Projet : Analyse de Sentiment et Modélisation de Thèmes des Commentaires YouTube

(Intégrer une lecture des émoji selon le temps disponible)

Comment utiliser l’api de youtube :

1. Avoir un compte google
2. Créer un projet et ses identitifiants
3. Autoriser le projet à utiliser l’api de youtube

API youtube :

Key = API\_KEY

Creation d’un environement virtuel :

python -m venv .venv

activation de l’environement virtuel

source .venv/bin/activate

installation des packages

pip install -r requirements.txt

POC sur l’extraction, collect, stockage des données

Poc sur l’analyse de sentiment avec model pre entrainé et modèle entrainé sur des données labellisé

Poc sur l’entrainemen t d’un modèle sur les données labellisées par le modèle pré-entrainé

Poc sur le topic modeling

Poc sur la restitution

Utiliser MLflow pour la reproductibilité, et le suivi des perfs du modèle et déployer le modèle

# BLOC 3 : ÉLABORATION ET CADRAGE DU PROJET DATA

**Problématique :**

* ~~Comment exploiter la valeur des commentaires youtube pour en faire un outil d’amélioration / en tirer des KPI et améliorer le service produit~~
* **Valorisation des Retours Clients via l'Analyse Automatisée**

Comment mettre en place un système d'analyse automatisé des commentaires YouTube pour extraire des insights permettant d'améliorer en continu le service produit ?

* **~~Optimisation de l'Expérience Client~~** ~~Comment transformer les retours qualitatifs des commentaires YouTube en données quantitatives pour améliorer l'expérience utilisateur et le support client ?~~

**Objectifs :**

Développer un outil permettant de tirer le maximum de valeur des commentaires laissés sur un vidéo YouTube. L’outils doit présenter un certain nombre de KPI pertinents et utiles pour l’amélioration du service. Cela doit être un outil utilisable pour tout utilisateur de YouTube, allant de l’influenceur au simple utilisateur occasionnel. Il doit permettre à l’utilisateur d’avoir une visibilité et une compréhension des retours de ses « spectateurs » (sous condition d’avoir une base de spectateurs importante). Il s’agit d’avoir une idée générale des sujets abordés, du sentiment général/ dominant et de sa polarité.

**Livrables :**

* Liste de KPI
* Modèle d’analyse de sentiment
* Modèle de topic modeling
* Traitements sur les données
* Script de collecte
* Base de données pour l’analyse de sentiments
* Plateforme
* Documentation technique
* Cahier des charges

Cadre réglementaire

<https://www.cnil.fr/fr/ouverture-et-reutilisation-de-donnees-personnelles-sur-internet-la-cnil-publie-ses-recommandations>

• [Recommandations pour les réutilisateurs de données publiées sur Internet](https://www.cnil.fr/fr/recommandations-reutilisateurs-donnees-internet)

[cnil.fr](https://www.cnil.fr/fr/recommandations-reutilisateurs-donnees-internet)  
• [Ouverture et réutilisation de données](https://www.cnil.fr/fr/ouverture-partage-donnees)

Les contraintes et les point de vigilance

Attention au phénomène des haters, mettre en place une modération si ce n’est pas déjà fait par YouTube.

Les enjeux RSE

1. Etat de l’art

* **Comprendre ce qui a déjà été fait** dans le domaine (techniques, outils, jeux de données).
* **Identifier les limites ou lacunes** des approches existantes.
* **Justifier tes choix** méthodologiques.
* **Proposer une amélioration, une innovation ou une approche originale.**

 **Introduction au domaine / situation fictive**

[**https://ia-data-analytics.fr/datascience/big-data-tirer-profit-reseaux-sociaux/**](https://ia-data-analytics.fr/datascience/big-data-tirer-profit-reseaux-sociaux/)

« les données générées sur le web, pour la plupart **via les réseaux sociaux**, contribuent nettement à alimenter le Big Data, lequel est une véritable mine d’or »

Les principaux acteurs, ceux qui sont à l’origine de ces données ce sont nous les utilisateurs attirés par des influenceurs / créateurs de contenus. On peut répartir ces acteurs entre l’offre et la demande, l’entreprise et le client, les influenceurs répondent à la demande de leurs followers. Ce sont des entreprises « unipersonnel », qui comme tout entreprise, pour orienter leur stratégie peuvent utiliser le Data.

« De nos jours, [la quantité de data collectée sur le web](https://ia-data-analytics.fr/datascience/) n’a jamais été aussi exponentielle […]Des données générées pour la plupart par les réseaux sociaux »

En effet, les plus gros influenceurs, tel que Squezzie comptabilise plus de 19 millions d’abonnés sur Youtube, pour plus de 10 milliards de vues sur 1735 vidéos, soit une moyenne d’environ 6 millions de vues par vidéo (**6 432 700)**. En termes de commentaires, on dénombre en moyenne \*. Si les grandes plateformes de réseaux sociaux proposent souvent un suivi sur la productivité des influenceurs en terme de contenu et d’interaction avec les abonnées, ce suivi ne permet que de visualiser que des indicateurs quantitatifs.

« Mais alors, comment les exploiter dans le cadre des réseaux sociaux ? »

​*« À l’ère du numérique, les réseaux sociaux sont devenus le théâtre d’une production massive de données. Chaque jour, des milliards de publications, de commentaires, de likes et de partages alimentent un flux continu d’informations. Ces données, multiformes et en constante expansion, représentent une mine d’or pour les entreprises et les créateurs de contenu.​*

*Parmi ces créateurs, les influenceurs occupent une place prépondérante. Souvent à la tête de véritables entreprises unipersonnelles, ils interagissent quotidiennement avec des communautés engagées. Les commentaires de leurs abonnés, en particulier, sont une source précieuse d’informations. Ils reflètent les perceptions, les attentes et les réactions du public, offrant ainsi un feedback direct et authentique.​*

*Cependant, exploiter efficacement cette richesse informationnelle nécessite des outils d’analyse adaptés. L’analyse de sentiment, par exemple, permet de décoder les émotions et les opinions exprimées dans les commentaires. En combinant les approches du big data et de l’intelligence artificielle, il est possible de transformer ces données brutes en insights stratégiques.​*

*Ce projet vise à développer une méthodologie innovante pour analyser les commentaires des followers, offrant ainsi aux influenceurs et aux marques une compréhension approfondie de leur audience. En mettant en lumière les tendances et les sentiments dominants, cette approche permettra d’optimiser les stratégies de communication et de renforcer l’engagement des communautés.​*

*Dans un monde où l’information est surabondante, savoir écouter et interpréter les voix qui s’expriment sur les réseaux sociaux devient un atout majeur. Ce projet s’inscrit dans cette dynamique, en proposant des solutions concrètes pour valoriser les données sociales et en tirer le meilleur parti.​ »*

D’autant plus qu’aucune plateforme de réseau sociale ne prévoit de fonctionnalité intégrée, « clé en main » permettant à ses utilisateurs d’avoir un suivi, une analyse qualitative des commentaires. En effet, les grandes plateformes, tel que You Tube, ne propose qu’une analyse quantitative avec quelques indicateurs comme le nombre de commentaires, avec un aperçu de quelques récents commentaires. Souvent sous forme de Dashboard avec d’autres indicateurs globaux comme le nombre de likes …

collecter, stocker, traiter et valoriser des données

piloter des projets complexes, intégrer les problématiques des entreprises et élaborer des solutions adaptées qui permettront d’atteindre les objectifs fixés

Véritable expert en apprentissage automatique, il conçoit des modèles de Machine Learning et des indicateurs pour aider la prise de décision dans la stratégie d’entreprise. Il utilise ces modèles prédictifs pour faire des recommandations et aider les dirigeants dans leur pilotage d’entreprise.

Automatiser la collecte de données en mettant en place des tâches planifiées et/ou des flux temps réel, en utilisant des logiciels d’automatisation afin de garantir l’actualisation des données

Développer un processus ETL

Définir la politique de sécurisation des données en évaluant les risques, en qualifiant leur niveau de sensibilité, en identifiant les droits d’accès selon les rôles des différentes parties prenantes et en respectant les exigences légales (ex : RGPD) afin de garantir la bonne utilisation et l’intégrité des données.

Analyser la problématique et le contexte d’un commanditaire en réalisant des entretiens exploratoires, des questionnaires et une analyse de l’existant afin de lui apporter une réponse appropriée.

Cadrer la stratégie de résolution du problème, en utilisant des algorithmes, en traduisant le problème en un problème d’optimisation afin de le résoudre avec les outils des modèles d’apprentissage automatique.

Sélectionner les technologies, les outils et les algorithmes en identifiant les différentes solutions disponibles et en comparant leurs avantages et leurs inconvénients afin de répondre à la problématique du commanditaire au regard des contraintes du projet.

Construire des variables en utilisant des langages de programmation (ex : Python, Scala, R, Julia…) en exploitant des bibliothèques d’analyse de données afin de fournir les meilleures variables au modèle d’apprentissage automatique.

Sélectionner les variables en identifiant les différentes méthodes de sélection de variables possibles, en utilisant des méthodologies d’apprentissage automatique afin d’optimiser la performance du modèle.

Entrainer un modèle d’apprentissage automatique à l’aide de librairies (ex : Scikit-learn XGBoost, TensorFlow, PyTorch) afin d’obtenir des modèles capables de prédictions sur de nouvelles données inconnues.

Optimiser la performance des modèles d’apprentissage automatiqueen modifiant les hyperparamètres et en analysant les prédictions afin de répondre au mieux à la problématique du commanditaire.

Sauvegarder le modèle d’apprentissage automatique entrainé à l’aide d’outils de sérialisation, virtualisation, containerisation, versioning afin de pouvoir le déployer dans des environnements de production.

Déployer des modèles d’apprentissage automatique en utilisant des API et des outils CI/CD afin de le mettre en production.

Superviser le système Machine Learning en sélectionnant des outils de monitoring et en les exploitant afin de détecter les dérives et les bugs du modèle d’apprentissage automatique.

Automatiser les tâches inhérentes au cycle de vie d’un système d’apprentissage automatique à l’aide de pipelines et des outils adaptés afin de maintenir la performance du modèle d’apprentissage automatique.

Mise en situation professionnelle réelle ou fictive. À partir de l’analyse d’une organisation réelle ou fictive de son choix, le candidat propose la gestion d’un projet Data. Il remet au jury un dossier écrit comprenant : Le cadrage du projet - Le dimensionnement du projet - La documentation projet - Le planning projet - Un outil de suivi de projet - Un tableau de bord - Un plan de développement des compétences - Les outils de communication et managériaux utilisés - La présentation d’un cas d’arbitrage rencontré au cours du projet - Une méthodologie de veille - Un plan d’actions relatif aux enjeux RSE, de sécurité, d’éthique et de confidentialité

 **Revue des méthodes d'analyse de sentiment**

https://www.ibm.com/fr-fr/topics/sentiment-analysis

* analyse basé sur des règles : utilisation d’un lexique annoté par le modèle qui ensuite attribue un score
* analyse basé du le ML : régression linéaire, classification naif bayese, svm, DL
* analyse hybride :

les différents types d’analyse de sentiment (repose toutes sur la capacité à identifier la polarité du texte = sentiment générale véhiculé par une partie du texte) :

* **Analyse fine (graduelle)** : regroupe le texte selon différentes émotions et indique le niveau d’émotion. L’émotion reçoit ensuite un score en 0 et 100 (intensité)
* **Analyse basée sur l’aspect (ABSA)** : *pas compris*
* **Détection émotionnelle** : vise à appréhender l’état psychologique de la personne derrière le texte, son état d’esprit et ses intentions au moment de l’écriture. Cette technique peu identifier des émotions plus fines et pas seulement la polarité.

CHATGPT

Les familles de méthode d’analyse de sentiments :

* Approches lexiconiques (basé sur des dictionnaires) : attribut un score de polarité à chaque mot à l’aide d’un lexique (SO-CAL, VADER, SentiStrength) et applique des règles pour gérer la négation.
* Les modèles statistiques supervisés : exploitent les représentation bow ou TFIDF et classifient via SVM Naif Bayse ou régression logistique.
* analyse basée sur les réseaux de neurones profond : rnn lstm, gru, cnn pour la prise en compte du contexte syntaxyque et sémantique de la phrase.
* analyse basée sur des transformers / modèle pré-entrainé : BERT ROBERTA DISTILBERT
* Les approches hybrid :
* L’aspect based analysis : utilise des architectures spécifiques pour détecter et classifier le sentiment par aspect, souvent via des variantes de BERT (par exemple Instruct‑DeBERTa)

 **Revue des techniques de modélisation de thèmes (Topic Modeling)**

 **Outils et bibliothèques existants**

 **Études similaires / projets passés**

 **Critique / Limites des approches**

 **Positionnement de ton projet**

Les projets existants

Quels modèles

Quels traitements

Cassendra aws, stockage forma colonne

1 modèle pour l’analyse de sentiment propre à chaque vidéo ou à améliorer en ajoutant des données (positif négatif ou neutre / ou score de sentiment)

1 modèle de topic modeling propre à chaque vidéo

POUVOIR ANALYSER DES GROUPES DE VIDÉOS

Outil d’analyse automatique : KPI, tableau de bord + génération d’un rapport

Faire attention à la réglementation RGPD, IA ACT … LES PROBLEMES D’ÉTHIQUE, DE CONFIDENTIALITÉ

Des chiffres sur la fréquentation de youtube la valeeur de youtube, les profits que cela dégage

Faire une plateforme sécurisée, compte : id, mp pour que seul les personnes dont on parle puissent voir ce qu’on dit sur elles.

PLANNING

Mars : étude de faisabilité (API, accès aux données, volume de données ...) et choix du sujet

Avril : terminer le cadrage du projet (état de l’art, méthodes, objectifs livrables, outils, limites…)

Mai : Premier POC + terminer la rédaction du bloc 3

Juin : amélioration du poc + rédaction du bloc 1 et 2

Juillet : amélioration et intégration méthode CI/CD

Aout : bloc 1 et 2

Septembre fin soutenance démo

SOURCES

Theses.fr

* <https://theses.fr/2015LYO22007>
* <https://theses.fr/2020LORR0037>
* <https://theses.hal.science/tel-04041435>

google scholar

* <https://link.springer.com/article/10.1186/s12889-020-8342-4>
* <https://serials.atla.com/theolib/article/view/2609/3271>
  + MALLET
  + Gensim : lda
  + « Le processus actuel de modélisation thématique comporte cinq étapes principales : (1) créer un corpus de documents, (2) prétraiter le texte pour rassembler les caractéristiques les plus informatives, (3) traiter le texte en un corpus utilisable par Gensim (c'est-à-dire transformer chaque document en un vecteur), (4) construire le modèle de sujet, et (5) analyser le modèle de sujet »