**Titre RNCP :**

**Élaborer et piloter un projet data**



Sommaire

[Introduction 3](#_Toc196994315)

[I. Cadrage du projet 4](#_Toc196994316)

[1. Contexte et Objectifs 4](#_Toc196994317)

[2. Enjeux et périmètre 5](#_Toc196994318)

[II. Le dimensionnement du projet 6](#_Toc196994319)

[III. La documentation du projet 6](#_Toc196994320)

[IV. Le planning du projet 6](#_Toc196994321)

[V. Un outil de suivi de projet 6](#_Toc196994322)

[VI. Un tableau de bord 6](#_Toc196994323)

[VII. Un plan de développement des compétences 6](#_Toc196994324)

[VIII. Les outils de communication et managériaux utilisés 6](#_Toc196994325)

[IX. La présentation, d’un cas d’arbitrage rencontré 6](#_Toc196994326)

[X. Une méthode de vaille 6](#_Toc196994327)

[XI. Un plan d’action relatif aux enjeux RSE, de sécurité, d’éthique et de confidentialité 6](#_Toc196994328)

# Introduction

Il y a environ 8.21 milliards de personnes sur Terre et 70.7% possèdent un smartphone, 68.7% utilisent internet et 64.7% ont un compte sur les réseaux sociaux. A l’ère du numérique, ces chiffres continuent d’augmenter notamment avec l’essor des réseaux sociaux, qui produisent toujours de données et comptent toujours plus d’utilisateurs. Ces plateformes répondent aux 3 V du Big data, vélocité, volume et variété. Chaque jour des milliards de vidéos, photos, messages, commentaires, likes et autres sont publiés et créent un flux continu d’informations.

*« 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 créateurs de contenus sont parmi les acteurs majeurs de ce flux de données constent. Chaque jour, \*\*

*« de nos jours, la quantité de data collectée sur le web n’a jamais été aussi exceptionnelle […] Des données générées pour la plupart par les réseau sociaux »*

De manière générale, pour les entreprises « classiques », les données sont utilisées pour orienter les décisions stratégiques des entreprises, mieux comprendre sa base client, améliorer un produit ou un service…. Les acteurs du monde de l’influence ne dérogent pas à cette règle. Leur influence repose sur leur capaciter à produire un contenu intéressant pour leurs abonnés et à être aimé. Ils se doivent donc d’être au courant de leur image, de la maintenir et de la faire évoluer ainsi que le contenu pour qu’il continue d correspondre aux envies. Sur les diverses plateformes, des outils leur permettent d’évaluer leur « influence », l’engagement de leurs abonnés grâces à divers indicateurs comme le nombre de likes, de vues, de partages, de commentaires …

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

# Cadrage du projet

## Contexte et Objectifs

Origine / contexte (un grand youtubeur veut avoir améliorer sa connaissance de sa base d’abonnés en utilisant les commentaires laissés sous ses vidéos, il aimerait savoir si les réactions sont positives ou non, et savoir globalement ce qui est dit, les sujets abordés) : Un Youtubeur / créateur de contenu souhaite en savoir plus sur ce qui est dit dans les commentaires à ses vidéos. Jusqu’à maintenant il ne se basait que sur des indicateurs quantitatifs pour évaluer l’engagement de ses abonnées (nombre de likes, de vues, de commentaires…), le nombre de like étant ce qui se rapproche le plus d’un indicateur qualitatif. Pour les influenceurs les plus influents le nombre de commentaires est souvent important et donc impossible à lire entièrement. Actuellement, il est déjà possible pour les créateurs de contenus sur YouTube d'analyser, d'en savoir plus sur leurs abonnés et leur engagement via YouTube studio. Cet outil de YouTube présente les principaux indicateurs quantitatifs mais ne permet pas de savoir ce qui est dit dans les commentaires. Les commentaires sont une donnée particulièrement porteuse d’informations. Ils contiennent les attentes, réactions et les ressenties des abonnés et sont ainsi un authentique feedback.

Utilité : comme pour toutes les entreprises, les feedbacks permettent d’avoir des éléments d’amélioration de l’expérience client sur un produit ou un service. Ainsi, on sait ce qui fait que le client est satisfait ou non et l’entreprise peut ainsi arbitrer entre poursuivre la vente du produit / service en priorisant les améliorations ou simplement stopper la vente. De la même façon un influenceur YouTube utilise les indicateurs du YouTube studio pour savoir si son contenu est aimé, vue (par ses abonnés, des spectateurs occasionnels …), combien de temps ses vidéos sont regardés, l’âge et le genre des spectateurs, les autres chaines qu’ils regardent …. Autant d’informations qui permettent de dresser le profil de ses spectateurs, de savoir ce qui marchent pour eux… Les informations contenues dans les commentaires ont donc tout à fait leur place aux seins de ces indicateurs.

Objectifs :

Mettre en place un pipeline d’extraction, de traitement, d’analyse de sentiments et de topic modeling des commentaires, pour plusieurs vidéos YouTube. Doit atteindre un score de classification d’au moins 85% (à voir pour le topic modeling en fonction de la cohérence).

Mettre en place un site / une API permettant d’accéder aux résultats du pipeline pour une vidéo donnée

Faire un prototype fonctionnel de bout en bout pour étudier la faisabilité

Livrables :

Les kpi : nombre de commentaires positifs vs négatifs / les sujets abordés dans 2 cotés

L’api

Le pipeline

## Enjeux et périmètre

* Enjeux stratégiques, opérationnels et réglementaires
* Limites contraintes et risques : le phénomène des haters + les règles RGPD
* Parties prenantes :
  + Youtubeur ou son équipe / manager
  + DataScientist
* Périmètre : Le périmètre du projet couvre uniquement l’extraction et l’analyse des commentaires des vidéos YouTube pour en tirer des indicateurs stratégiques / porteur de valeur. Il s’agit dans un premier temps d’extraire les commentaires en utilisant l’API de YouTube et de les traiter en utilisant les techniques de traitement de texte. Puis, dans un second temps, de faire de l’analyse de sentiment selon la technique qui sera jugé la plus pertinente et de faire du Topic Modeling pour identifier les sujets principaux selon le sentiment, grâce aux meilleurs modèles. Enfin, générer un rapport/ tableau de bord automatisé qui présentera les principaux résultats. Les résultats ne seront pas visibles sur YouTube Studio et ne pourront donc pas être mis en lien avec les indicateurs déjà existants. \*limites organisationnelles et fonctionnelles\*

# Le dimensionnement et planning du projet

* Charge de travail (le volume d’effort pour chaque phase du projet)
* Ressources nécessaires :
  + Humaines : data scientsit avec des compétences de chef de projet et de data analyst (dasboard), qui connait la réglementation RGPD, idéalement un data engineer pour gérer tout la partie infrastructure.
  + Matérielles : langage de programmation python, bibliothèque de traitement NLTK et spacy, bibliothèques de modélisation gensim, transformers, outils de visualisation \* outils de déploiement docker, streamlit, flask / fastAPI, base de données mangoDB / postgresql, serveurs / puissance de calcule\*, Mlflox pour la CI/CD
* Répartition et ordonnancement des tâches
* Planning prévisionnel avec les ressources nécessaires

# 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

# Une méthode de vaille

# Un plan d’action relatif aux enjeux RSE, de sécurité, d’éthique et de confidentialité