



الْمَعْهَدُ الْوَطَّانِيُّ لِلْبَرِيدِ وَالْمَوَاسِلَاتِ
Institut National des Postes et Télécommunications



agence nationale de réglementation
des télécommunications
الوَكَالَةُ الْوَطَّانِيَّةُ لِتَقْنِيَّنَ الْمَوَاسِلَاتِ
+٢١٥٦٠٣٩٥٤ | +٢١٥٦٠٣٩٥٥٤

Youtube Engagement and virality

analysis

Data Mining

PRESENTED BY

AABIL Rime

DIHAJI Youssed

EL MIZ Niama

MOHSEN Mohammed

SUPERVISOR

Mme. El Asri

MAJOR

Data Science

Problématique

En moyenne, plus de **500 heures de vidéos sont publiées chaque minute** sur YouTube, ce qui équivaut à environ **80 000 vidéos mises en ligne chaque seconde**.

Répondre aux questions suivantes :

- Qu'est-ce qui fait qu'une vidéo a plus de chances de devenir virale ?
- À quel moment publier pour toucher un maximum de personnes ?
- En quoi le titre influence-t-il la viralité d'une vidéo ?

Plan

01 Objectifs du projet

02 Extraction de données

03 Feature engineering

04 Analyse et insights
Perspectives:

- Modélisation
- Evaluation
- Déploiement

06 Ouverture

Objectifs du projet

Prédiction

Prédire la viralité potentielle d'une vidéo avant sa publication.

Analyse

Analyser l'influence des caractéristiques du titre, du moment de publication et du canal.

Recommendations

Fournir des recommandations actionnables aux créateurs

Technical stack

Data collection



YouTube Data API v3

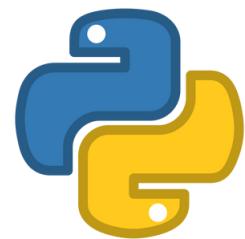
Feature engineering



Reproducibility



Data preprocessing



Visualisation



Extraction de données et Scrapping

Youtube Data API v3

Restrictions et quota:

- 10 000 units / jour / projet (défaut)



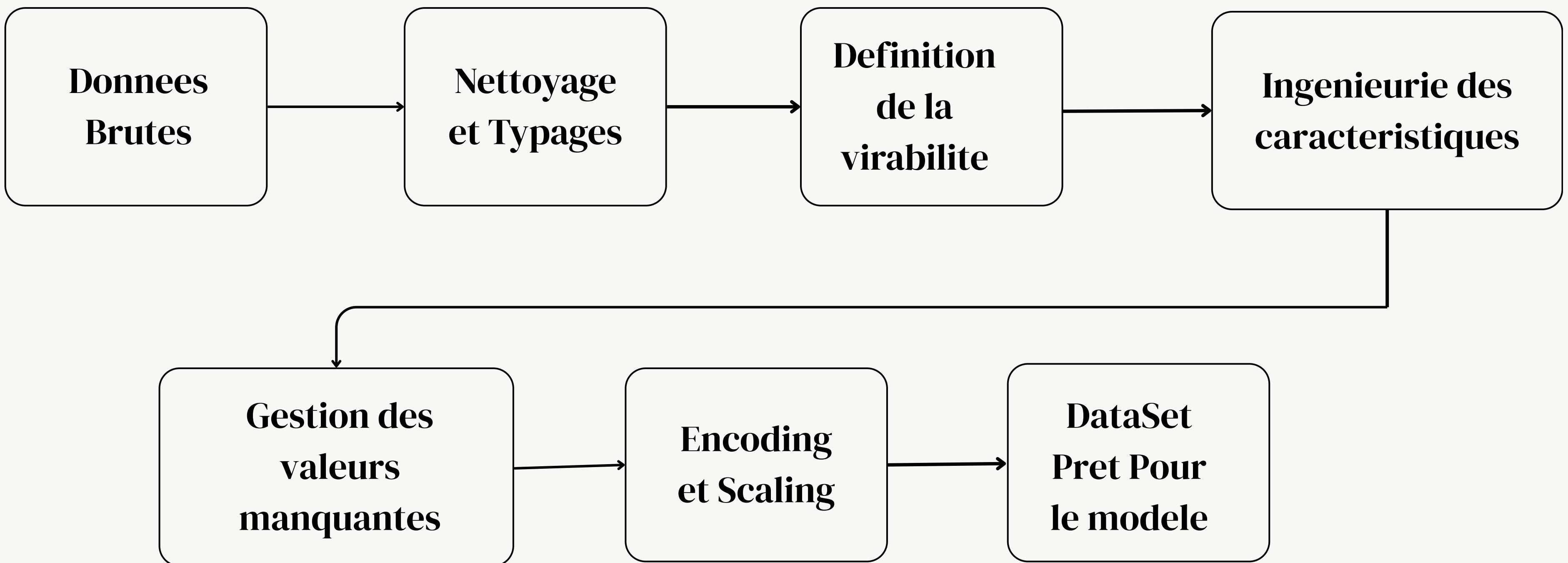
Caractéristiques importantes

- Recherche de vidéos, chînes, playlists....
- Metadata (titre, description, vues, likes, commentaires...)

Données collectées

	Eléments	Métadonnées
D A T A	Vidéo	<ul style="list-style-type: none">• video_id• title• published_at <ul style="list-style-type: none">• comment_count• like_count• duration ...
	Chaîne	<ul style="list-style-type: none">• view_count• subscriber_count• video_count

Pipeline EDA



Nettoyage des données

Fondation des Données:

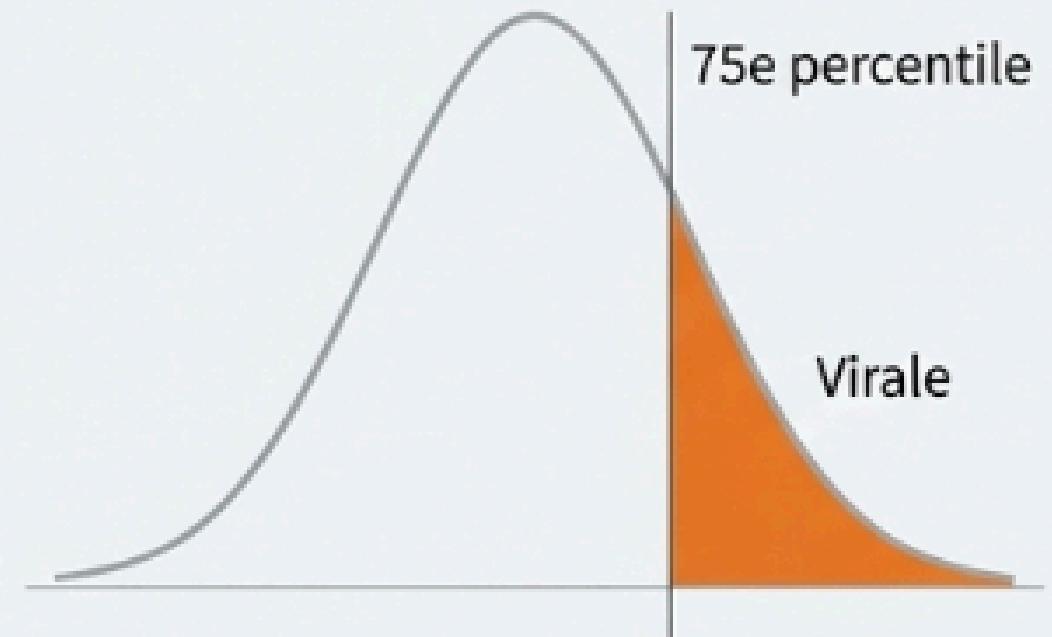
Avant toute Analyse, deux actions fondamentales sont menées pour assurer la qualité et la pertinence de nos données.

1. Nettoyage et Typage (fix_data_types)

- Conversion systématique des colonnes numériques (`view_count`, `like_count`, etc.) et temporelles (published_at) vers les formats corrects.
- Gestion des valeurs manquantes : les valeurs nulles dans les colonnes de comptage sont remplacées par la médiane pour préserver la distribution.

2. Définition de la Cible : Qu'est-ce qu'une Vidéo "Virale" ? (calculate_virality_score)

- Nous définissons la viralité de manière relative au jeu de données pour garantir une analyse pertinente quel que soit le contexte.
- Une vidéo est classée comme virale (`is_viral = 1`) si son nombre de vues (`view_count`) se situe dans le quartile supérieur (75e percentile) de l'échantillon.
- **Seuil calculé:** `threshold = df['view_count'].quantile(0.75)`



Comment transformer les titres en signaux Prédictifs ?

Caractéristiques Extraites :

Quantitatives :

- title_length : Nombre total de caractères.
- title_word_count : Nombre de mots.

Stylistiques :

- has_exclamation / has_question : Présence de "!" ou "?".
- has_emoji / emoji_count : Présence et nombre d'emojis.
- caps_word_count : Nombre de mots entièrement en majuscules.

Contenu :

- has_numbers : Présence de chiffres.
- special_char_count : Nombre de caractères spéciaux.
- clickbait_score : Score basé sur la présence de mots comme "secret", "incroyable", etc.



Le ‘QUAND’ de la publication

Le moment de la publication peut influencer la performance d'une video. Nous avons extraire 13 Caracteristiques remporelle .

Caractéristiques Extraites :

Basiques :

- publish_hour
- publish_dayofweek
- publish_month

Contextuelles :

- is_weekend : La publication a-t-elle eu lieu un samedi ou un dimanche ?
- is_peak_hour : La publication a-t-elle eu lieu durant les heures de pointe (17h-21h) ?
- days_since_publication : Ancienneté de la vidéo.

Feature Engineering

Features Engineering

Fusioner des colonnes pour créer une métrique composite. (ex: Engagement Rate)

1. Données d'Origine (AVANT)

Nous partons de métriques isolées ('likes', 'commentaires', 'vues').

post_id	title	like_count	comment_count	view_count
101	"Incroyable découverte !"	1,200	45	30,000
102	"Mon avis sur le sujet"	50	2	1,500
103	"Le guide complet"	8,500	310	250,000
104	"Une question pour vous"	350	80	9,000

2. Logique de Création (PROCESSUS)

Nous les combinons avec une formule pour créer un indicateur normalisé et plus riche.

$$\left(\frac{\text{like_count} + \text{comment_count}}{\text{view_count}} \right) * 100$$

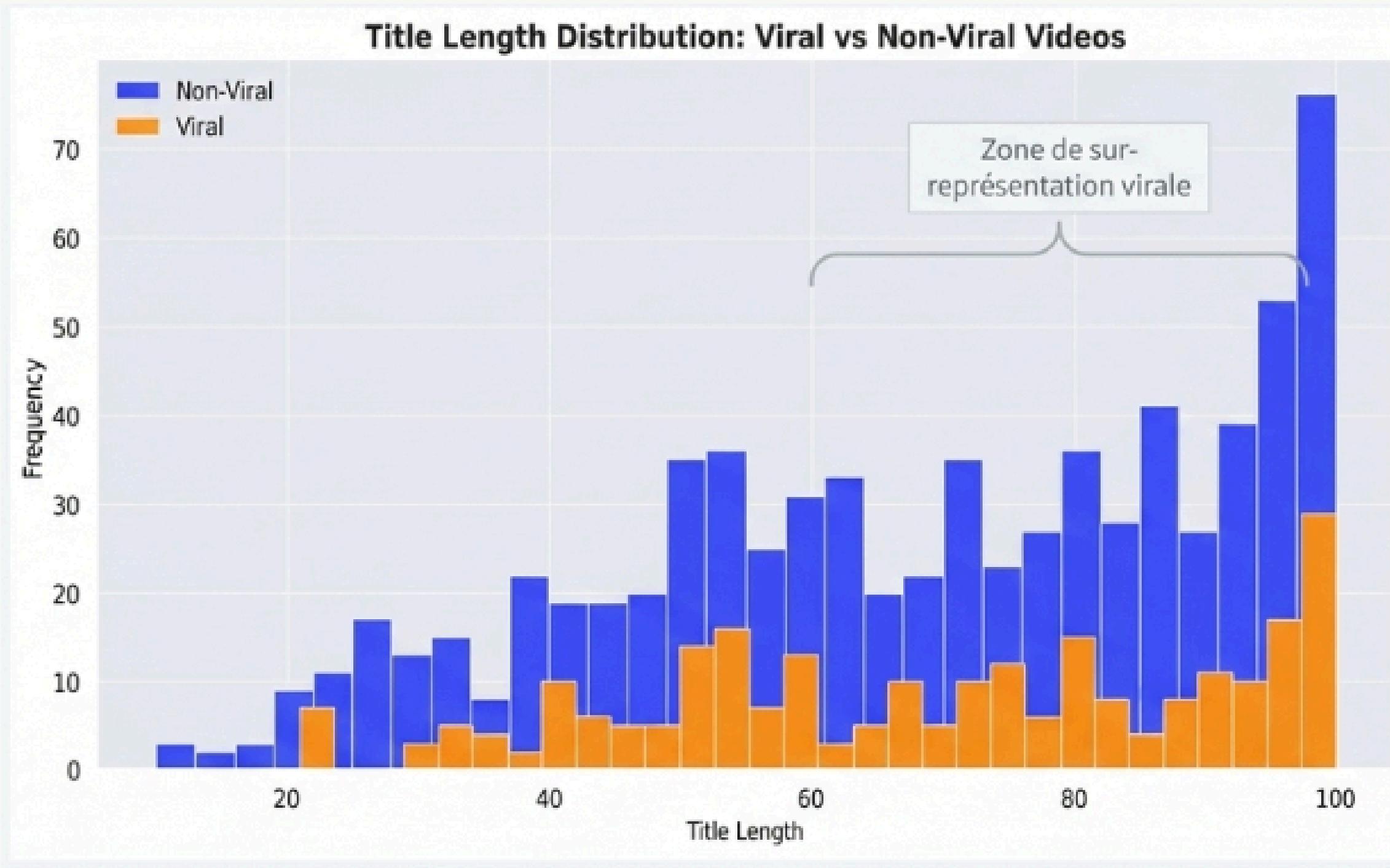
```
view_count_safe = df['view_count'].replace(0, 1)  
  
df['engagement_rate'] = (  
    (df['like_count'] + df['comment_count'])  
    / view_count_safe  
) * 100
```

3. Données Enrichies (APRÈS)

Le jeu de données final contient une nouvelle colonne qui capture une information de plus haut niveau.

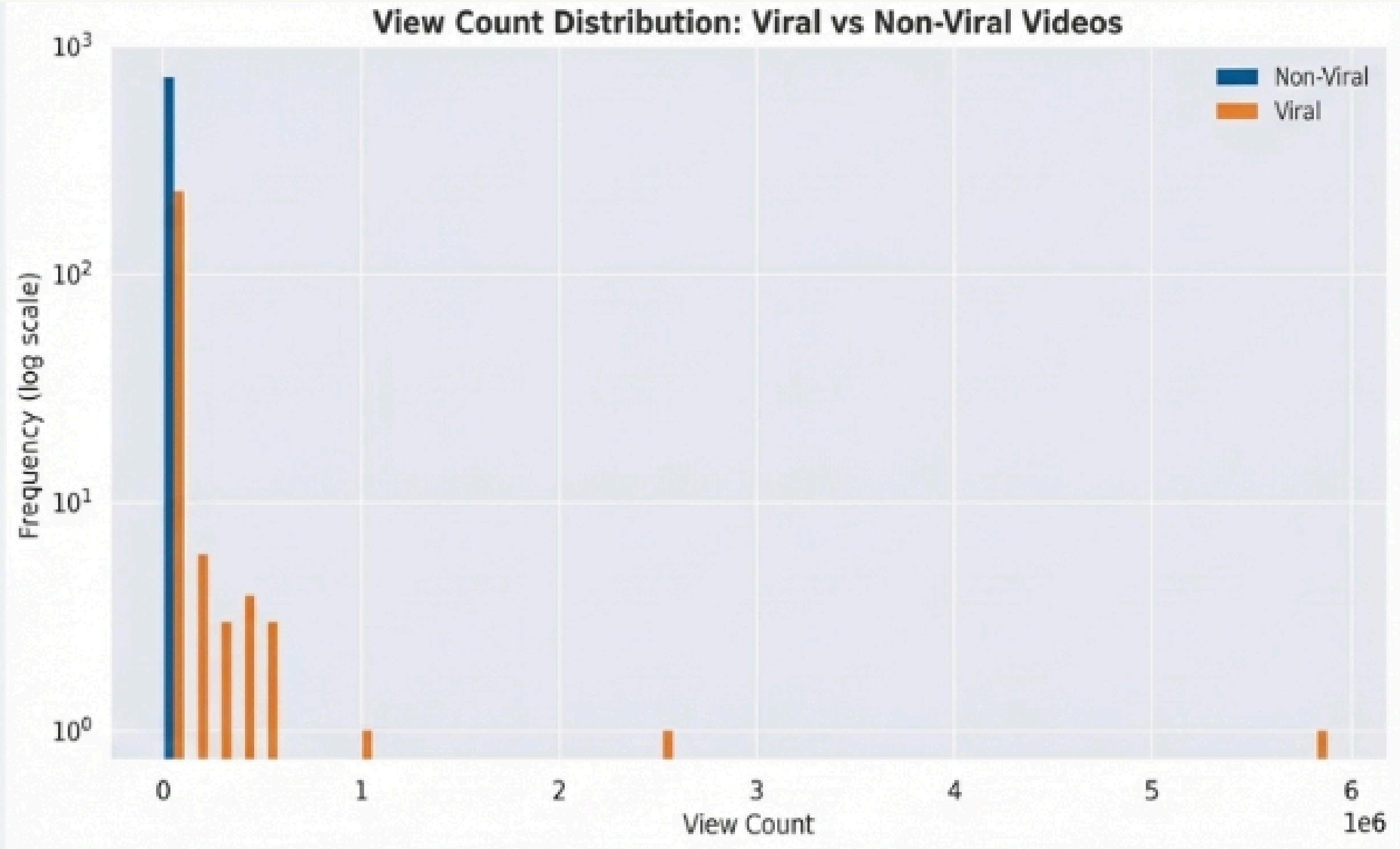
post_id	title	like_count	comment_count	view_count	engagement_rate
101	"Incroyable découverte !"	1,200	45	30,000	4.15
102	"Mon avis sur le sujet"	50	2	1,500	3.47
103	"Le guide complet"	8,500	310	250,000	3.52
104	"Une question pour vous"	350	80	9,000	4.78

La longueur du titre: Un levier subtile pour la Viralite



- Ce que nous voyons : Cet histogramme compare la distribution de la longueur des titres pour les vidéos virales (orange) et non virales (bleu).
- Ce que cela signifie : Il n'y a pas de 'longueur magique', mais on observe une tendance claire : les vidéos virales sont sous-représentées dans les titres très courts (< 40 caractères) et sur-représentées dans les titres plus longs (50-100 caractères).
- Hypothèse : Des titres plus descriptifs et informatifs, qui nécessitent plus de caractères, pourraient être plus efficaces pour attirer un public qualifié et stimuler le partage.

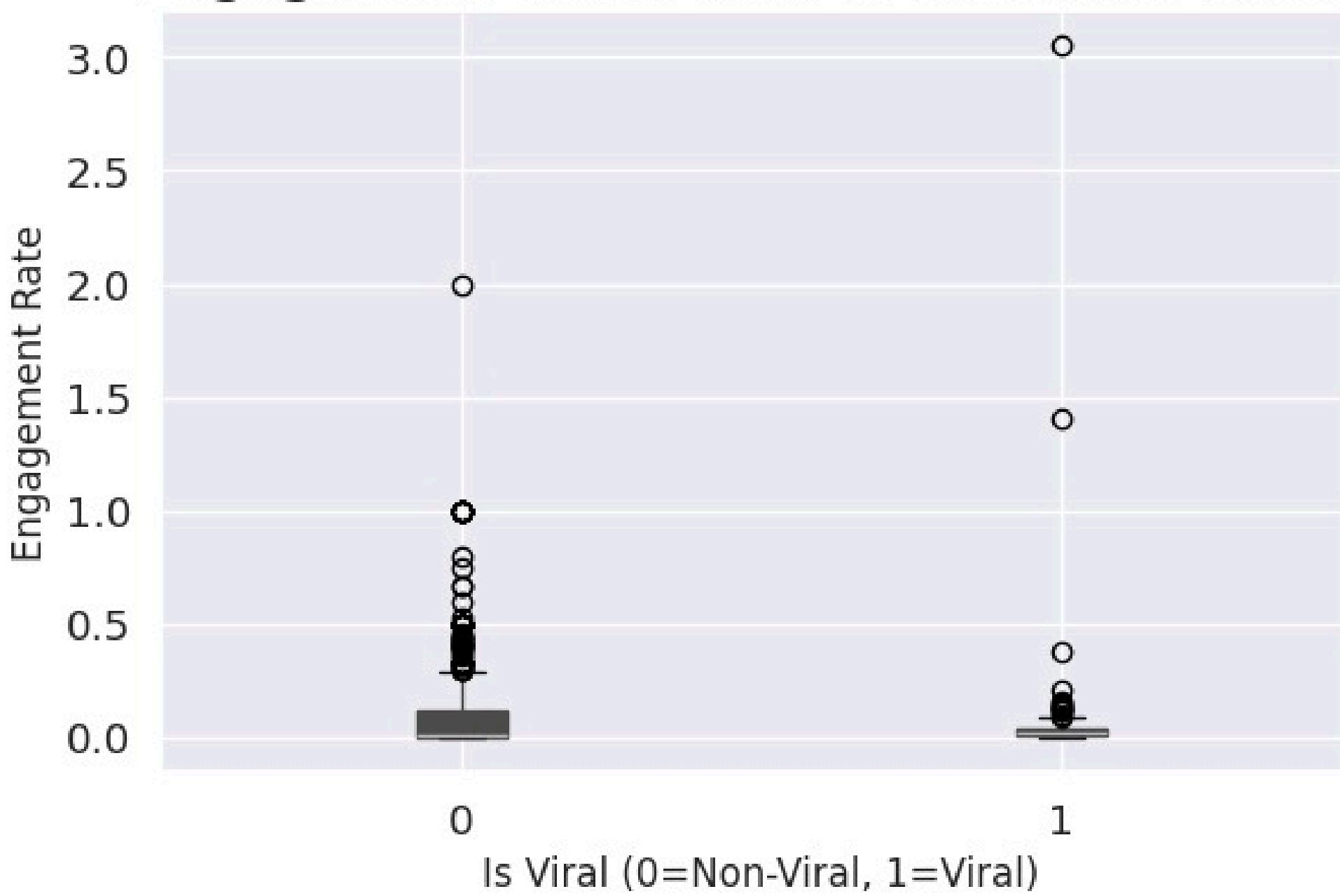
Distrubition des Vues



- Ce que nous voyons : La distribution du nombre de vues, avec une échelle logarithmique en ordonnée. Une écrasante majorité de vidéos (en bleu) cumule un faible nombre de vues.
- Ce que cela signifie : Les vidéos virales (en orange) sont des exceptions statistiques. Ce graphique illustre parfaitement pourquoi définir la viralité par un seuil relatif (le 75e percentile) est une approche plus robuste que de choisir un seuil absolu (ex: 1 million de vues), car cela s'adapte à la performance globale du dataset.

Le PARADOX du taux d'engagement

Engagement Rate: Viral vs Non-Viral Videos

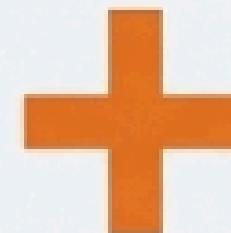


(Likes + Comments)
/ VIEWS

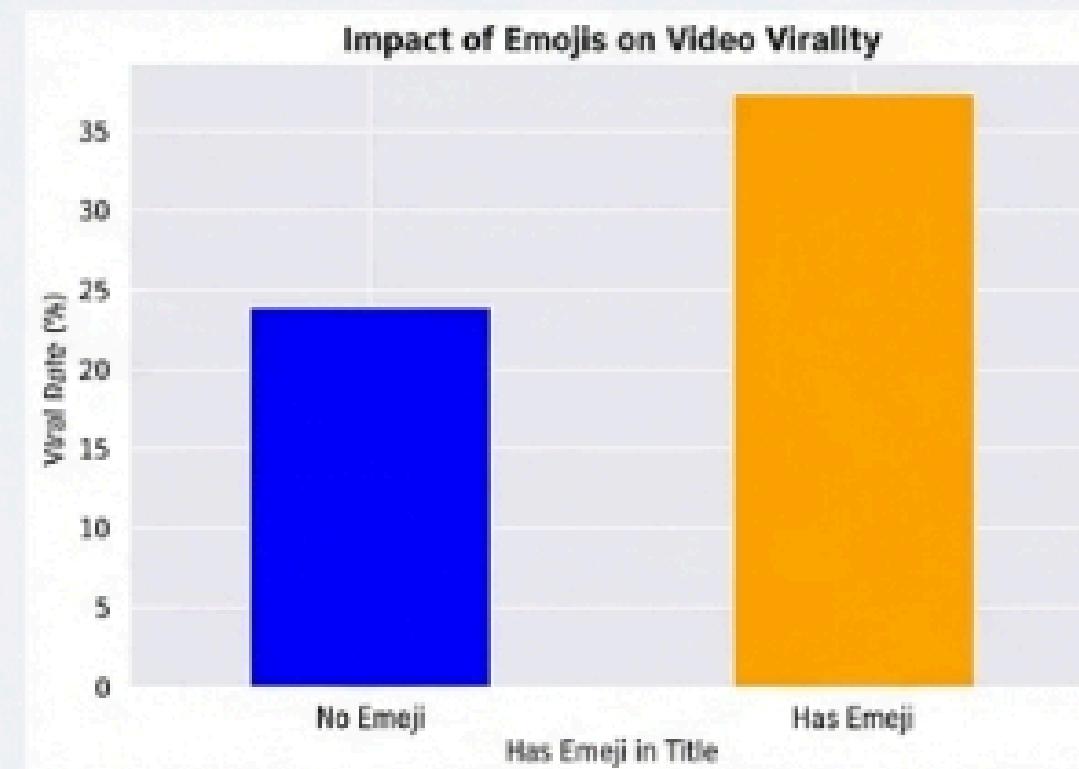
- Ce que nous voyons : Ce boxplot compare le taux d'engagement ((likes + comments) / views) pour les vidéos non virales (0) et virales (1).
- Ce que cela signifie : De manière surprenante, la médiane du taux d'engagement est plus faible pour les vidéos virales.
- Hypothèse : "L'effet de dénominateur". Les vidéos virales atteignent une audience si massive (le dénominateur views est énorme) que le ratio est mécaniquement dilué par des millions de spectateurs "passifs". Le taux d'engagement brut est donc un indicateur trompeur lorsqu'il est utilisé seul.

L'Impact des Titres : Émotion et Affirmation vs. Questionnement

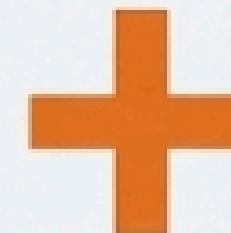
Ces graphiques mesurent l'impact de la ponctuation et des emojis sur le taux de viralité.



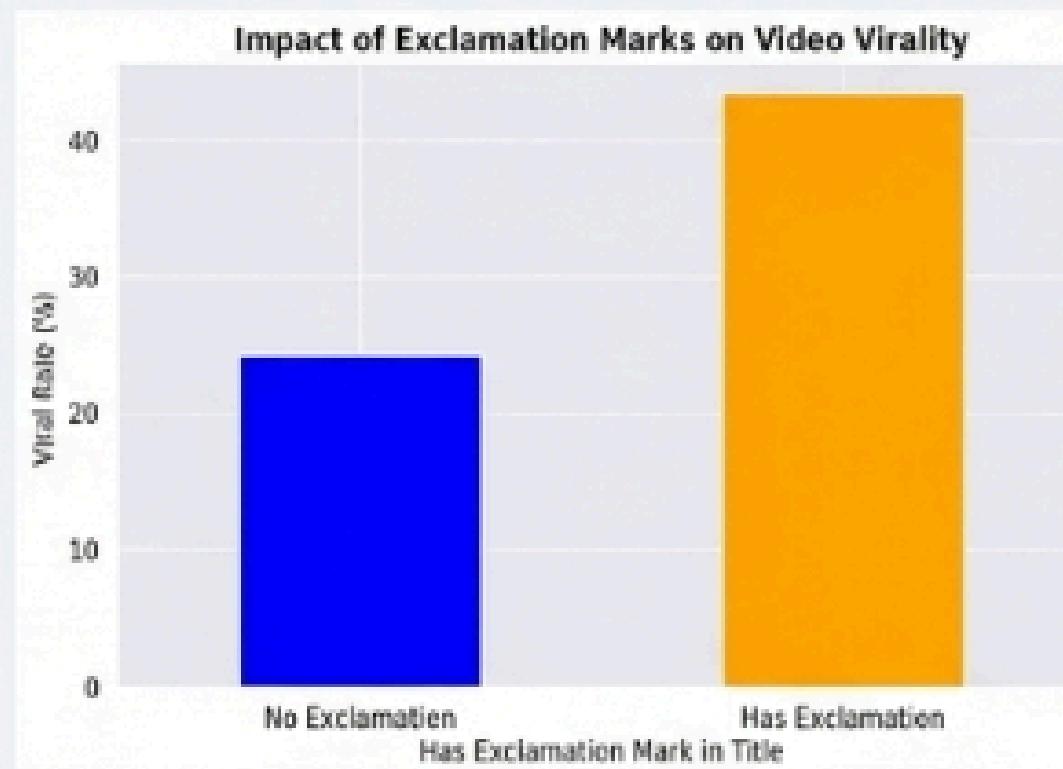
Positif - Emojis



La présence d'un emoji est associée à un taux de viralité nettement plus élevé.



Positif - Exclamations



Un point d'exclamation signale l'enthousiasme et l'émotion, corrélés à la performance.



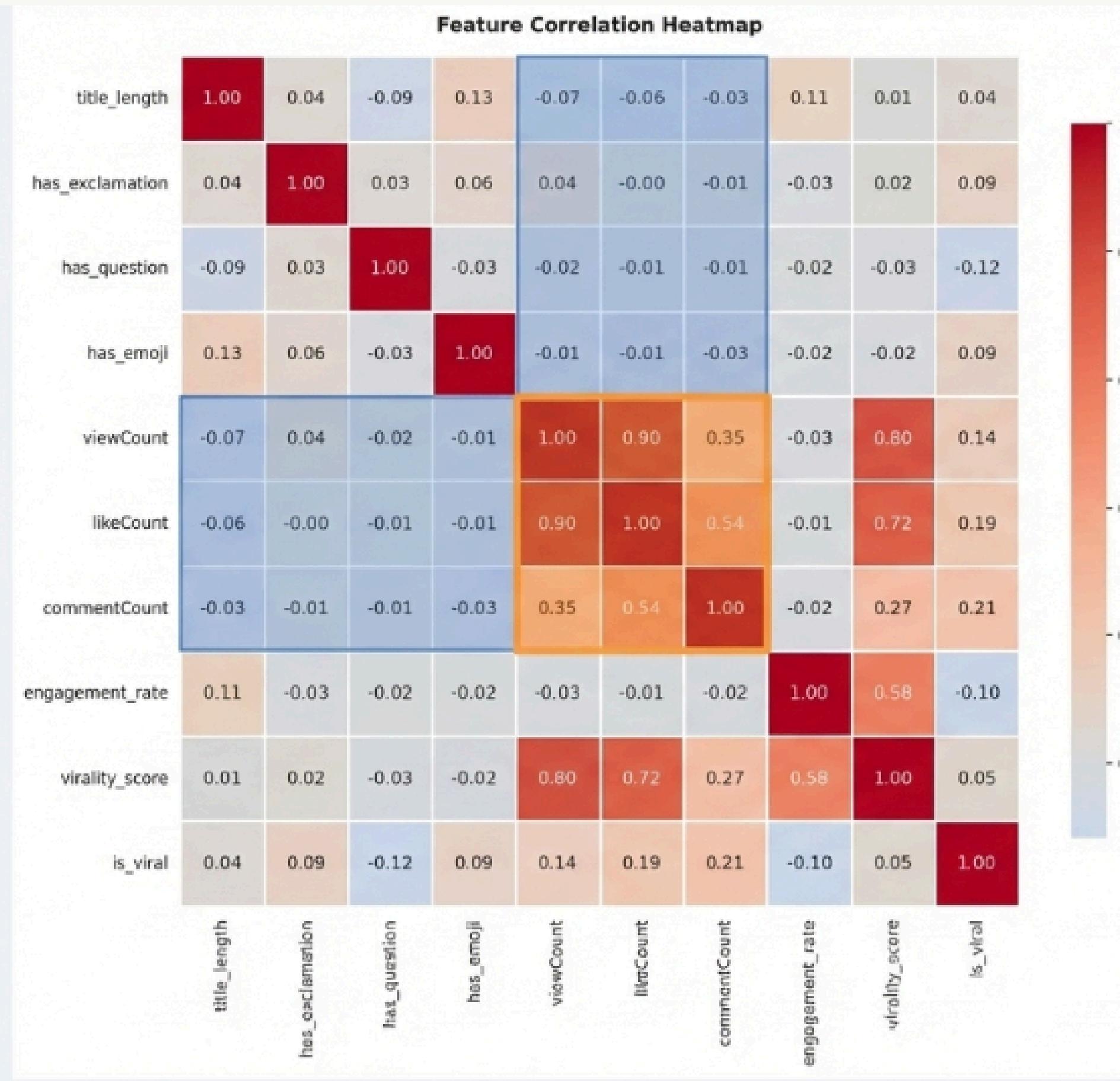
Négatif - Questions



La présence d'un point d'interrogation est corrélée à un taux de viralité plus faible.

Conclusion : Les titres qui affirment un message avec énergie semblent plus performants que ceux qui posent une question à l'audience.

Analyse des Interdépendances : Que nous dit la Matrice de Corrélation ?



- Ce que nous voyons : Cette matrice thermique montre la corrélation linéaire entre nos principales caractéristiques. Les couleurs chaudes (rouge) indiquent une corrélation positive, les froides (bleu) une corrélation négative.
- Insights pour la Modélisation :
- Redondance attendue : viewCount, likeCount et commentCount sont très fortement corrélés entre eux (ex : 0.90 entre vues et likes). Pour certains modèles, il faudra gérer cette multicolinéarité.
- Information unique : Les caractéristiques du titre (title_length, has_exclamation, has_emoji) ont de faibles corrélations entre elles et avec les métriques d'engagement. Cela confirme qu'elles apportent une information nouvelle et complémentaire.

Perspectives

Perspectives: Modélisation

Modèle de Viralité (Classification)

- **Objectif:** Prédire si une vidéo va devenir virale
- **Algorithme:** XGBoost Classifier
- **Pourquoi XGBoost?** Performance sur données tabulaires, gestion des relations non-linéaires

Modèle de Timing Optimal (Régression)

- **Objectif:** Recommander le meilleur moment de publication
- **Algorithme:** LightGBM Regressor
- **Pourquoi LightGBM :** Prédire Performance (views/jour) pour chaque time slot (168 combinaisons: $24h \times 7$ jours)

Perspectives: Evaluation

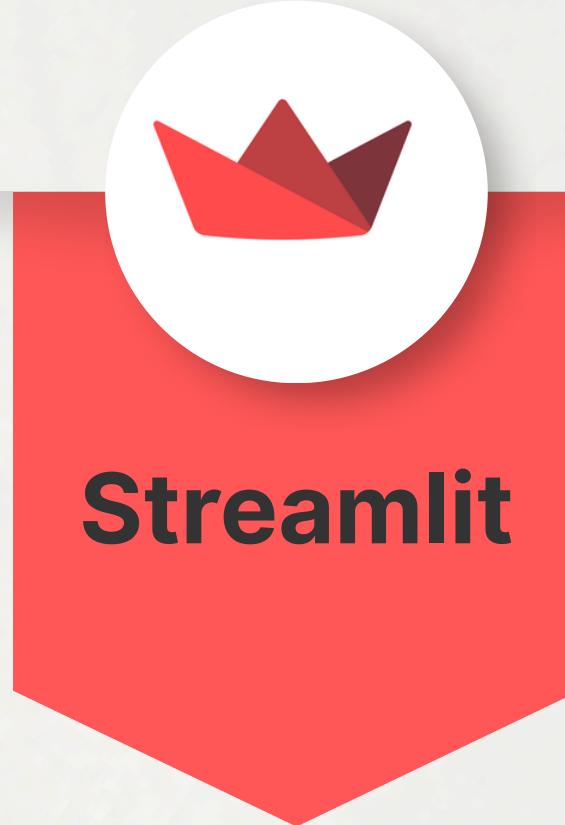
Modèle de Vitalité (Classification)

- Métriques techniques: ROC-AUC, Precision, Recall, F1-Score
- Métrique business: Top-K accuracy mesure si la bonne réponse figure parmi les K meilleures prédictions du modèle, plutôt que d'exiger qu'elle soit la première.

Modèle de Timing Optimal (Régression)

- RMSE, MAE, MAPE pour mesurer l'erreur de prédiction
- R^2 pour la qualité d'ajustement
- Métrique business: estimation de l'impact revenue

Perspectives: Déploiement



Streamlit :Framework Python qui transforme des scripts en applications web interactives .

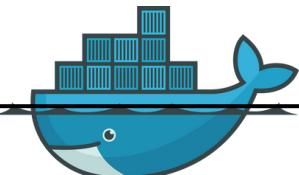


Airflow Plateforme d'orchestration qui automatise et monitore l'exécution planifiée de pipelines de données via des workflows Python

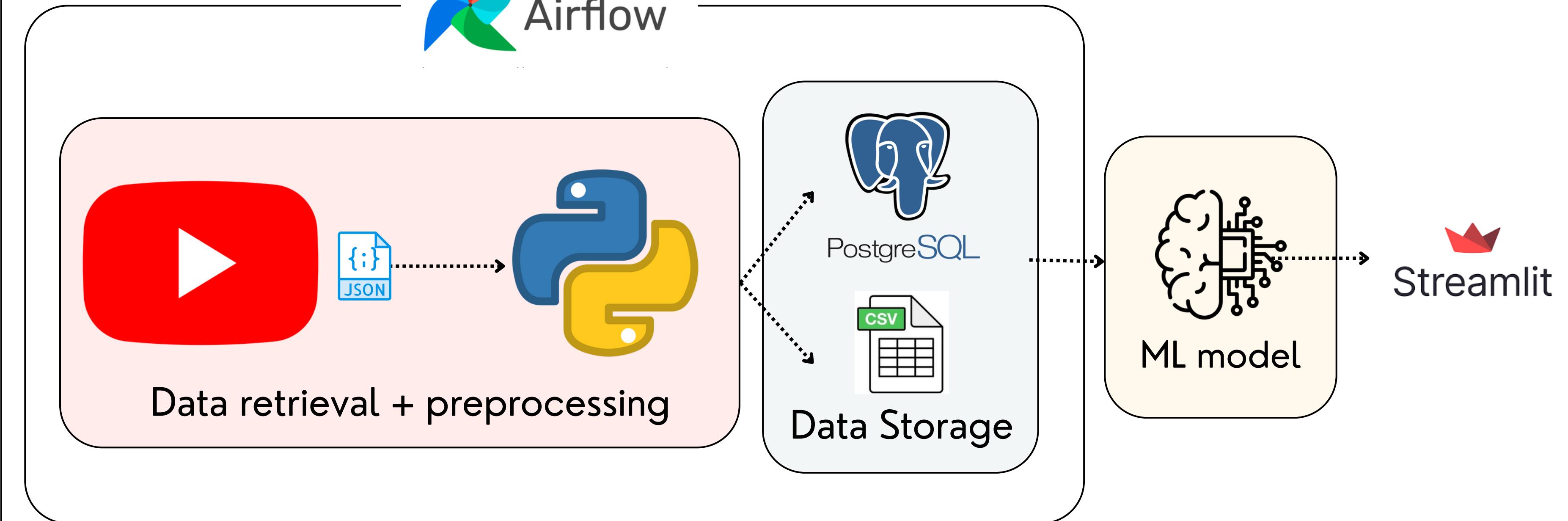


Docker : crée des environnements d'exécution isolés (conteneurs) contenant une application et toutes ses dépendances, garantissant la portabilité et la reproductibilité.

Architecture finale du projet



docker



Ouverture

ANALYSE APPROFONDIE DU CONTENU

- Analyse des commentaires (sentiment, sujets récurrents)
- Extraction audio pour identifier les patterns viraux
- Recommandations basées sur les éléments de succès

ENRICHISSEMENT DES DONNÉES

- Expansion multi-niches pour détection de tendances
- Analyse saisonnière du contenu optimal

INTELLIGENCE AUGMENTÉE

- Assistant IA pour génération de titres optimisés

Merci pour votre attention !

PS. Veuillez consulter le repo github suivant pour voir la version finale:
<https://github.com/RimeAabil/Youtube-Virality-Prediction>