Titre RNCP :

INGENIEUR EN SCIENCE DES DONNEES SPECIALISE EN APPRENTISSAGE AUTOMATIQUE – YNOV (RNCP39586 - niv 7).

Bloc 3 :
Élaborer et piloter un projet data

Sommaire

Introduction

Aujourd'hui, à l'ère du numérique, 70,7% de la population mondiale possède au moins un smartphone, 68.7% utilise Internet et 64,7% a un compte sur les réseaux sociaux. Ces chiffres continuent d'augmenter, notamment avec l'essor des réseaux sociaux, qui produisent continuellement du big data et comptent toujours plus d'utilisateurs. Chaque jour, des milliards de vidéos, photos, messages, commentaires, likes et autres sont publiés, créant 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 constant. Ils interagissent parfois chaque jour avec leur communauté, car la régularité est importante pour eux. Il est essentiel de maintenir ce lien, de continuer à capter l'attention et l'intérêt de leur audience. C'est ce qui leur donne de la visibilité sur les réseaux et leur permet de générer du chiffre d'affaires. Ainsi, chaque jour, des messages, commentaires, vidéos, images et autres sont postés.

« 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 une entreprise, les données sont utilisées pour orienter les décisions stratégiques, mieux comprendre sa base client et améliorer un produit ou un service. Les acteurs du monde de l'influence ne dérogent pas à cette règle. Leur influence repose sur leur capacité à produire un contenu intéressant pour leurs abonnés et à être appréciés (ou non). Ils se doivent donc d'être au courant de leur image, de la faire évoluer, ainsi que leur contenu, afin qu'il continue de correspondre aux envies. Sur les diverses plateformes, des outils leur permettent d'évaluer l'engagement de leurs abonnés grâce à divers indicateurs comme le nombre de likes, de vues, de partages ou de commentaires, sans prendre compte le fond des commentaires.

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

Ce document présente l'élaboration d'un outil d'analyse automatisé des commentaires You Tube, le cadrage, la gestion et le pilotage de ce projet.

I. Cadrage du projet

1. Contexte et Objectifs

L'état des lieux nous révèle qu'il est déjà possible, pour les créateurs de contenus sur YouTube, d'analyser l'engagement de leurs abonnées via YouTube studio. Cet outil intégré à la plateforme présente les principaux indicateurs quantitatifs mais, ne permet pas de savoir ce qui est dit dans les commentaires. Or c'est une donnée particulièrement porteuse d'informations. Ils contiennent les attentes, réactions et les ressentis des abonnés, ainsi ce sont d'authentiques feedback. Cependant, pour les influenceurs avec le plus de succès, les commentaires sont souvent très nombreux, il est donc impossible de les lire entièrement.

C'est dans ce contexte et dans le cadre de son activité, qu'un grand youtubeur souhaite enrichir 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. Jusqu'à maintenant il ne se basait que sur des indicateurs quantitatifs pour évaluer l'engagement de ses abonnés (nombre de likes, de vues, de commentaires…), le nombre de likes étant ce qui se rapproche le plus d'un indicateur qualitatif.

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

En effet, 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 pourquoi le client est satisfait ou non et on peut arbitrer entre poursuivre la vente du produit / service en priorisant les améliorations ou simplement stopper la vente. De la même manière, un influenceur YouTube utilise les indicateurs à sa disposition pour savoir si son contenu est aimé, vu (par ses abonnés, des spectateurs occasionnels …), combien de temps ses vidéos sont regardées, 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 et de savoir ce qu'ils apprécient. Les informations présentes dans les commentaires ont donc tout à fait leur place au sein de ces indicateurs.

Par conséquent, notre objectif en tant que consultants d'une entreprise de consulting, est de mettre en place une infrastructure d'extraction, de traitement et d'analyse

automatisée des commentaires YouTube relatifs à une vidéo donnée. Puis de rendre les résultats disponibles et accessibles via une application pour le client.

Ainsi, les livrables de ce projet sont les suivants :

- La base de données des commentaires
- Les modèles d'analyse
- L'agent IA pour l'interprétation des topics
- La liste des KPI
- L'interface utilisateur et guide d'utilisation (ensemble des fonctionnalités)
- Le code source de tous les modules du projet
- Les tests et les rapports associés
- Conditions générales d'utilisation
- Manuel technique
- Manuel d'utilisation

2. Enjeux et périmètre

Comme évoqué précédemment, les enjeux de ce projet sont dans un premier temps de valoriser les retours « clients ». En effet, les commentaires sont des feedbacks qui révèlent les clés, les leviers de la satisfaction « client » et permettant d'aligner le contenu avec les attentes des clients. De manière plus large, cela permet d'améliorer la connaissance de sa communauté et donc d'orienter / de faire évoluer le contenu et de mieux choisir les partenariats.

Dans un second temps, le flux de feedbacks étant continu et important, il s'agit de pouvoir mettre en place un processus de traitement et d'analyse scalable avec des mises à jour automatisées.

Au-delà des enjeux stratégiques et opérationnels, il y a également des enjeux réglementaires et éthiques. Les commentaires postés sur les réseaux sociaux étant des données personnelles, leur traitement doit *« respecter la loi Informatique et Libertés et le règlement général sur la protection des données personnelles (RGPD) »* et « *être loyal et licite »*. En d'autres termes, l'analyse doit se faire selon une base légale qui l'autorise, traiter des données personnelles sans base légale est interdit. De plus, conformément aux règles RGPD sur la confidentialité / l'anonymat, l'analyse des commentaires doit se faire indépendamment de l'identité de son émetteur. Certains commentaires pourraient être qualifiés de données sensibles car il est possible de déterminer l'orientation sexuelle, l'opinion politique, l'origine ethnique ou l'état de santé d'une personne, d'où la nécessité d'anonymiser pour rendre impossible le profilage des personnes. L'anonymat est d'autant plus important que tous les commentaires ne sont pas toujours « valorisables ». On fait ici référence au phénomène des « haters », des utilisateurs qui

publient des commentaires, négatifs, insultants voire haineux de manière systématique, sans raison. Ces personnes profitent de la protection de leur anonymat pour attaquer, critiquer en masse les personnes qui s'exposent sur les réseaux sociaux. Dans ce contexte, l'analyse de ces commentaires n'apportent aucune valeur, car ces critiques ne sont pas constructives. Ici l'anonymat assure qu'ils ne seront pas identifiés par l'utilisateur.
L'utilisation de l'IA est également réglementé, en fonction de son niveau de risque. Si le risque est trop élevé et donc porte atteinte aux droit fondamentaux, son utilisation est interdite. En fonction du niveau de risque, la réglementation diffère, il est donc essentiel d'identifier le niveau de risque associé au projet pour adopter la bonne approche.

Un autre point de vigilance concerne l'une des parties prenantes, You Tube / Google en tant fournisseur de données. L'ensemble de notre projet repose sur l'API de You Tube et notre capacité à récupérer d'analyser ces données. Nous sommes donc soumis aux règles d'accessibilités et d'utilisation de ces données. Les principaux risques sont la fermeture de l'API, la fin de son accès gratuit, les quotas d'extraction et les mis à jour de l'API.

Ainsi, le périmètre du projet couvre l'extraction, le traitement, l'analyse et la mise à jour des analyses de commentaires français. L'extraction se fait via un pipeline automatisé d'ELT (extraction, chargement et traitement) avec un appel de l'API You Tube Data V3, vidéo par vidéo. Le stockage se fait dans une base de données adapté au format json. L'analyse fait intervenir un second pipeline pour l'analyse de sentiments, le topic modeling et l'interprétation automatisée des classes. Les résultats de cette partie permettront de labelliser les données et d'automatiser le reporting des résultats. Il est exclu du périmètre l'analyse croisée avec les indicateurs de You Tube car il n'y a pas d'intégration prévu avec la plateforme. De plus, le projet couvre uniquement l'analyse des commentaires français (dans la limite de 200 commentaires minimum) et n'intègre donc pas l'analyse pour les autres langues. L'analyse du contenue de la vidéo est également exclu et l'interface client n'intégrera pas de chatbot.

En définitive, notre projet s'inscrit dans un cadre réglementaire régit par le RGPD, l'IA Act et les règles d'accès et d'utilisation de l'API de You tube.

II. Le dimensionnement du projet

1. Les ressources matérielles

Les ressources matérielles comprennent dans un premier temps l'espace de travail. C'est-à-dire, les bureaux dans lesquels l'équipe travaillera et le matériel associé (chaises, bureaux, écrans ...). Chaque membre devra disposer d'un PC adapté à ses tâches (développement, analyse de données, création de dashboard, gestion / pilotage de projet). Ces ressources sont gérées pas la boîte de consulting.

En ce qui concerne les licences et logiciels pour les développeurs, leur environnement de travail doit se composer d'un IDE, tel que Visual Studio Code, des bibliothèques de base pour coder en python et faire de l'analyse textuelle.

Un serveur hébergera les modèles d'analyse et la base de données sera stocké par un conteneur Docker. L'ensemble du projet sera également hébergé sur une solution cloud (Amazon AWS) pour la production, la partie locale servira principalement pour les tests et mises à jour futures. Pour la partie développement Git hub permettra de sauvegarder les codes, l'environnement et de versionner le travail. C'est également un bon moyen de travailler en équipe sur un même projet. En termes de logistique, tous les membres de l'équipe devront avoir un accès à la plateforme collaborative de gestion de projet (Jira) et à un outil de communication (Microsoft Teams, Outlook).

2. Les ressources humaines

L'organisation, le pilotage et la gestion du projet nécessiterons les compétences d'un chef de projet. Son rôle, une fois l'analyse de besoin effectuée, est de coordonner les tâches entre les différentes parties prenantes, d'assurer la compréhension du projet et son suivi. Il travaillera en collaboration avec :

- 3 Data Scientist, responsables du nettoyage et de l'analyse des données.
- Un Data Engineer : responsable de l'architecture technique du projet, de l'intégration continue (CI/CD) et de la gestion de la base de données.
- Un Data analyst : élaboration du tableau de bord et le choix des KPI.

Précisons que l'équipe affectée à ce projet travaillera à temps plein, selon un contrat de 35h par semaine sur 3 mois.

3. Chiffrage & analyse de faisabilité

L'estimation précise des coûts permet d'assurer le succès du projet. Nous distinguons les couts associés à la main-d'œuvre, au matériel, l'utilisation de licences logiciels, les couts liés à la communication (réunion, déplacements...) et les frais généraux. Les coûts de main d'œuvre comprennent les salaires et avantages sociaux, les couts matériaux comprennent les serveurs, PC et équipements de bureau, les licences et outils de conception. Les frais généraux : loyer des bureaux, fournitures de bureau

Ressources humaines	Quantité	Salaire /an	Total sur 3vmois
Data scientist	3	46 K/an	11k €
Data engineer	1	48K/an	12k €
Data analyst	1	45K/an	11k €
Chef de projet	1	47K/an	11k €
Total = 67k €			

Ressources matérielles	Quantité	Coût
PC	1 /personnes	1200 x 6 = 7200 €
Serveur	1	10-30€/ mois
Amazon aws (hebergement cloud, berrock, CI/CD)	1 compte /prs	
Licences de développements (IDE, bibliothèques)	1 compte /prs	0 -45$/mois
Git hub	1 compte /prs	0-21$/mois
Plateforme de gestion de projet	1 compte /prs	0-13.53$/mois
Outil de communication	1 compte /prs	2.1€/mois
Total = K €		

Soit un budget final d'environ X€ sur la durée du projet, sachant qu'ils continueront de courir sur la durée de vie du projet.

L'ensemble des frais généraux, des couts de main-d'œuvre et de matériels seront gérées par l'entreprise et non par le client. D'autant plus que certains de ces coûts (main d'œuvre, pc, outils de communication...) sont pris en charge par l'entreprise indépendamment du projet. Ainsi, seuls les couts de server et d'utilisation de l'API, voire de formation, sont inhérents au projet.
En termes de faisabilité techniques (main-d'œuvre et technologies), nous disposons de tous les éléments nécessaires ou pouvons y avoir accès sans problème. Au niveau des délais, la livraison est prévue début septembre, nous disposons de 3 mois.

III. Documentation du projet

Notre projet fait intervenir 3 parties prenantes directes et indirectes :

- Le client, dont on doit satisfaire les attentes (youtubeur)
- Google / You tube, dont on exploite les données et l'API
- Notre équipe de projet
- Les utilisateurs de You Tube, auteurs des commentaires

- Exigences fonctionnelles et techniques

Pour répondre aux attentes du client, nous allons déployer un Software As A Service (SaaS) qui regroupe plusieurs fonctionnalités clés de base. Pour une vidéo donnée l'outil permet, après authentification du client, de :

- Voir le nombre de commentaires traités à une date donnée
- Connaitre la répartition des commentaires par catégorie de sentiment
- Visualiser les commentaires (une dizaine) selon la catégorie de sentiments grâce à un filtre
- Connaître la répartition des commentaires par sujet
- Visualiser quelques commentaires appartenant à chaque sujet (une dizaine)
- Visualiser les termes caractéristiques de chaque sujet
- Les noms attribués à chaque sujet par l'agent IA (pour assistance)
- Renommer les sujets si le nommage par défaut ne semble pas en adéquation avec les exemples et les termes caractéristiques
- Choisir de faire l'analyse des sujets pour l'ensemble des commentaires ou par type de sentiment
- Améliorer le nettoyage des commentaires, en permettant à l'utilisateur de renseigner / signaler du vocabulaire inapproprié.

Il s'agit donc d'un tableau de bord interactif, qui du point de vue technique doit répond à certaines exigences. La première est l'extraction de données, via l'API You Tube V3 de GCP, qui doit se faire de manière automatique, dès lors que l'utilisateur (le client) précise la vidéo dont il veut analyser les commentaires. Cependant, l'ensemble des commentaires ne sera pas visible par l'utilisateur, uniquement quelques exemples. Ensuite, en vue d'être analysée, les données doivent être traitées, nettoyées de manière adaptée pour optimiser les résultats de l'analyse. Après les différentes étapes du traitement (tokenisation, nettoyage des stop-words, lemmatization, stemming ...) les données propres sont stockées dans la base de données (selon le délai de stockage autorisé). Les données ainsi stockées seront labellisées d'un type / d'une catégorie de sentiment et d'un sujet. L'analyse de sentiments reposera sur un modèle préalablement sélectionné, après entrainement. Le topic modeling, étant un type d'apprentissage non supervisé, le modèle sera propre à chaque ensemble de commentaires, il n'y aura pas de sauvegarde du modèle. De plus, puisque le nombre de commentaires évolue avec le temps et conformément à la réglementation, la base de commentaires doit aussi se mettre à jour de même que l'analyse. Par conséquent, en fonction du nombre de commentaires ajoutés à chaque mis-à-jour, soit ils seront intégrés aux résultats soit l'analyse sera relancée. En effet, s'il ne s'agit que de 10 commentaires, cela n'est pas suffisant pour dégager un nouveau sujet, ils seront donc intégrés aux anciens. Au-delà de 200 (chiffre donnée à titre d'exemple), on considère que l'on peut relancer l'analyse et

donc mettre à jours les résultats. De la même façon il faudra mettre à jour le modèle d'analyse de sentiments en testant régulièrement ses résultats.
Ainsi, soit le tableau de bord affiche les résultats des analyses enregistrés dans la base de données, soit il faut mettre à jour la base. Dans ce cas, soit il y a peu de commentaires et on applique le modèle d'analyse de sentiments et de topic modeling sur le même nombre de sujets, soit il y a beaucoup de commentaires et on relance la recherche du nombre de sujets.

- Exigences éthiques et réglementaires

En ce qui concerne la réglementation, nous sommes soumis aux règles en vigueurs localement au niveau européen, dans un premier temps, puis aux conditions générales d'utilisation de l'API.
Conformément aux règles du RGPD et de l'IA Act, notre analyse s'inscrit dans le cadre de l'intérêt légitime (article 6 RGPD). L'extraction sera donc limitée aux informations pertinentes pour l'analyse, tel que le texte et la date de publication (principe de minimisation des données). L'accès à la base de données sera sécurisé grâce à une méthode d'authentification et les données stockées seront anonymisé. Ainsi, même si nous n'utilisons pas les pseudonymes et que nous ne faisons pas d'analyse ciblée, l'anonymat sera préservé. Une stratégie d'évitement sera également mise en place, afin de minimiser le phénomène de « haters » autant que possible et d'éviter l'analyse de données sensibles. Par exemple grâce à la détection de certaines émotions caractéristiques ou type de contenue. Dans le même temps, l'ensemble de nos traitements concernant ces données seront répertoriés.
Notre projet appartient à la catégorie de risque limité puisqu'il prévoit une interprétation assistée par un agent IA, des sujets issues du Topic modeling. De ce fait, le système d'IA influence la décision, même s'il n'y a pas d'interaction direct avec l'utilisateur ou de décision automatique pouvant affecter une personne spécifique (analyse non-ciblé), ses libertés fondamentales, la discriminer ou la dignité humaine. Il s'agit uniquement d'un outil d'assistance, d'aide à la compréhension. Pour éviter que l'interprétation introduise des biais politique, culturel… un filtre sera appliqué en amont, lors du nettoyage des données.
Pour garantir la transparence, nous devons également documenter l'utilisation de l'IA et présenter les conditions et conseils d'utilisation et fournir une charte d'usage. L'utilisateur doit comprendre que l'analyse de l'agent IA n'est pas absolue, ce n'est qu'une assistance qu'il peut contester. L'agent IA peut se tromper, ainsi pour laisser la place à l'interprétation humaine, une fonctionnalité permet à l'utilisateur de nommer un sujet si l'interprétation automatique ne lui semble pas cohérente. C'est pour ces raisons que nous affichons des exemples de commentaires pour illustrer les sujets identifiés.
En ce qui concerne le règlement de l'API YouTube V3 et de Google, les principes qui s'appliquent particulièrement à notre projet et que nous devons respecter sont les suivants :

- Quota d'1 millions de commentaires par jour et par projet, accessibles gratuitement sur GCP.
- Délais de stockage : les données ne peuvent être stockées plus de 30 jours.
- Principe de consistance des données : les données doivent toujours être conformes à ce qui existe sur You Tube, en cas de modification ou de suppression des commentaires.
- Afficher les données les plus récentes.
- Fournir une politique de confidentialité avec mention de l'utilisation du service YouTube V3, les références vers les Terms of Service et la politique de Google/YouTube.
- Ne pas faire d'analyse sur l'agrégation de plusieurs chaînes appartenant à des personnes distinctes.

Par conséquent, afin de garantir une utilisation responsable de l'outil et pour qu'il soit conforme aux réglementations :

- Les appels d'API seront optimisés, afin de ne pas dépasser le quota, sans limiter les fonctionnalités de l'outil.
- Garder des données à jour et conformes à l'existant, tout en prévoyant une mise à jour des données tous les 30 jours afin de respecter les délais de conservation.
- Rédiger une politique de confidentialité mentionnant l'utilisation de l'API You Tube V3 et des conditions d'utilisation. Ce document sera accessible depuis l'outil.

Ainsi les principes moraux fondamentaux du projet sont la transparence, la protection et le respect de la confidentialité et la non-discrimination.

IV. Planning du projet

1. Méthode de gestion de projet et outils de planification

Une bonne gestion de projet garantie son succès, compte tenu des contraintes de temps, de budgets... Parmi les méthodes existantes, nous avons privilégié les méthodes agiles plus flexibles. En effet, en plus de permettre l'amélioration continue de chaque composante du projet, elles permettent de travailler en collaboration avec le client. De ce fait, on s'assure à chaque étape clé que les attentes soient respectées.

La méthode Scrum se base sur des sprints, des cycles de courte durée bien définis (1 à 4 semaines) itératifs et fait intervenir / collaborer l'ensemble des parties prenantes. L'idée est de favoriser l'implication / la collaboration entre elles, optimiser la gestion du temps, et la productivité. En effet, l'idée est qu'à la fin de chaque sprint on produise un livrable utilisable ou à améliorer jusqu'à satisfaction. Un schéma explicatif est disponible en annexe.

Le projet est découpé en plusieurs sprints qui s'organise autour de 5 étapes.

- Product backlog : le Product Owner (c'est l'intermédiaire entre l'équipe de développement et le client dont il représenter les intérêts) et le client définissent les fonctionnalités du projet et réalisent un cahier des charges.
- Sprint planning meeting : réunion à chaque début de sprint où l'équipe détermine l'objet du sprint, les tâches et objectifs à réaliser et la durée du sprint.
- Le daily scrum : point quotidien rapide sur la progression du sprint (fait, à faire et obstacles rencontrés), si nécessaire on réoriente les tâches de la journée.
- Le sprint review : à la fin de chaque sprint on fait un bilan et on test les fonctionnalités livrées, on détermine ce qui est à améliorer et les fonctionnalités à ajouter, conjointement avec le Product Owner.
- Répétition du cycle en améliorant et en ajoutant les fonctionnalités en fonction des retours du client / Product Owner.

Cette organisation favorise les échanges fréquents entre les différentes parties prenantes étant réguliers, ce qui garantit un alignement du projet avec les attentes du client. Elle permet également une certaine flexibilité puisque le cahier des charges peut évoluer au cours des sprint. Cependant, les réunions quotidiennes, les reviews et planning meeting nécessitent une forte implication, un fort engagement des équipes.

La méthode Kanban se base plutôt sur un flux continue de travail avec des ensembles de tâches qui évoluent selon 3 états dans un tableau : « à faire », « en cours », « à valider », « terminé ». Les tâches sont traitées par ordre de priorité et selon les capacités de l'équipe, sans délais associé. Les fonctionnalités sont livrées dès que possible, pas besoin d'attendre la fin d'un sprint pour se réorienter ou livrer. Cette méthode permet de visualiser la progression des tâches, la productivité et les goulots d'étranglement. Ainsi, la priorité est de diminuer le nombre de tâches « en cours ». Cependant, l'absence de délais, de structure temporelle représente un risque pour le respect des délais de livraison.

D'où le choix de la méthode hybride Scrumban qui allie la flexibilité de la méthode Kanban et la structure de la méthode Scrum. On y retrouve le tableau de visualisation des tâches, « backlog », « à faire », « en cours », « review » et « terminé » avec l'idée de limiter le nombre de tâches « en cours ». On l'associe à des sprints plus légers avec des objectifs définis en fonction des priorités et des capacités de l'équipe, c'est-à-dire un backlog flexible. Les réunions quotidiennes et les reviews sont également conservés.
Ainsi, cette méthode hybride nécessite également un outil de planification hybride. Nous allons allier le tableau Kanban au diagramme de Gantt. Le tableau permettra de visualiser l'évolution des tâches et de limiter le nombre de tâches en cours et le diagramme de

Gantt, de visualiser leur enchainement et les échéances. On retrouve dans les 2 visuels les mêmes tâches / objectifs, elles évoluent de manière simultanée grâce aux mis à jour faites lors des daily scrum au cours desquels on fait l'état des lieux des travaux et on réoriente la trajectoire si besoin.

2. Planning et assignation des tâches

Formellement, Jira permet d'établir une « chronologie » des différentes tâches ou, comme ici, des principales étapes du projet à l'image d'un diagramme de Gantt. Dans notre cas il s'agit plus d'une première estimation du planning. En fonction de l'évolution des tâches les délais pourraient évoluer avec de nouvelles fonctionnalités et donc de nouvelles tâches. Pour autant, cela ne doit pas retarder la date de livraison.

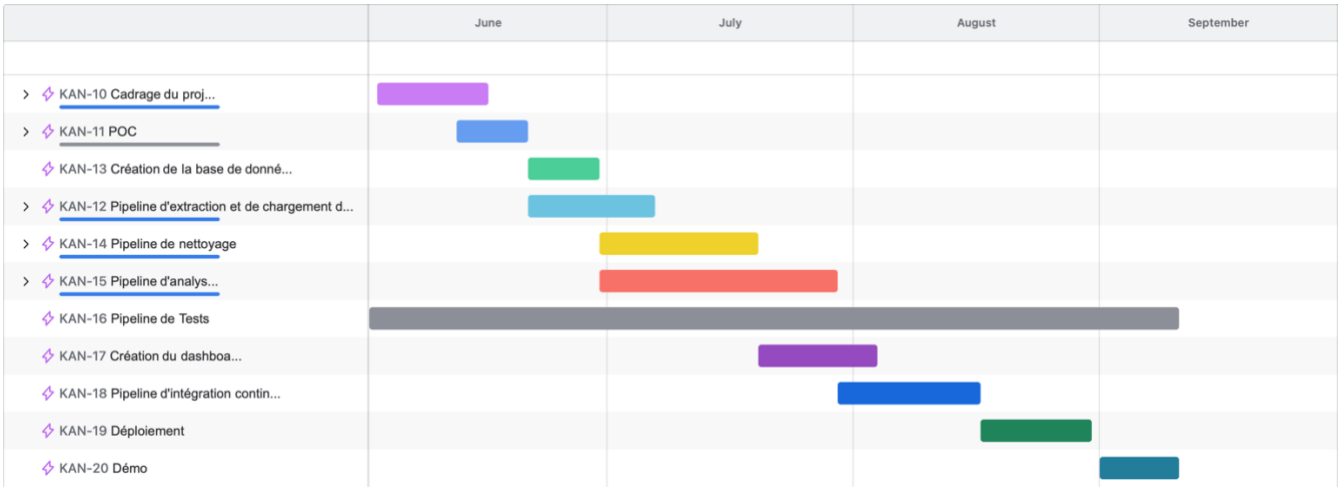

Figure 1: diagramme de Gantt

On peut ainsi découper le projet en 5 phases :

- La phase de cadrage : il s'agit l'analyse de besoin, écriture du cahier des charges… et réalisation d'un premier POC.
- La phase d'extraction, de chargement et de nettoyage (ELT) : on y trouve la création de la base de données et du pipeline d'extraction et nettoyage des données.
- La phase d'analyse : entrainement et sélection des différents modèles d'analyse de sentiments, topic modeling et choix de l'agent IA. Ainsi que la création des graphiques pour le dashboard.
- La phase de test : cette phase commence dès le début des travaux de développement, tout au long du projet tout est testé et amélioré.
- La phase de déploiement et d'intégration continue : la phase finale au cours de laquelle on développe l'interface utilisateur avec les résultats, en suivant respectant une organisation CI/CD avant de déployer, tester et de faire la démonstration au client.

En termes d'ordonnancement des phases et tâches, toutes les phase dépendent de la phase de cadrage ca en en découle les objectifs, fonctionnalités de base et le périmètre pour éviter des dérives. La phase d'analyse dépend de la phase d'ELT ce qui en fait la phase clé ou critique du projet. La dernière phase de déploiement est soumise à l'évolution de la phase d'analyse et la production des résultats. La phase de test est la seule qui est indépendante.

Pour la planification des tâches, conformément à la méthode gestion de projet sélectionnée, nous utilisons un tableau Kanban pour chaque sprint. L'image en annexe présente un exemple. La matrice ci-dessous, présente un premier exemple de répartition des tâches pour chaque rôle selon les phases sans entrée dans le détail.

Phases / tâches	**Chef de projet**	**Data Scientist**	**Data Engineer**	**Data Analyst**
Phase de cadrage				
• *Analyse de besoin*	R	I	I	I
• *Rédaction du cahier des charges*	R	C	C	C
Phase d'ELT				
• *Accès API Youtube (tokens)*	C	I	R	I
• *Code d'extraction des données*	A	C	R	I
• *Création de la base de données*	A	C	R	I
• *Nettoyage des données*	A	R	C	I
• *Tests*	A	R	C	I
Phase d'analyse				
• *Entrainement et sélection de modèles d'analyse de sentiments*	A	R	I	I
• *Topic modeling*	A	R	I	I
• *Choix de l'agent IA*	A	R	I	I
• *Création des tableaux de bord et choix des kpi*	A	C	C	R
• *Tests*	A	R	C	I
Phase de test				
• *Tests en continues*	A	R	R	I
• *Organisation d'un pipeline de tests*	A	C	R	I
Phase de déploiement				
• *Développement du pipeline CI/CD*	A	C	R	I
• *Démonstration et livraison*	R	I	I	I

Légende :

A *= approbateur* ***C*** *= consulté* ***I*** *= informé* ***R*** *= réalisateur*

Poste	Charge de travail	Ressources
Data scientist 1	Traitement, nettoyage des données	1 PC, IDE avec les bibliothèques python de base,
Data scientist 2	Analyse de sentiment	1 PC, IDE avec les bibliothèques python de base
Data scientist 3	Topic modeling	1 PC, IDE avec les bibliothèques python de base
Data engineer	Architecture du projet Extraction, chargement des données dans la base de données, intégration continue	1 PC, IDE avec les bibliothèques python de base
Data analyst	Développement des visualisations	1 PC, IDE avec les bibliothèques python de base
Chef de projet	Cadrage, suivi du projet, organisation des sprint	1 PC

3. Points de vigilances

Cette méthode a ses inconvénients, notamment le fait de devoir gérer les 2 tableaux en même temps pour les garder à jour. Cela induit un risque d'erreur dans le suivi du projet si on oublie de mettre à jour le Gantt au fur et à mesure de l'évolution du tableau Kanban et donc que le projet diverge. Il est également important que toute l'équipe comprennent bien le fonctionnement de cette méthode pour qu'elle puisse être efficace. Ainsi, il faudrait que le chef de projet soit à l'aise avec cette méthode et puisse l'expliquer à l'équipe et veiller à son application tout au long du projet si nécessaire.

Au-delà des risques inhérents aux méthodes et outils de gestion du projet, la nature même du projet créé des risques. Puisque on laisse la possibilité au cahier des charges d'évoluer avec de nouvelles fonctionnalités, les tâches à réaliser ne sont, à priori, pas toutes connues, d'où cette méthode de gestion de projet. Par conséquent, le diagramme de Gantt donne les délais pour chaque tâche, mais ils sont susceptibles d'évoluer. Il y a donc une incertitude sur la durée de chaque tâche et ce manque de visibilité pourrait engendrer des retards de livraison. Ainsi, il est important dans un premier temps de bien définir les fonctionnalités de base qui doivent obligatoirement être livrées au client. Puis de spécifier le périmètre en excluant toutes fonctionnalité trop longues, complexes à développer dans les temps. L'équipe se concentrera sur le développement et

l'amélioration des fonctionnalités de base selon les attentes du client puis sur leurs évolutions.
On peut distinguer 2 types de tâches critiques, celles avec une forte dépendance et celles qui passent trop de temps au statut « en cours ». Les tâches appartenant à la première catégorie sont les tâches appartenant à la phase ELT, car sans les données il est impossible d'avancer sur les autres tâches. Les tâches du second type seront identifiées grâce à l'outil tout au long du projet.

V. Un outil de suivi de projet (tableau de bord)

1. Outil de suivi

Pour suivre l'avancement du projet, nous avons privilégier l'outils de gestion de projet Jira d'Atlassian. Il intègre dans un même outil, la planification, le pilotage et le suivi de projet. Nous avons déjà présenté les fonctionnalités relatives à la planification précédemment avec le tableau « chronologie » similaire au diagramme de Gantt et le Tableau Kanban pour l'affectation et la réalisation des tâches au cours des sprints.
Jira intègre également des fonctionnalités de pilotage et de suivi de projet. Chaque ticket peut être affecté et commenté et poss possède également un historique qui ressence l'ensemble des actions relatives au ticket.

2. Indicateurs et tableau de bord

Jira intègre un tableau de bord « Résumé » permettant d'apprécier l'avancement du projet. Il se met à jour automatiquement au fur et à mesure de la définition des tâches / tickets, de leur affectation et réalisation. Ainsi, il présente différents indicateurs qui aide à l'évaluation de la productivité de l'équipe, comme le nombre d'éléments terminés (sur les 7 derniers jours), ou encore un rappel des échéances conformément à ce que qui est indiqué dans le diagramme de Gantt.

On retrouve également une vue d'ensemble sur la répartition des tâches selon leur état. On peut donc visualiser la productivité et les possibles goulots d'étranglement avec le nombre de tâches « terminé » et « en cours ».
Il intègre également un graphique permettant d'évaluer la charge de travail des membres de l'équipe, grâce au nombre de tickets/ tâches affectées à chaque membre. Ainsi, au début de chaque sprint, lors du meeting de planification, les tâches / objectifs sont identifiés et affectés, de manière à ne sur chargé personnes. Par la suite, tout au long du sprint / du projet, le chef de projet pourra si nécessaire ré-évaluer l'allocation des ressources au cours du projet.
Une vision globale du tableau de bord est disponible en annexe.

Cet outil propose également différentes formes de rapports permettant de suivre / évaluer la productivité. Notamment, le diagramme de flux cumulé présente l'évolution de l'état des tickets. Ce rapport sert à identifier les goulots d'étranglement et les tâches

critiques. Le graphique brun down du sprint permet d'apprécier l'évolution du travail restant au cours du sprint.

VI. Pan de développement des compétences

1. Identification et évaluation des compétences

Le tableau suivant identifie les compétences et niveaux requis pour ce projet, les métiers mobilisés et leur niveau actuel. Donc, après sélection des collaborateurs affectés à ce projet selon leurs compétences, nous évaluons leurs compétences.

Compétences	Poste	Niveau requis	Niveau actuel
Gestion de projet (Scrumban /Jira)	Chef de projet	Avancé	Débutant
Communication		Avancé	Avancé
Juridique /éthique		Avancé	Avancé
Python	Data scientiste	Avancé	Avancé
NLP		Avancé	Avancé
Analyse de sentiments		Intermédiaire	Intermédiaire
Topic modeling		Avancé	Avancé
Tests		Intermédiaire	Débutant
Python	Data analyste	Avancé	Avancé
Data visualisation		Intermédiaire	Intermédiaire
Streamlit		Intermédiaire	Intermédiaire
Python	Data engineer	Avancé	Avancé
Extraction de données, API		Intermédiaire	Avancé
Base de données		Intermédiaire	Intermédiaire
Intégration continue / MLOps/ AWS		Avancé	Avancé
Docker		Intermédiaire	Intermédiaire
Tests		Intermédiaire	Débutant

Malgré son expérience en gestion de projet, le chef de projet n'est pas expérimenté en ce qui concerne la méthode Scrumban et ne maitrise pas l'outil Jira. Le reste de l'équipe maitrise globalement ses sujets, mais doit monter en compétence pour l'implémentation des tests. Plus particulièrement, notre Data Engineer manque d'expérience en intégration continue.

2. Plan de développements de compétences

Dans un premier temps, il est nécessaire de former toute l'équipe à la méthode Scrumban. La formation expliquera les méthodes Scrum et Kanban dans un premier

temps pour amener à Scrumban avec des exercices de mise en situation. Cela associée à une formation à l'outil Jira basique pour les data scientists, engineer et analyst et un peu plus poussé pour le chef de projet. Une formation de 2 à 3 jours organisée en amont du projet.

Ensuite, une formation à l'implémentations de tests unitaires pour l'équipe de développement, avec présentation du module de test pytest. Il s'agira de préférence d'une formation gratuite en ligne, puisqu'il en existe et que la documentation de pytest est disponible en ligne. Il n'y aura pas de temps dédié à cette formation, elle se fera en autonomie, puisque la documentation est disponible, les collaborateurs pourront la consulter autant de fois que nécessaire.

VII. La présentation, d'un cas d'arbitrage rencontré

Pendant la phase de développement d'une « proof of concept » (POC), la question s'est posée du choix du modèle d'analyse de sentiments. A ce stade, il s'agit uniquement de s'assurer de la faisabilité du projet, donc d'implémenter différents modèles et de tester le champ des possibles. Ainsi, nous avons constaté plusieurs choses :

- Pour nos données non étiquetées, il faut un modèle pré-entrainé.
- On pourrait utiliser un modèle simple d'apprentissage supervisé, entrainé sur un jeu de données similaire labelisé (en français avec des textes courts...).
- Il est également possible d'utiliser un transformer, notamment disponibles chez Huggins face, et de l'entrainer sur des données similaires si besoin.
- Ou bien faire appel à une IA via une API

La problématique porte donc sur le choix de la méthode d'analyse de sentiments.

1. L'apprentissage automatique et base de données adaptées

En ce qui concerne l'utilisation d'un modèle d'apprentissage automatique pré-entrainé sur des données similaires, il est assez difficile de trouver ce genre de données. Le jeu de données le plus populaire concerne des reviews de films issus du site d'Allociné. Il s'agit globalement des textes plus longs et mieux rédigés avec un style différent que ceux que l'on trouve sous les vidéos youtube. Ce qui laisse penser que le modèle ne serait pas efficace sur nos données. Une solution serait de définir des règles spécifiques à des référence, expression non connue des modèles, mais cela serait très couteux en temps.

2. Transformers et performances incertaines

A l'inverse, certains transformers semblent donner des résultats satisfaisants. Parmi les modèles transformers utilisés, l'un était pré-entrainé sur les reviews mais n'était pas le plus performant. Étant donné que nos données ne sont pas labelisées, pour qualifier les résultats de « satisfaisant », il a fallu vérifier à la main les résultats obtenus sur plusieurs

textes pour les différents modèles. Cependant, nous n'avons aucune garantit que cela sera efficace sur l'ensemble des commentaires. En effet, nous ne disposons d'aucun moyen de vérifier la qualité des résultats.

3. L'intelligence artificielle et les couts infinis

La solution semble être de faire appel à une IA. Nous avons testé certains textes avec la version gratuite de ChatGpt et ils sont très satisfaisant. Les expressions particulières, et le sarcasme sont compris. Néanmoins, ce service est limité puisque son utilisation est payante via l'API et nous serions dépendant de la plateforme sans possibilité de faire évoluer le modèle, ou d'expliquer ses résultats. Les couts associés seraient infinis.

4. Une solution hybride

Si on veut allier performance et explicabilité, sans engendrer de couts trop élevés il est possible d'allier apprentissage supervisé et IA. En effet, les IA tel que Chatgpt ou Claude sont très performante en compréhension du langage et de ces subtilités. Une utilisation ponctuelle d'un agent IA pour labelliser les données pour ensuite entrainer un modèle d'apprentissage supervisé, permet d'allier performance et explicabilité. Même sans contexte ces agent peuvent labeliser efficacement nos commentaires, ce qui assure une bonne base s'entrainement pour les modèles de machine learning classique à fort pouvoir explicatifs. Cette solution n'engendrera pas d'augmentation des couts ni des délais. En effet, des agents IA sont accessible dans l'environnement d'Amazon AWS.

1 Veille, éthique et gouvernance des données

1. Méthodologie de veille

Une bonne méthode de veille garantie la performance du projet, le respect des couts et délais. Il est donc essentiel de rester informé des nouvelles avancées dans le domaine de l'apprentissage automatique, de l'intelligence artificielle, qu'elles soient technologiques ou réglementaires. Plus particulièrement, la veille pour ce projet concerne :

- L'actualité sur l'API You Tube V3 et ses conditions d'utilisation
- Les règlementations sur l'utilisation des données personnelles et de l'IA
- Les techniques de NLP

Un abonnement à la newsletter de le Commission Nationale de l'Informatique et des Libertés (CNIL) permet d'être informé de l'actualité. La CNIL propose également plusieurs ressources pédagogiques pour tous les profils. Pour être plus actif dans la veille, l'organisation de webinaires ou d'ateliers permet d'intégrer plus efficacement les informations / connaissance grâce aux échanges et mises en situations. D'autant plus que ce sont des temps dédiés à l'apprentissage, la prise d'information. La CNIL organise notamment des webinaires de décryptage de sujets ou d'actualités liées à la protection

des données, gratuitement. Ainsi, participer régulièrement à ce webinaire permet de rester à jour sur la réglementation de d'assurer leur bonne application.

En ce qui concerne les technologies, la participation annuelle à des conférences tel que Viva tech est un bon moyen d'être au courant des grandes avancées techniques et technologiques. Cela permet également de découvrir des solutions innovantes et d'échanger et d'apprendre de la communauté Data science. De même, l'abonnement à Médium, une plateforme collaborative de partage de connaissances, permet de consulter des documents scientifiques ou techniques relatifs à l'utilisation d'outils. Cependant, il faut rester vigilent quant à la pertinence et la fiabilité des documents publiés.

En ce qui concerne l'actualité sur les conditions d'utilisation de l'API YouTube v3, l'abonnement à la newsletter de Google Cloud plateforme est un moyen d'être notifié des actualités. Néanmoins, les informations concerneront l'ensembles de outils GCP.

De manière générale, la configuration d'alertes Google permet d'automatiser la veille dans tous les domaines. Grâce à la recherche des mots clés, les liens des articles associés sont regroupés et envoyés par mail. Cette méthode permet un réel gain de temps, c'est l'information qui vient à nous. Cependant, le risque est d'être noyé dans le « flow » de mails. De plus, puisqu'il s'agit d'une simple recherche de mots clé, les résultats ne sont pas toujours pertinents. S'il y a peu de mots clés recherchés il y a moins de risque de se perdre parmi toutes les informations, sinon un mail récapitulatif des informations pertinentes serait plus efficace.

Toutes ces méthodes de veille permettent aux équipes d'anticiper les changements. Ainsi, d'adapter les projets au cadre réglementaire, d'utiliser les technologies les plus efficaces ou d'en éviter certaines. L'anticipation entre particulièrement en jeu lorsqu'il s'agit d'évolution concernant des projet déjà lancés.

2. Un plan d'action relatif aux enjeux RSE, de sécurité, d'éthique et de confidentialité

Notre projet, comme de nombreux projets en data science, comprend des enjeux. De responsabilité sociétale et environnementale, d'éthique, de confidentialité et de sécurité.
Au niveau environnementale, l'utilisation des modèles IA, des LLM et modèles d'apprentissage automatique ont un réel impact, même s'il est difficilement mesurable.

Leur utilisation mais surtout leur entrainement consomme énormément d'énergie et produit beaucoup de gaz à effet de serre.

*« Selon le MIT Technology Review 5, la seule phase de pré-entraînement de GPT-3 a généré l'équivalent de 626 000 kg de CO_2, soit **71.9 tours de la terre en voiture ou la fabrication de 3 244 ordinateurs portables**… »*

Plus le modèle est complexe avec un nombre de paramètres important, plus l'impact environnementale sera important. D'autant plus qu'ils nécessitent une grosse quantité de données, stockées dans des data centers particulièrement énergivores et consommateurs d'eau pour les refroidir.

« Il faudrait 4 à 6 fois la consommation annuelle du Danemark en eau rien que pour refroidir les centres de données d'IA d'ici à 2027 »

Par conséquent nous avons décidé d'utiliser des modèles dit légers et pré-entrainé autant que possible pour ce projet. Dans le même temps, nous allons mener un travail d'optimisation des pipelines, notamment de ceux qui font des appels d'API. En ce qui concerne l'hébergement cloud, il existe de cloud « vert ».

Au niveau éthique, l'utilisation d'IA comporte des risques pour lesquels un plan d'action réglementaire à déjà été mis en place. Les actions mise en place pour limiter les risques et respecter la réglementation sont présentés dans la documentation du projet. De même pour les données personnelles. Lorsque l'on traite des données personnelles, potentiellement sensibles, la sécurité et la confidentialité des données sont nécessairement en jeu. Le plan d'action dans ce cas est également principalement dicté par la réglementation comme expliqué dans la documentation du projet. Les données anonymisées sont stockées dans une base de données sécurisée. De plus, l'accès à l'outil d'es pas ouvert, il est soumis à une authentification de l'utilisateur.

Sources

https://www.cnil.fr/fr/entree-en-vigueur-du-reglement-europeen-sur-lia-les-premieres-questions-reponses-de-la-cnil
https://www.cnil.fr/fr/me-mettre-en-conformite/rgpd-par-ou-commencer
https://www.mailjet.com/fr/blog/bonnes-pratiques-emailing/methode-agile-scrum/#chapter-2
https://slack.com/intl/fr-fr/blog/productivity/methode-de-gestion-de-projet
https://blog-gestion-de-projet.com/gestion-de-projet/outils-methodes-gestion-projet/
https://institut-superieur-environnement.com/blog/lintelligence-artificielle-une-pollution-cachee-au-coeur-de-linnovation/

Annexe

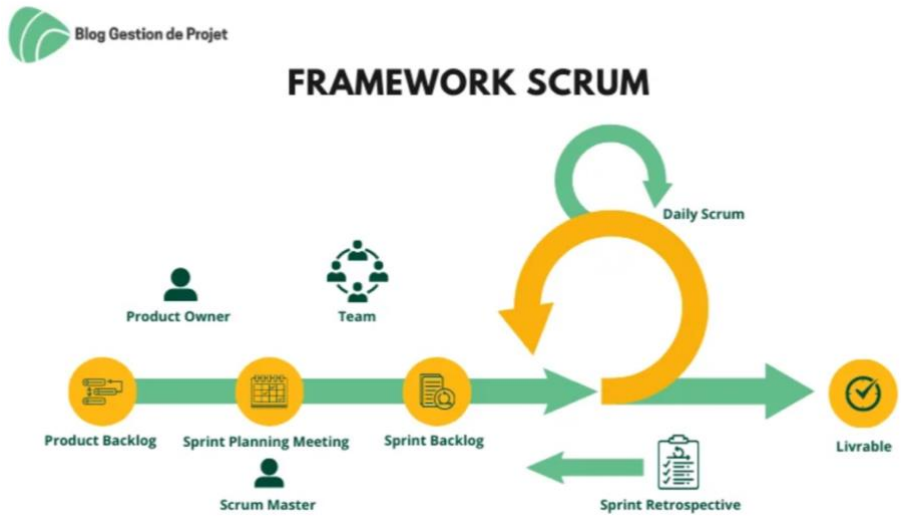

Figure 2: schéma explicatif de la méthode SCRUM

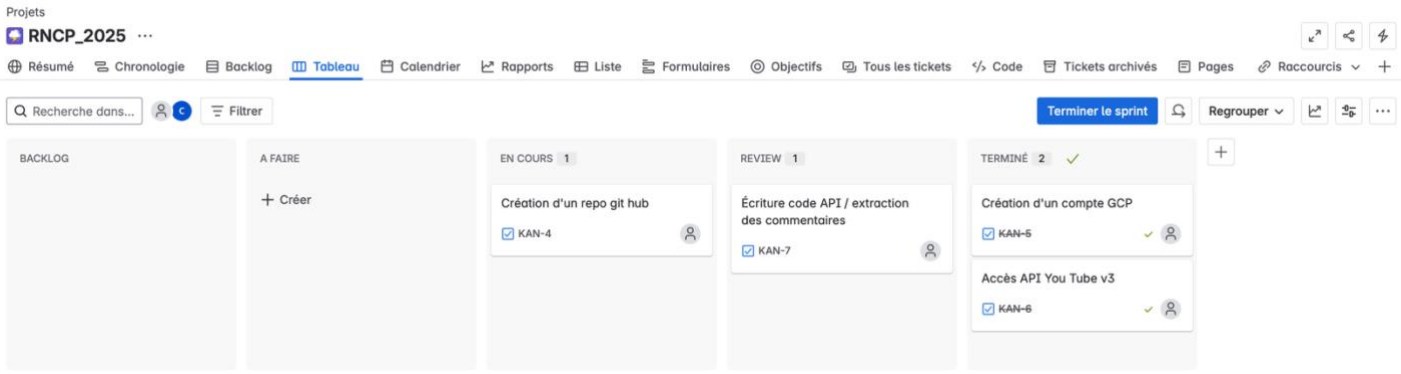

Figure 3: Exemple de tableau Kanban de Jira

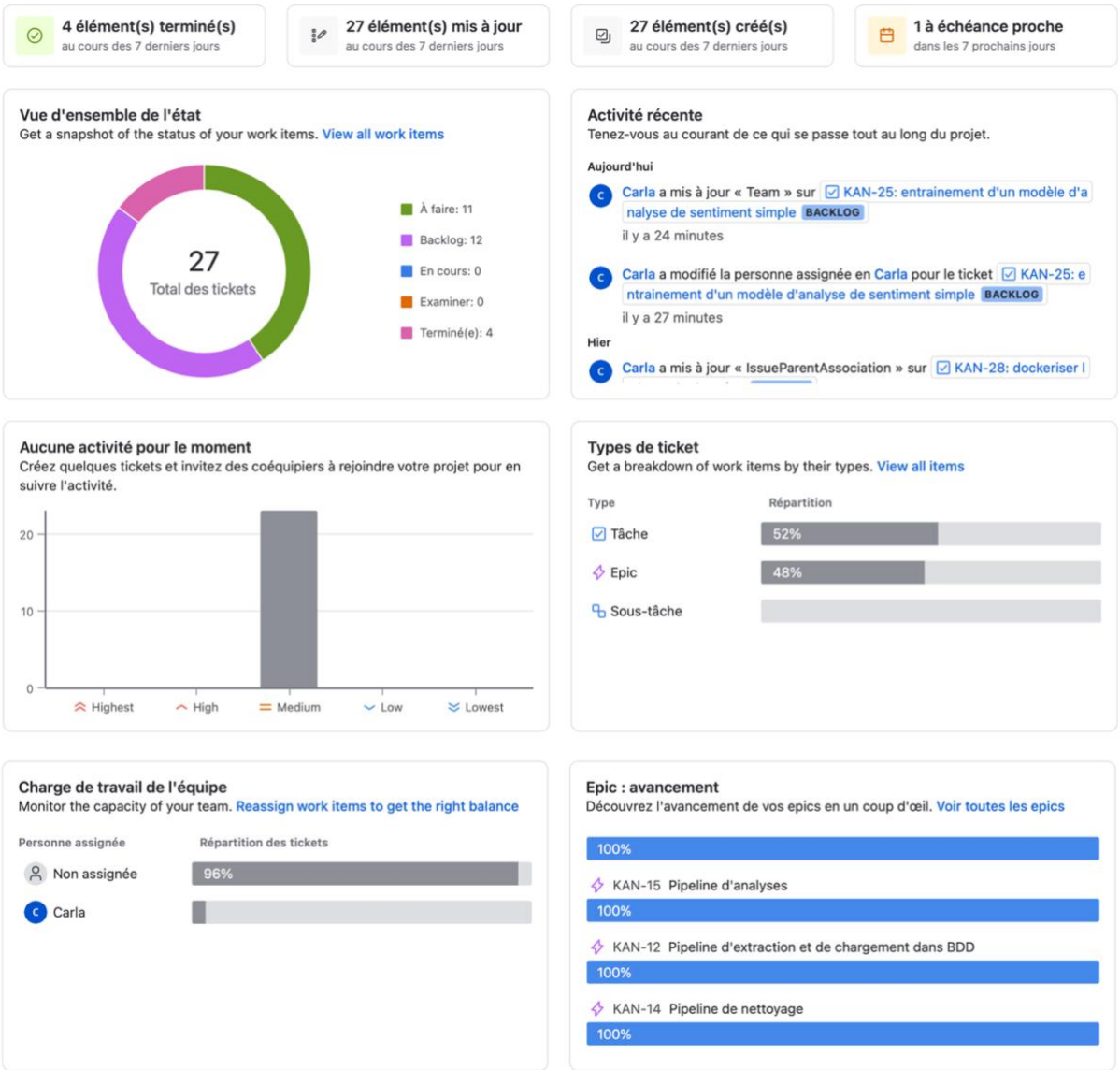

Figure 4: Tableau de bord résumé de Jira