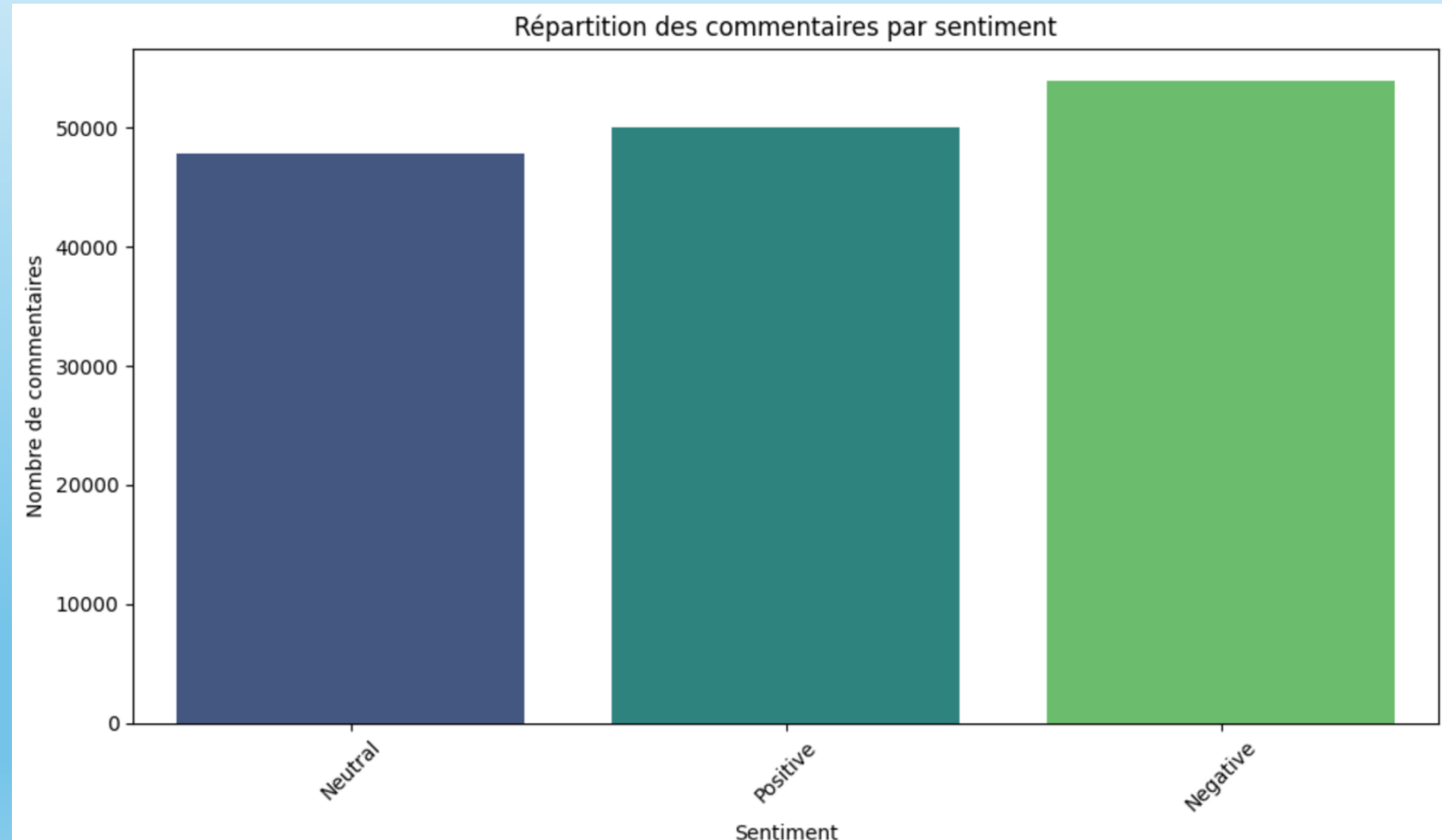
Source : Kaggle

Origine : API You Tube v3

Langue : Internationale

Transformations :

- Traduction
- Preprocessing
- Vectorisation



# Données pour l'apprentissage & construction de variables



	Variables dépendantes	Variables dépendantes	Variable prédictive	Variables dépendantes		
Méta-données	CommentText	tokens_clean_lem	Sentiment	Text_fr	w2vec_vector	Tfidf_vector
Identifiants vidéo, chaîne, auteur, titre, likes, réponses, date ...	Commentaires original rédigés en anglais		Negative Neutral Positive	Commentaire traduit		

Extrait de 151 884 sur un total de 1 032 225 commentaires

# Des méthodes de sélection de variables

Retrait des stop-words et  
ponctuations

: je, tu, un, des.... - : , ; / ? ! ...

Écrire 2 pages → écrire page

Retrait des chiffres et  
des accents

Nettoyage des liens

Regex sur les « http »

Avait, avaient, ont → avoir

Lemmatisation

Retrait des répétitions

Mmmmmdr → mdr ; loooooool → lol

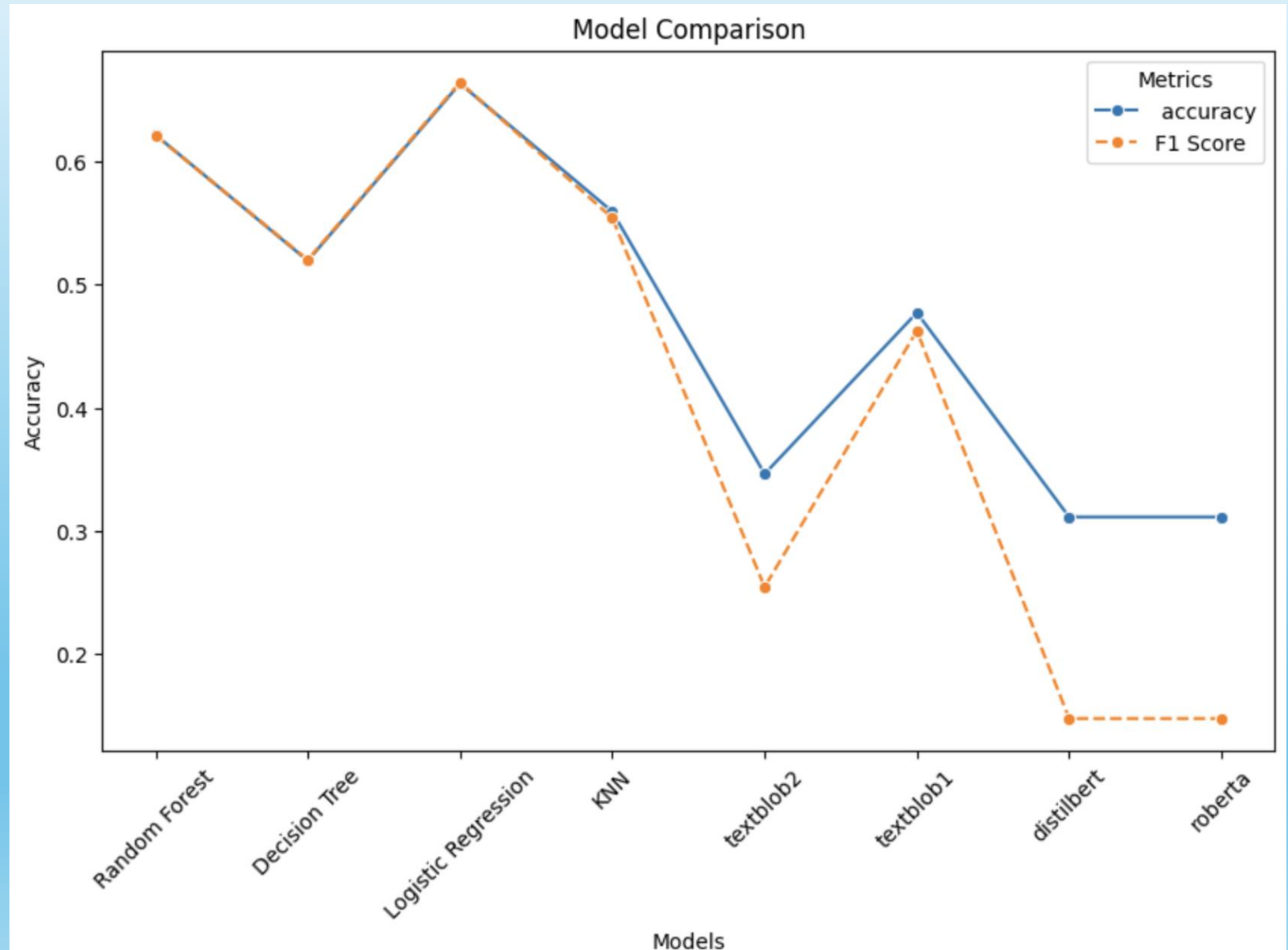
Mdr → mort de rire

Traitement des  
expressions

# Entrainement du modèle d'analyse de sentiment

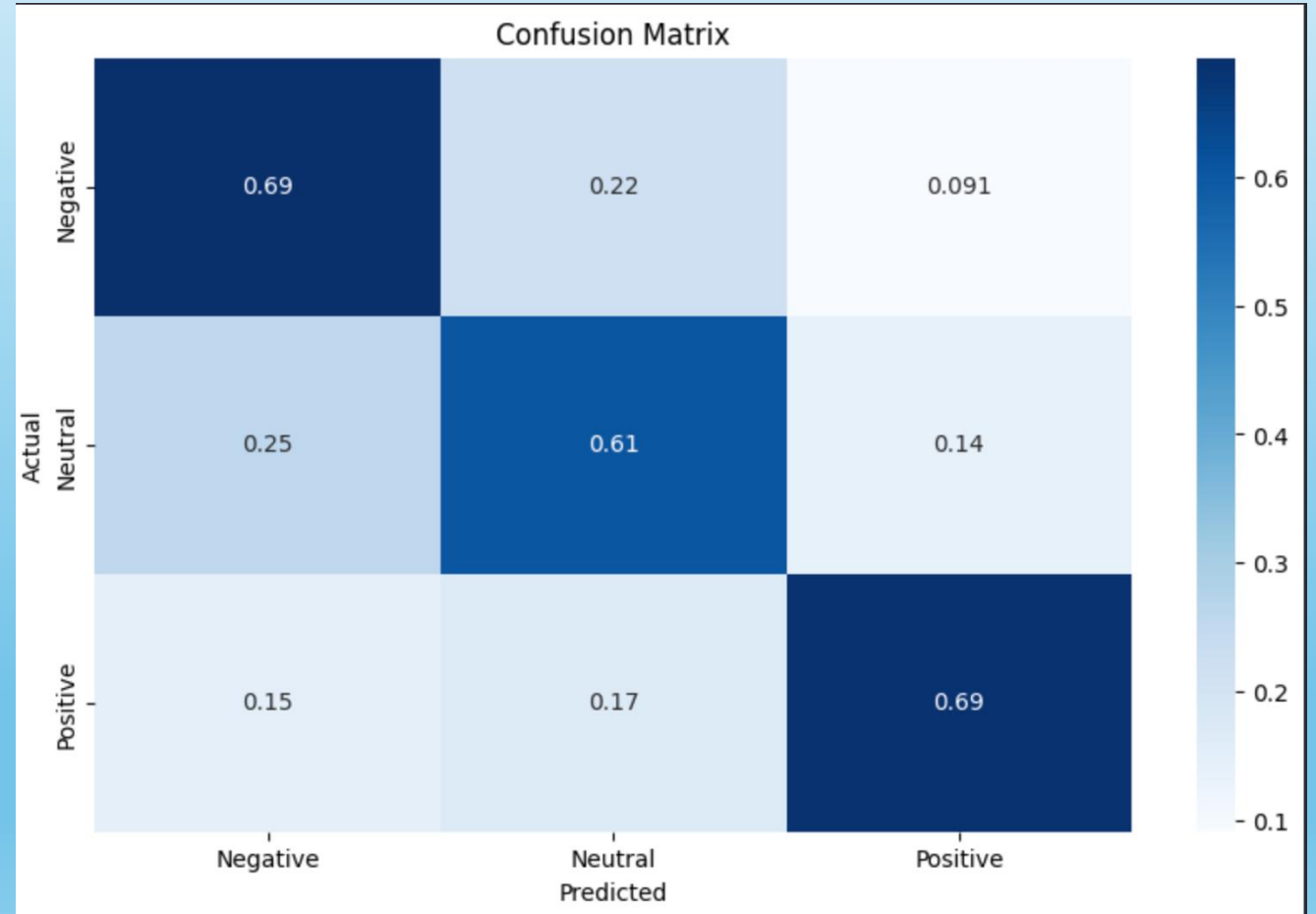
Technologies :

Sklearn  
Transformers



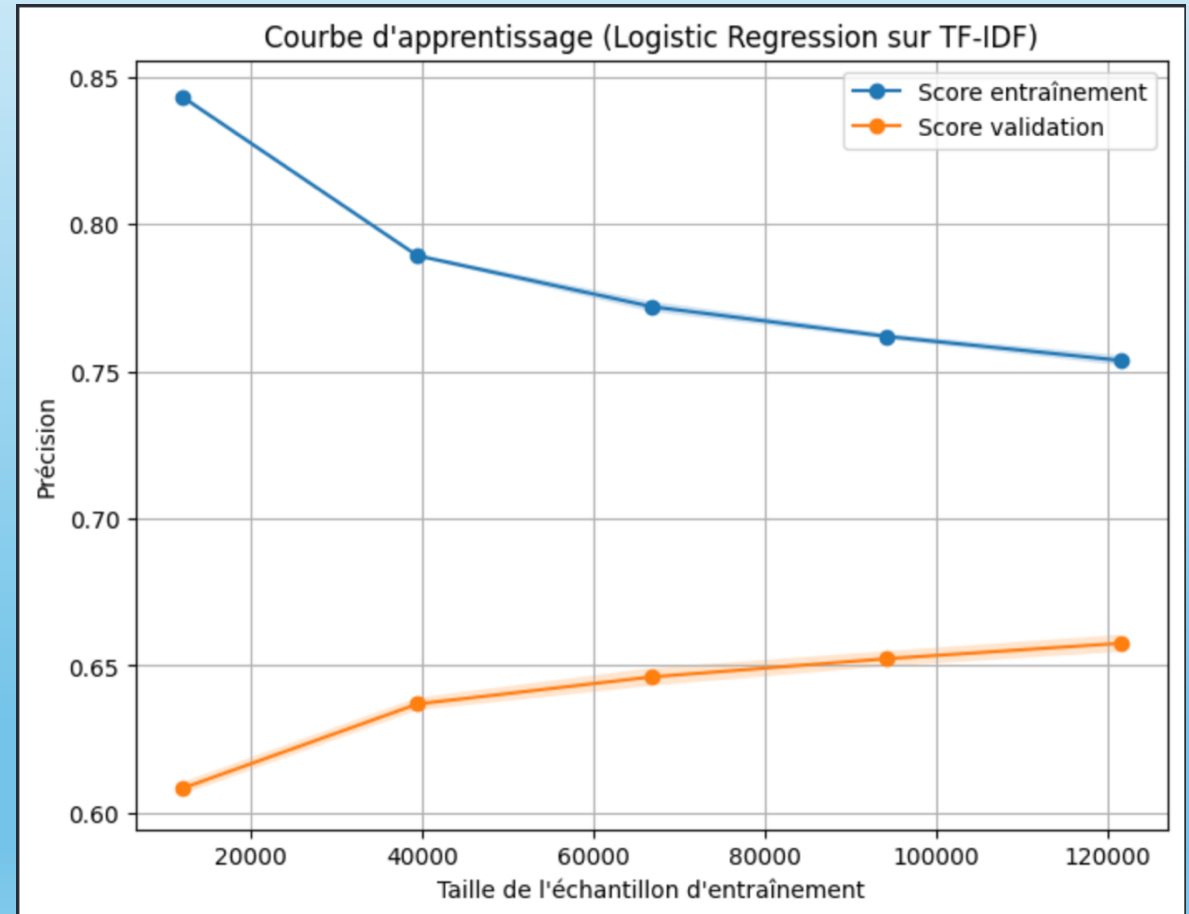
# Entraînement du modèle d'analyse de sentiment

Accuracy à 67%  
Plus d'erreurs lorsqu'il prédit les  
commentaires négatifs  
Très peu de commentaires positifs  
classés comme neutre



# Entraînement du modèle d'analyse de sentiment

Pas de problème d'overfitting ou d'underfitting





# Une méthode d'optimisation des modèles

Faire du finetuning sur  
un texte labellisé à la  
main

Chronophage  
Possible sur le long terme

Recherche des  
meilleurs paramètres

Couteux en temps  
De performance pas  
significativement améliorées

Faire du finetuning

Couteux en temps  
Amélioration significative des  
performances



# Une méthode d'optimisation des modèles

Faire du finetuning

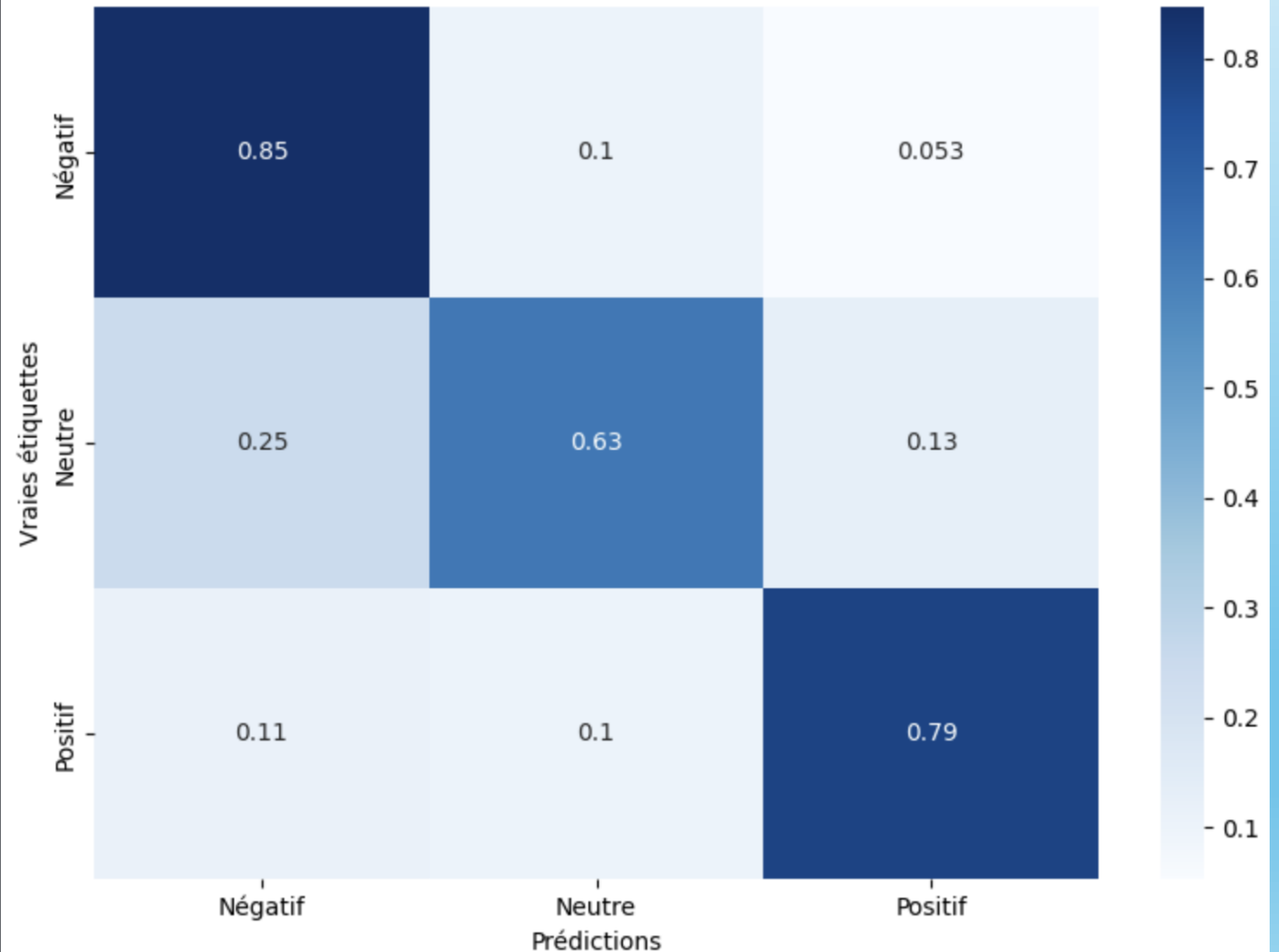
Couteux en temps

Amélioration significative  
des performances

- Accuracy : 75%



Matrice de Confusion - Modèle Fine-tuné



# Topic modeling

Objectif

Trouver le **nombre optimal** de topic

**Latent Dirichlet Allocation (LDA)**

Bibliothèque spécialisée dans le traitement de texte avec des métriques d'évaluation intégrées

Model



Méthode

Calcul de métrique

**Cohérence** : mesure la cohérence sémantique dans un topic

→ Entre 0.4 et 0.7

# Déploiement et automatisation des modèles

Méthode de sauvegarde

Processus CI/CD

Système de monitoring

Système de collecte de données

# Une méthode de sauvegarde



**Hugging Face**

Models 2

↑↓ Sort: Recently updated

🦊 Carlito-25/sentiment-model-logistic

Updated 30 minutes ago

🦊 Carlito-25/sentiment-model-finetuned

📄 0.3B • Updated 30 minutes ago

## Avantages

- Stockage de modèles volumineux (difficile sur GitHub)
  - Accès sécurisé grâce à des tokens
  - Utilisation simple via « pipeline »
    - Versionning « manuel »

# Un processus CI/CD



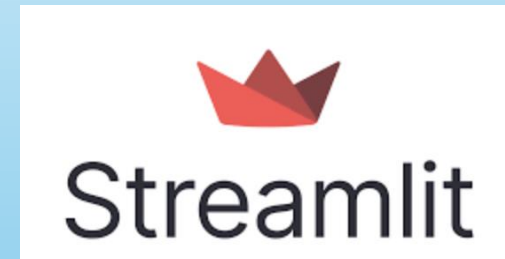
Intégration continue des modifications grâce à GitHub actions

- Déploiement et sauvegarde automatiques des modèles sur Hugging face
- Test de chargement des modèles avant utilisation
- Mise à jour automatique toutes les heures de la base de données
- Génération de rapport de performance des modèles
- Monitoring et alerting en cas d'échec des jobs



# Un processus CI/CD

Application accessible pour l'utilisateur grâce au déploiement sur Streamlit cloud qui se base sur le repository GitHub.

Accessible via un simple lien.



**carla-fl's apps**

 [projet-rncp](#) • [main](#) • [src/app/6\\_app.py](#) 

# Un système de monitoring de la performance

## Métriques de monitoring

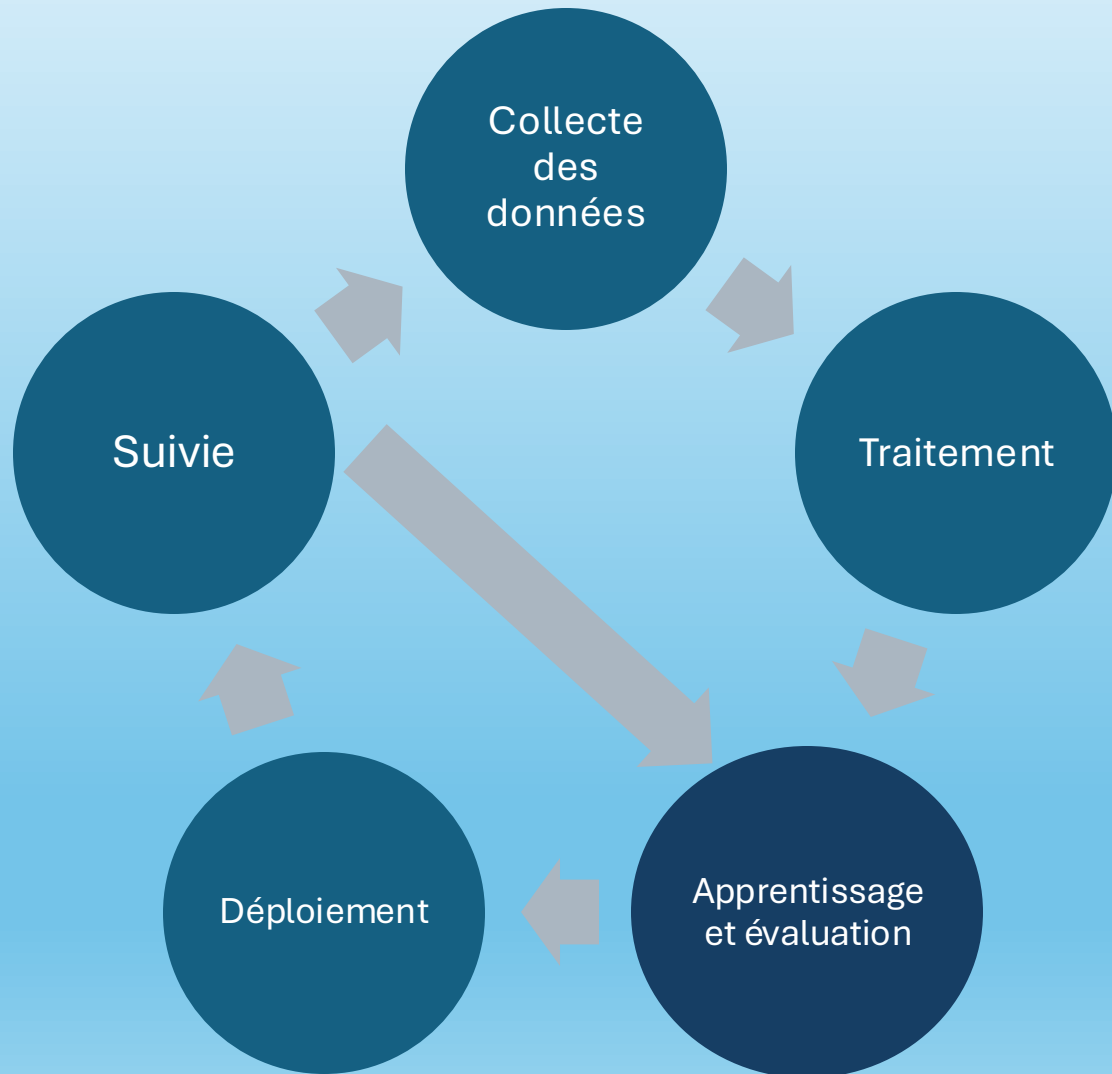
- Distribution de la taille des commentaires
- Distribution des labels
- Similarité du vocabulaire

## Outils

- GitHub actions avec exécution périodique d'un rapport de monitoring



# Un système de collecte de données



## Cycle du système d'apprentissage

- Processus ETL automatise la collecte et le traitement
- Processus CI/CD automatise le suivi et le déploiement des modèles
- Il n'y a apprentissage et évaluation qu'après analyse des résultats du suivi

The background of the slide features a crowd of stylized human figures, possibly made of felt or cardboard, holding hands in a circle. The figures are in various shades of grey and blue, set against a dark green background. The text is overlaid on this image.

# Merci pour votre attention

---

Avez-vous des questions ?