

Répertoire national des certifications professionnelles (RNCP)

Développeur en intelligence artificielle

Projet chef d'oeuvre

# Customer Intent

Projet soutenu par : Mirai IIDA

Organisme d'alternance : Devoteam

Organisme de formation : Simplon

## Table des matières

<b>Remerciements</b>	<b>4</b>
<b>Introduction</b>	<b>5</b>
<b>Compréhension besoin client</b>	<b>5</b>
<b>Etat de l'art</b>	<b>6</b>
Types de l'analyse des sentiments	6
Analyse des sentiments basée sur les aspects	7
<b>Éléments de conception technique</b>	<b>8</b>
Conceptions comportementaux	8
User stories	8
Diagramme comportemental	9
Conceptions structurelles	10
Diagramme des composants	10
Modélisation des données	10
Diagramme Modèle Conceptuel de Données (MCD)	10
Diagramme Modèle Physique des Données (MPD)	12
Définition des routes Front et Back	13
<b>Choix techniques</b>	<b>15</b>
Choix des technologies	15
Choix de l'intelligence artificielle	16
<b>Réponse finale apportée ; ce qui a été réalisé</b>	<b>17</b>
Sources à récolter	17
SGBD	17
MongoDB	17
SnowflakeSQL	18
Solution de l'IA	19
Application	20
Monitoring	21
<b>Mise en oeuvre du projet</b>	<b>22</b>
Organisation technique	22
Domaine d'hébergement	22
Exigence de programmation	22
Accessibilité	22
Compatibilité navigateurs	22
Types d'appareil	22
Sécurité	22
Logging et monitoring	22
Environnement de développement tout au long de la production	23
Gestion de projet	23
Composition de l'équipe	24
Retours d'expérience sur les outils, techniques et compétences à l'œuvre	24

<b>Bilan du projet et les améliorations envisageables</b>	<b>24</b>
<b>Conclusion</b>	<b>24</b>

## Remerciements

Je remercie de blablabla

## Introduction

De nos jours, il est devenu courant pour les consommateurs de consulter les avis en ligne concernant un produit ou un service afin de guider leur décision d'achat. Les entreprises ont désormais la possibilité d'analyser ces avis pour adapter leur offre aux attentes des clients. L'évaluation des avis d'utilisateurs constitue pour ces dernières un moyen de stimuler leurs ventes. Cependant, le nombre d'avis disponibles en ligne augmente énormément chaque année et nécessite le recours à des outils qui permettent d'automatiser leur analyse.

C'est dans ce contexte que la méthode d'**analyse des sentiments** par intelligence artificielle se développe. C'est un outil puissant pour mieux comprendre ce que les consommateurs pensent de la marque et des produits d'une entreprise. L'analyse des sentiments permet de comprendre rapidement les motivations et les raisons qui expliquent les comportements des consommateurs (points forts ou faibles d'un produit, positionnement par rapport à la concurrence, etc.).

Ce projet vise à analyser les contenus textuels publiés sur Internet en vue d'**étudier les intentions** d'achats ou de consommation d'un produit ou d'un service. Il s'agit d'une analyse des commentaires et des échanges des internautes sur leurs souhaits ou impressions. Les données sont extraites de sites marchands, des réseaux sociaux ainsi que des forums de consommateurs et médias web. Elles sont étudiées quantitativement et qualitativement pour en générer des indicateurs d'appréciation.

## Compréhension besoin client

Devoteam souhaite développer en interne une application d'analyse de l'intention dans le domaine des assurances qui répondrait aux besoins suivants :

- Connaître l'intention d'un consommateur sur un produit, un service
- Générer des indicateurs d'appréciation
- Prédire les sentiments et comportements des consommateurs

A partir d'avis de consommateurs sur diverses assurances, nous allons procéder à plusieurs types d'analyse de sentiments afin de prédire les sentiments, les émotions et les comportements des clients. Nous allons également fournir de nombreux indicateurs d'appréciation qui porteront sur la satisfaction, l'évolution et la répartition de ces sentiments. Les données collectées proviendront de différents canaux tels que des sites web spécialisés sur les avis de consommateurs, des forums et des réseaux sociaux (Facebook et Twitter)

## Etat de l'art

### Types de l'analyse des sentiments

L'analyse des sentiments repose sur l'utilisation du traitement du langage naturel (NLP) et peut être réalisée suivant de nombreuses méthodes différentes, dont les quelques exemples suivants.

- **L'analyse des sentiments par polarité**

L'analyse des sentiments par polarité permet de classer des phrases selon les mots qui les composent, de manière à leur attribuer une "polarité" qui peut être positive, neutre ou négative, par exemple, un avis "j'aime beaucoup le nouveau design de votre site web" est analysé une expression des sentiments "positif". On peut également appliquer cette méthode aux émotions (colère, joie, tristesse) , à l'urgence et aussi aux intentions (intéressé versus non intéressé).

- **L'analyse fine des sentiments**

L'analyse fine des sentiments permet d'améliorer la précision de la polarité en étendant les catégories pour inclure, très positif, positif, neutre, négatif et très négatif.

- **L'analyse des sentiments basée sur les aspects**

L'analyse des sentiments basée sur les aspects (ABSA) consiste à analyser le texte pour identifier les différents aspects qui le constituent et déterminer pour chacun d'entre eux les sentiments correspondants. Cette analyse permet d'obtenir des résultats plus détaillés et précis car l'analyse basée sur les aspects examine de manière précise les informations contenues dans un texte, par exemple, un texte "le sushi est mauvais" est analysé une expression de sentiment "mauvais" apparaît près de l'expression d'aspect "sushi", respectivement.

## Analyse des sentiments basée sur les aspects

Dans ce chapitre, on examine les tendances techniques d'ABSA et de ce pré-processing. Une tâche essentielle de cette méthode est d'attribuer aux mots d'un texte un ou plusieurs aspects, pour ensuite déterminer le sentiment correspondant à l'ensemble de ces aspects.

Jusqu'en 2014 environ, les chercheurs utilisaient principalement deux algorithmes classiques qui permettaient d'extraire les aspects et de classer les polarités. Les principaux algorithmes alors utilisés étaient le Support Vector Machine (SVM), le naive bayes, le random forest et le Conditional Random Fields (CRF). Le prétraitement était communément réalisé avec des méthodes telles que le Part-Of-Speech (POS) tag, le bag of word ou encore le Term frequency-inverse document frequency (Tf-idf) [1][2][3]. Les aspects dans les documents pouvaient ainsi être exprimés par un nom, un adjectif, un verbe ou un adverbe, mais en pratique 60-70 % des termes d'aspect étaient des noms explicites [4]. Les termes d'aspect pouvaient également consister en des entités de plusieurs mots telles que *"battery life"* et *"spicy tuna rolls"*, etc. Par conséquent, en étiquetant chaque mot, le modèle pouvait extraire l'aspect du groupe de mots.

Un changement majeur a eu lieu vers 2014, lorsque les principales technologies de deep learning ont été publiées : c'est l'année où ce domaine a commencé à s'épanouir avec la publication de la technologie des Generative Adversarial Networks (GANs) et la sortie de Tensorflow [5][6]. D'autre part, dans le domaine du NLP, un modèle de traduction automatique appelé "Seq2Seq" a été publié par Google en 2014 [7], et l'année suivante, une technique appelée "Attention mechanism" a été présentée par Bahdanau et al. pour compenser l'inconvénient de Seq2Seq [8]. Cette technologie a joué un rôle majeur dans l'amélioration des performances de Google Translate [9]. Elle a également été utilisée dans le modèle Bidirectional Encoder Representations from Transformers (BERT) de Google [10].

Depuis quelques années, les chercheurs montrent un grand intérêt pour l'identification simultanée des aspects et des sentiments [1]. Ils ont commencé à étudier l'ABSA en utilisant des techniques de deep learning telles que le recursive neural network, l'architecture Long Short-Term Memory (LSTM), et des techniques de prétraitement avec Word embedding [11][12]. Les techniques utilisées dans d'autres domaines du langage naturel, comme la traduction automatique, commencent actuellement à être utilisées dans l'ABSA. De nombreuses recherches ont été menées sur la prédiction simultanée de l'aspect et de l'extraction des sentiments à l'aide de modèles qui adoptent le modèle Seq2seq avec l'Attention mechanism[13][14].

## Éléments de conception technique

Ce chapitre détermine la conception technique à partir des diagrammes structurels et comportementaux.

### Conceptions comportementaux

#### User stories

Nous choisissons de lister les User stories en les regroupant par fonctionnalités. Nous définissons différentes catégories d'utilisateur de l'application afin d'établir les User stories. Nous distinguons les catégories suivantes :

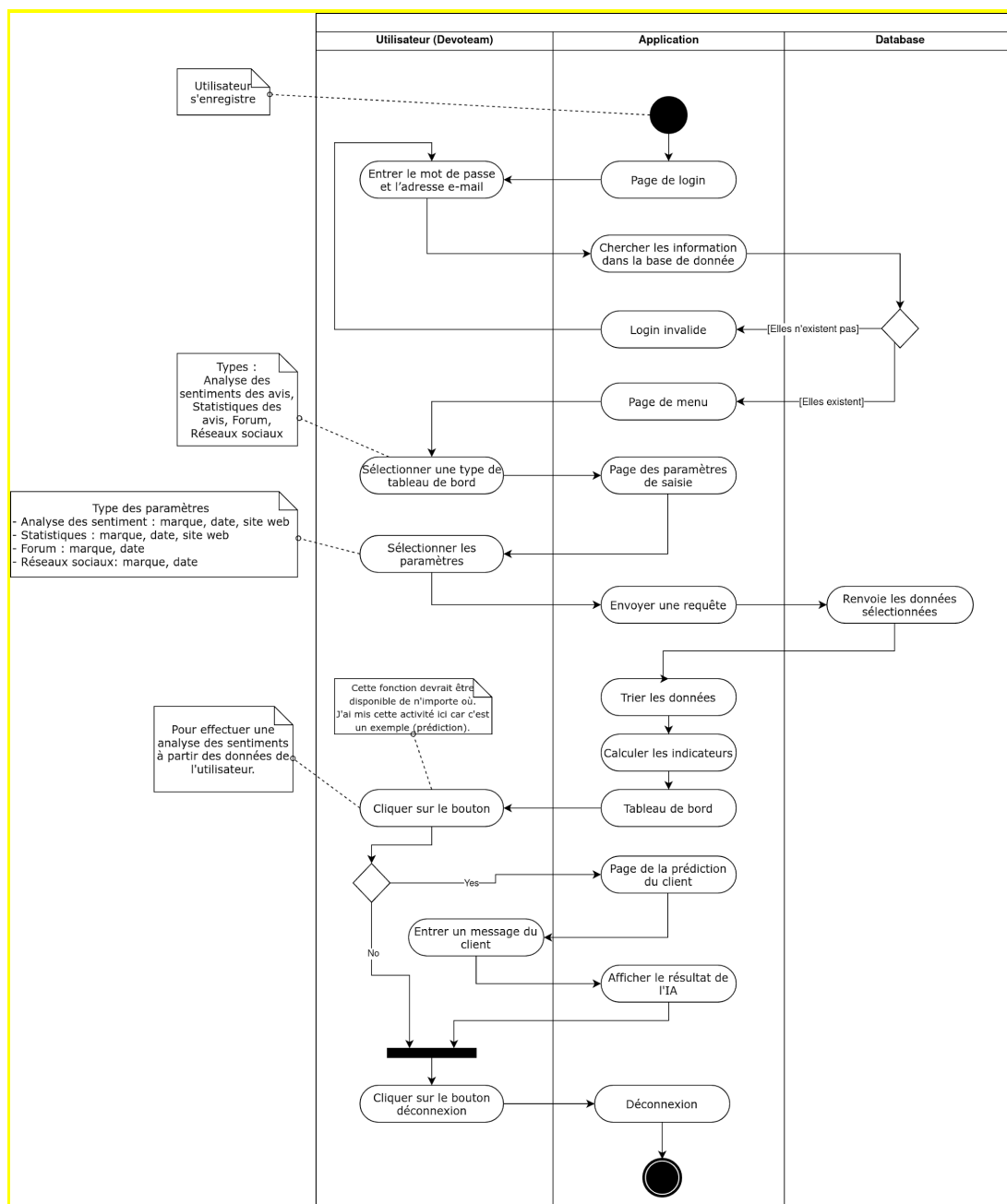
- **Visiteur** : Usager anonyme de l'application, ne disposant pas de compte ou n'étant pas connecté à son compte.
- **Utilisateur** : Usager disposant d'un compte d'utilisateur et connecté à ce dernier (authentification réussie).

Fonctionnalité	En tant que...	Je souhaite...	Afin de....
Page	Visiteur	Accéder à une page 404	Etre informé et redirigé en cas d'erreur
Compte utilisateur	Visiteur	Créer un compte d'utilisateur	Accéder à l'ensemble des fonctionnalités
Compte utilisateur	Visiteur	Me connecter à mon compte d'utilisateur	Obtenir l'autorisation d'accéder aux fonctionnalités d'utilisateur
Compte utilisateur	Utilisateur	Me déconnecter de mon compte	Fermer l'accès à mon compte d'utilisateur
Dashboard	Utilisateur	Saisir les critères	Visualiser les données
Dashboard	Utilisateur	Voir les figures des statistiques	Etudier les intentions d'achats ou de consommation d'un produit ou d'un service
Analyse de sentiment	Utilisateur	Saisir le nouveau avis	Tester l'ABSA à partir de cet avis



## Diagramme comportemental

Sur la base des User stories, nous définissons pour l'application les diagrammes comportementaux afin de spécifier, visualiser et documenter les procédés dynamiques d'un système. Le processus se compose de trois parties (utilisateur, application et base de données) et le processus envisagé est illustré dans le diagramme ci-dessous :



**Figure 1. Diagramme d'activités.**

## Conceptions structurelles

### Diagramme des composants

Le diagramme des composants présente chaque technologie employée pour développer le système. Les composants de l'application communiquent par l'intermédiaire du framework **Flask**. Par conséquent, la technologie Python prend en charge tous les **modèles** et **contrôleurs**.

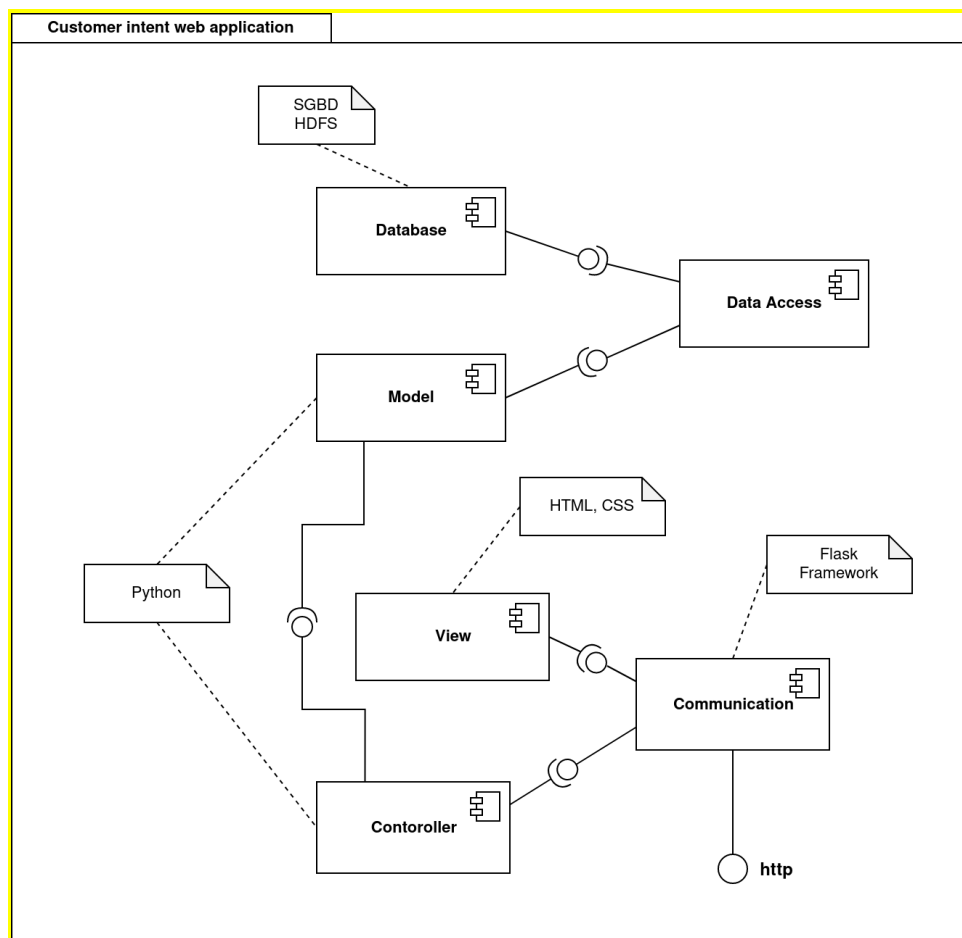


Figure 2. Diagramme des composants. L'application est conçue selon le Design Pattern MVC.

## Modélisation des données

### Diagramme Modèle Conceptuel de Données (MCD)

Sur la base des User Stories, nous définissons les entités suivantes :

- **Website** : Page web. On lui attribue une page web.
- **Company** : Entreprise. On lui attribue une entreprise d'assurance .
- **Review** : Avis. On lui attribue un avis, une date publiée, une étoile, un sujet, et un sentiment.
- **Social media** : Réseaux sociaux. On lui attribue une publication, une date publiée, un nombre de likes et nombre de partages.
- **Comment** : Commentaire. On lui attribue un commentaire, une date commentée et un sentiment.
- **Forum** : Forum. On lui attribue un titre du forum.
- **Discussion** : Discussion. On lui attribue une discussion, une date discutée.

Cette définition de la structure de données se traduit par le diagramme MCD ci-dessous :

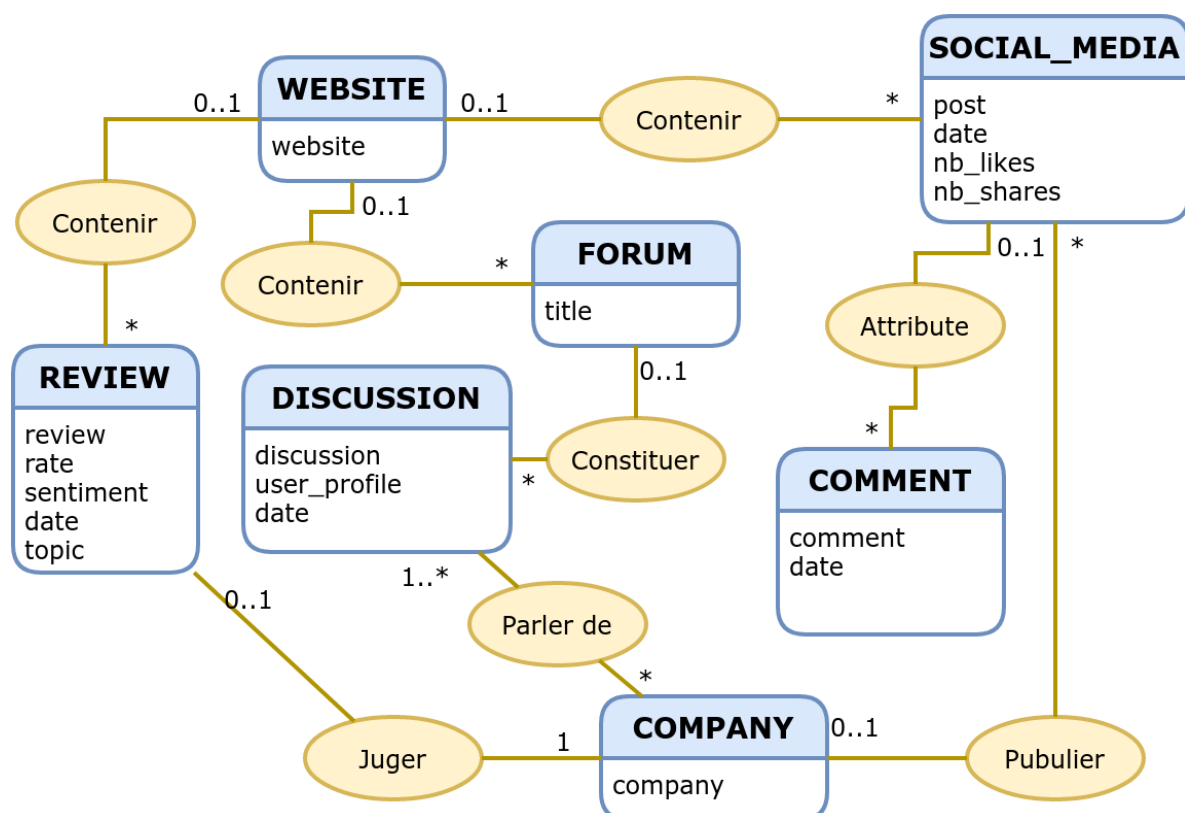


Figure 3. Diagramme MCD.

Les cardinalités des différentes relations sont déduites des hypothèses suivantes :

- Un site web peut contenir plusieurs sujets de forum.
- Un sujet de forum est contenu dans un seul site.
- Un sujet de forum peut constituer aucun, un ou plusieurs discussions.
- Dans une discussion, on parle de aucun, d'un ou de plusieurs entreprises et une entreprise est parlé dans un ou plusieurs discussions. Cela justifie la relation "**Many To Many**" entre les entités Company et Discussion.

## Diagramme Modèle Physique des Données (MPD)

Le diagramme MPD est rédigé à partir du diagramme MCD en respectant les trois règles suivantes :

- Toute entité du diagramme MCD devient une table à part entière du diagramme MPD.
- Dans le cas d'une relation où l'une des cardinalités maximales vaut 1 (relation de type **"One To Many"**), une clé étrangère sera placée du côté de l'entité portant la cardinalité maximale valant 1.
- Dans le cas d'une relation où les cardinalités maximales valent toutes deux \* (relation **"Many To Many"**), une nouvelle table d'association est créée.

On ajoute par ailleurs un attribut "id" à chaque entité permettant d'identifier chaque enregistrement de manière unique dans la base de données.

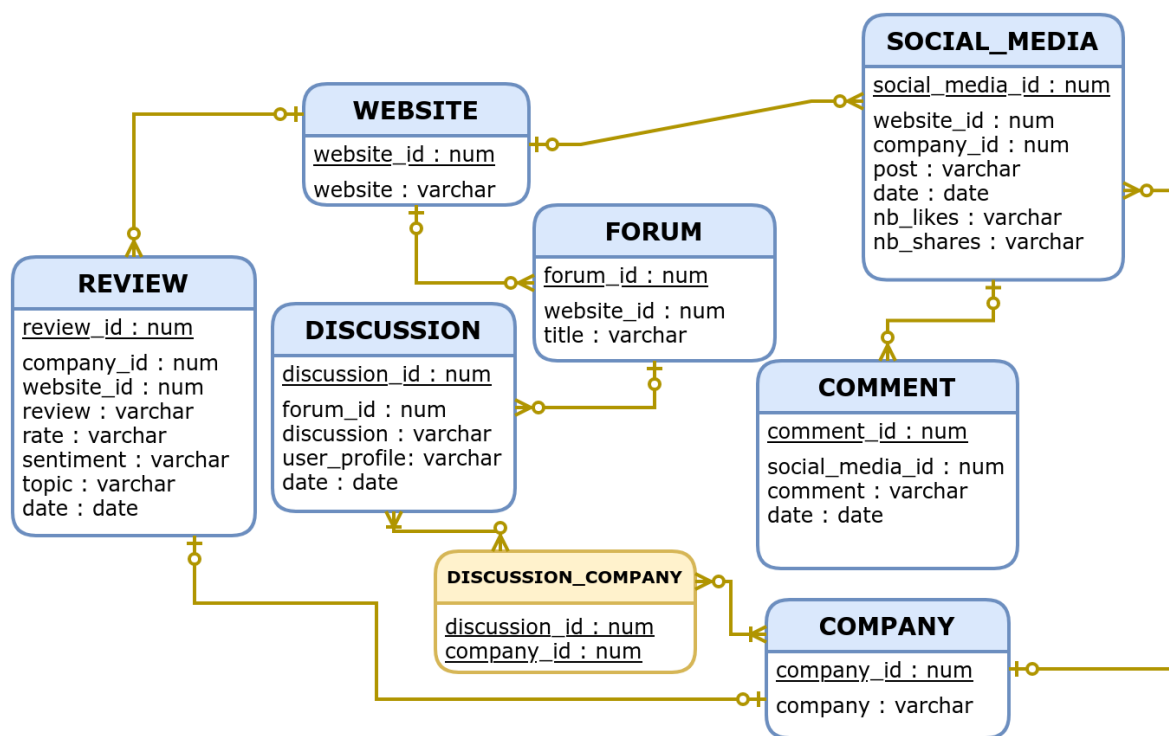


Figure 4. Diagramme MPD

## Définition des routes Front et Back

Les routes suivantes ont été définies durant la phase de conception.

### Routes Back

	GET	POST	PATCH	DELETE
<b>/login</b>		X		
Description : Doit permettre l'envoi d'un formulaire de connexion permettant à l'utilisateur de se connecter à son profil d'utilisateur. Renvoie le token en cas de succès.				
<b>/signup</b>		X		
Description : Doit permettre à un visiteur d'envoyer un formulaire de création de compte utilisateur (username, adresse mail, password).				
<b>/logout</b>		X		
Description : Doit permettre à l'utilisateur connecté de se déconnecter.				
<b>/index</b>		X		
Description : Doit permettre l'envoi d'une liste des types du tableau de bord.				
<b>/review</b>		X	X	
Description : Doit permettre l'envoi des affichages des statistiques de l'avis.				
<b>/sentiment</b>		X	X	
Description : Doit permettre l'envoi des affichages des sentiments de l'avis.				
<b>/forum</b>		X	X	
Description : Doit permettre l'envoi des affichages des statistiques du forum.				
<b>/socialmedia</b>		X	X	
Description : Doit permettre l'envoi des affichages des statistiques et des sentiments des réseaux sociaux.				
<b>/prediction</b>		X	X	
Description : Doit permettre de prédire un sentiment et un sujet à partir d'un text que l'utilisateur met				

### Routes Front

URL	Page
<b>/login</b>	Page de login
Description : l'url qui dirige vers la page de login	
<b>/signup</b>	Page d'inscription
Description : l'url qui dirige vers la page d'inscription	
<b>/error</b>	404 / not found
Description : l'url qui dirige vers la page 404 / not found	
<b>/:username/review?website=":input"&amp;company=":input"&amp;date=":input"</b>	Tableau de bord des statistiques d'avis
Description : l'url qui dirige	
<b>/:username/sentiment?website=":input"&amp;company=":input"&amp;date=":input"</b>	Tableau de bord de l'analyse des sentiments
Description :	
<b>/:username/forum?company=":input"&amp;date=":input"</b>	Tableau de bord du forum
Description :	
<b>/:username/socialmedia?company=":input"&amp;date=":input"</b>	Tableau de bord des réseaux sociaux
Description :	
<b>/:username/prediction?text=":input"</b>	Page de la prédiction des sentiments
Description :	

## Choix techniques

### Choix des technologies

Nous choisissons les technologies par étape sur Tableau 1.

*Tableau 1. Choix techniques*

	Étape	Méthode	Technologies
1	Collecte des données	Crawling	Selenium, BeautifulSoup, Tweepy, Facebook-scraper
2	Stockage des données	ETL	Python, MongoDB
3	Exploration, visualisation des données	Data Engineering	Pandas, Matplotlib, Seaborn, Numpy, Plotly
4	Nettoyage des données	Pré-processing	Pandas, Numpy, Spacy, NLTK, Gensim
5	Construction du modèle prédictif	Data Science	Textblob, Scikit-Learn, Pytorch, Tensorflow
6	Optimisation de la performance	Hyperparameter optimization (tuning), feature engineering, data augmentation	Textblob, Scikit-Learn, Pytorch, Tensorflow, Keras
7	Stockage de la donnée nettoyée	SGBD *1	SnowSQL
8	Développement de l'application web	API restful	Flask, HTML, CSS, JS
9	Automatisation	Déploiement	Docker, Azure, serveur
10	Surveillance de l'application	Monitoring et logging	logging, evidently
11	Maintenance de l'application	Mise à jour	Python

\*1 Système de Gestion de Base de Données

## Choix de l'intelligence artificielle

Nous proposons l'ABSA afin d'étudier les intentions d'achats ou de consommation d'un produit ou d'un service. Cette analyse consiste 2 objectives :

- **Analyse des sentiments polarités**
- **Extraction des sujets**

Le tableau 2 énumère les algorithmes, les librairies, les méthodes d'entraînement et d'évaluation pour chaque objectif.

*Tableau 2, Description des choix de l'IA*

Objectif	Algorithme	Libraries	Méthode d'entraînement	Méthode d'évaluation
Analyse des sentiments	Rule-based	Textblob, NLTK	None (règles prédéfinies)	None
	Classification	Scikit-learn	3 classes selon les 1, 3 et 5 étoiles	precision, f1, recall, accuracy
Extraction des sujets	Topic modeling	Gensim, NLTK	Clustering	mesure de topic coherence
	Classification	Scikit-learn	Multi-classification (labels devant prédéfinir)	precision, f1, recall, accuracy



## Réponse finale apportée ; ce qui a été réalisé

### Sources à récolter

Nous avons obtenu des données d'au moins deux sites de trois types différents (**avis, forums et réseaux sociaux**). Les types de ressources, les noms de sites et les liens sont énumérés ci-dessous :

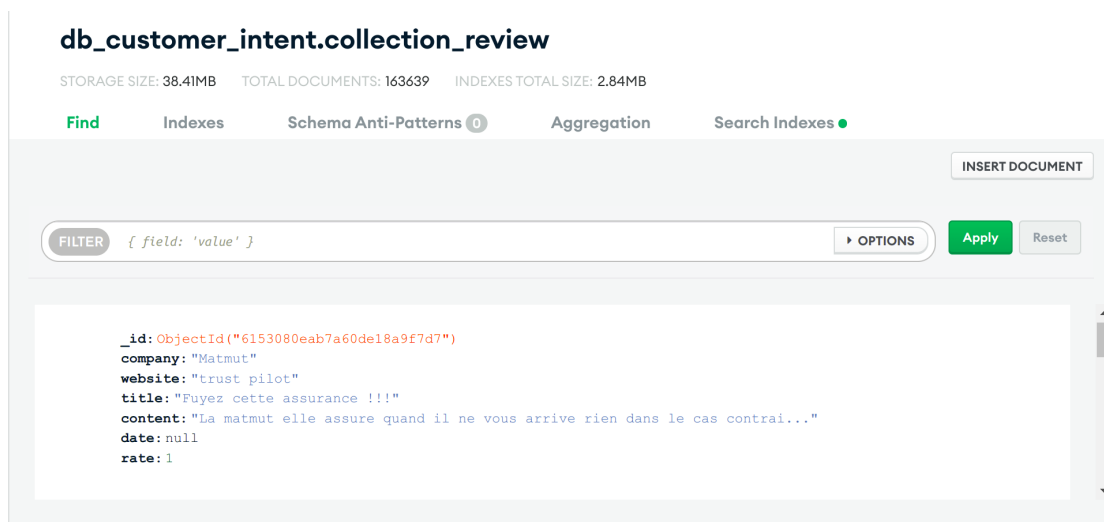
**Tableau 3, Sources de données**

Type	Source de données	Lien
Review	Opinion assurance	
	TrustPilot	
	Google Maps	
Forum	60 millions consommateur	
	QueChoisir	
Réseaux sociaux	Twitter	
	Facebook	

### SGBD

### MongoDB

Le jeu de données brutes a été stocké dans NoSQL avant la transformation des données, car il pourrait potentiellement être utilisé à l'avenir pour différents projets (ex : différents objectifs, différentes méthodes de transformation etc). Avant de le stocker, l'information d'utilisateurs a été anonymisée. Nous l'avons stocké dans MongoDB Atlas.



The screenshot shows the MongoDB Atlas web interface. At the top, the collection name is **db\_customer\_intent.collection\_review**. Below it, statistics are displayed: STORAGE SIZE: 38.41MB, TOTAL DOCUMENTS: 163639, INDEXES TOTAL SIZE: 2.84MB. There are tabs for Find, Indexes, Schema Anti-Patterns, Aggregation, and Search Indexes. The 'Find' tab is active. A filter bar shows a filter: { field: 'value' }. Below the filter bar, a document is displayed in a code editor:

```
{
  "_id": ObjectId("6153080eab7a60de18a9f7d7"),
  "company": "Matmut",
  "website": "trust pilot",
  "title": "Fuyez cette assurance !!!",
  "content": "La matmut elle assure quand il ne vous arrive rien dans le cas contrai...",
  "date": null,
  "rate": 1
}
```

Figure 5, Exemple de données insérées dans une collection de MongoDB Atlas.

```
def insert_data_for_reviews(input_data:list):
    '''Insert new data in review collection'''

    collection = DB().review_collection()

    for row_data in input_data:
        # loop for data

        try:
            # if
            cursor = collection.find({'website': { "$eq" : row_data['website']},
                                     'company': { "$eq" : row_data['company']},
                                     'content': { "$eq" : row_data['content']}})

            if cursor.count() == 0:
                print(row_data,': does not exist in our database')
                # insert row_data in database
                collection.insert_one(row_data)
                print('add data')

        except Exception as e:
            print(e)
```

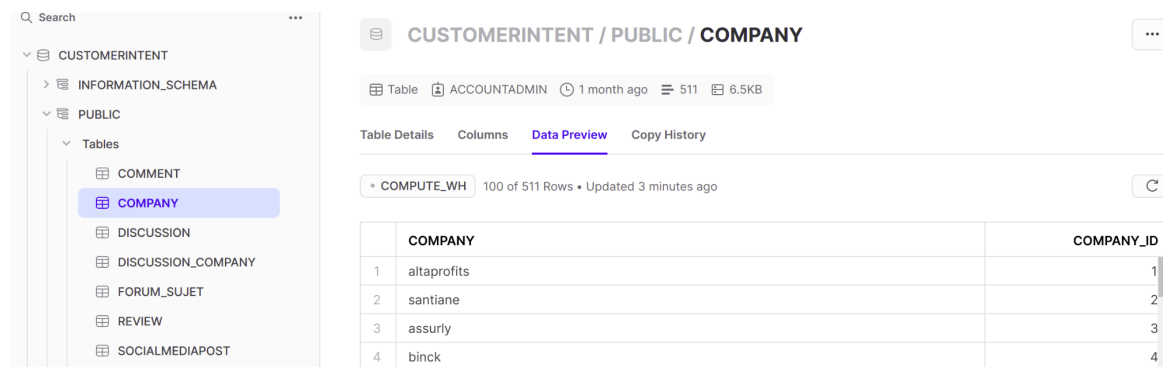
Figure 6, Exemple de code pour l'insertion de données dans une base de données. Insérer de nouvelles données dans une base de données, sauf s'il s'agit de la même entreprise, du même site ou du même contenu.

## SnowflakeSQL

Selon le diagramme MPD conçu, les données MongoDB ont été converties et insérées dans Snowflake SQL. Les commandes de création sont les suivantes :

```
CREATE TABLE IF NOT EXISTS REVIEW (
    "review_id" number PRIMARY KEY,
    "review" varchar,
    "date" date,
    "rate" number,
    "sentiment" number,
    "topic" varchar,
    "website_id" number,
    "company_id" number)
```

Figure 7, Exemple de commande de création de table. La table review a une clé primaire, review\_id, et contient les id des tables website et company, qui sont des relations "One to many".



	COMPANY	COMPANY_ID
1	altaprofits	1
2	santiane	2
3	assurly	3
4	binck	4

Figure 8, Exemple de données insérées à la BDD Snowflake (Interface).

## Solution de l'IA

Chaque développeur devait en développer un. On décrit l'IA d'extraction des sujets qui a été développée.

### Extraction des sujets

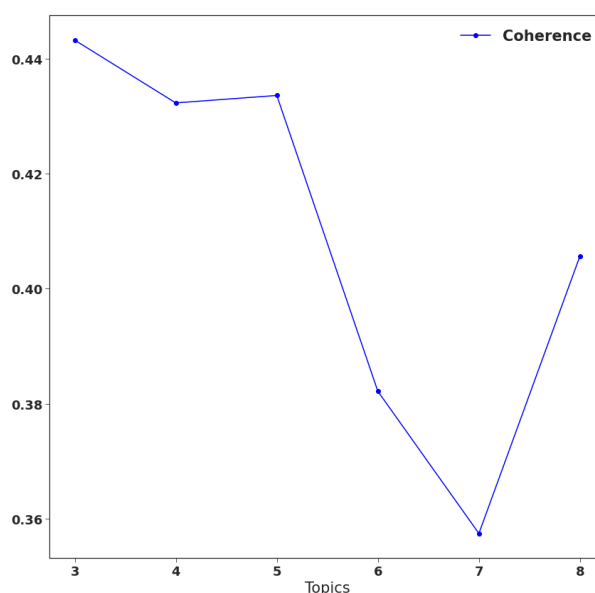
La différence entre la classification et le clustering réside dans le fait que les étiquettes sont attachées manuellement à toutes les données ou que les étiquettes sont définies après des regroupements. Nous avons finalement adopté le topic modeling (clustering). Les raisons sont données ci-dessous :

- Lorsqu'une tendance topique change soudainement (par exemple, les coronavirus), le modèle de classification doit être ré-étiqueté, et alors le clustering permet d'être reconstruit rapidement.
- La classification est une procédure longue dans laquelle on analyse les données à l'avance, par exemple en dressant la liste des sujets.

L'algorithme le plus connu en matière de topic modeling, **Latent Dirichlet Allocation** (LDA), a été utilisé pour regrouper le jeu des données d'entraînement. Les détails des données et des hyperparamètres sont donnés ci-dessous :

<b>Jeu de données</b>	Données des avis (AXA) : 2745 rows
<b>Hyper parametres</b>	Nombre des topics (k), alpha, beta

La **mesure de Topic Coherence** ont été utilisées pour optimiser les hyperparamètres dans le LDA.



*Figure 9, Résultat de la mesure de Topic Coherence. Lorsque le nombre des topics est 3, la mesure a obtenu le meilleur score.*

Comme la mesure maximum de topic cohérence est atteint lorsque **k = 3**, **alpha = asymmetric** et **beta = 0.91**, ce résultat a été visualisé et les noms de chaque cluster ont été définis.

- **Tarif** : tres, augment, bon, agent, garantir ...
- **Automobile** : véhicule, voiture, expert, chargé, accident ...
- **Service** : envoie, résilier, appel, recevoir, mail ...

## Application

A partir des spécifications techniques, nous avons développé 3 tableaux de bord, une page qui permet de tester l'IA.

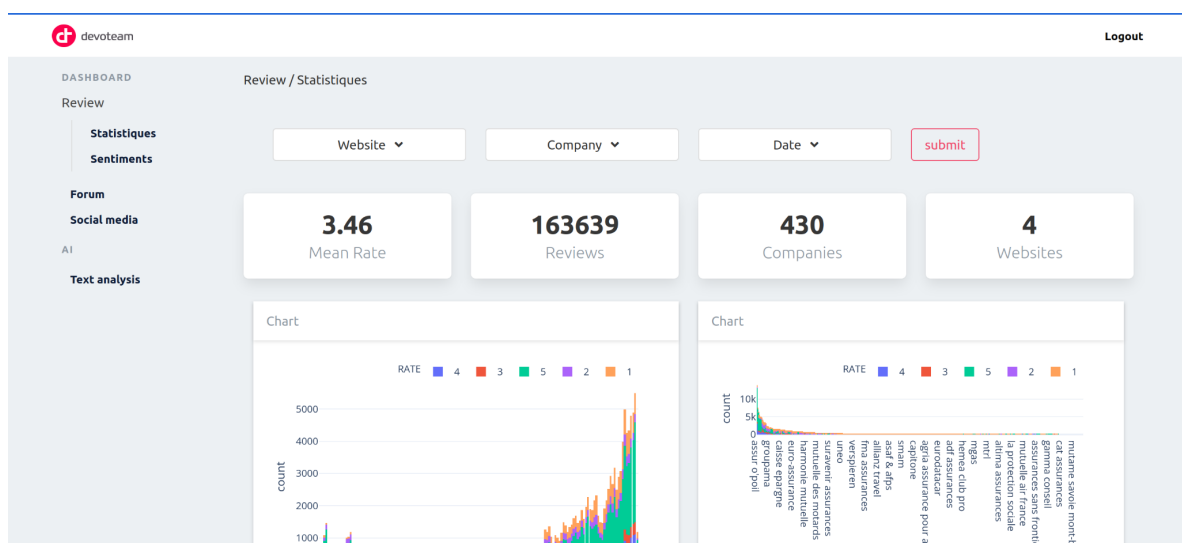


Figure 10, Page Tableau de bord (review).

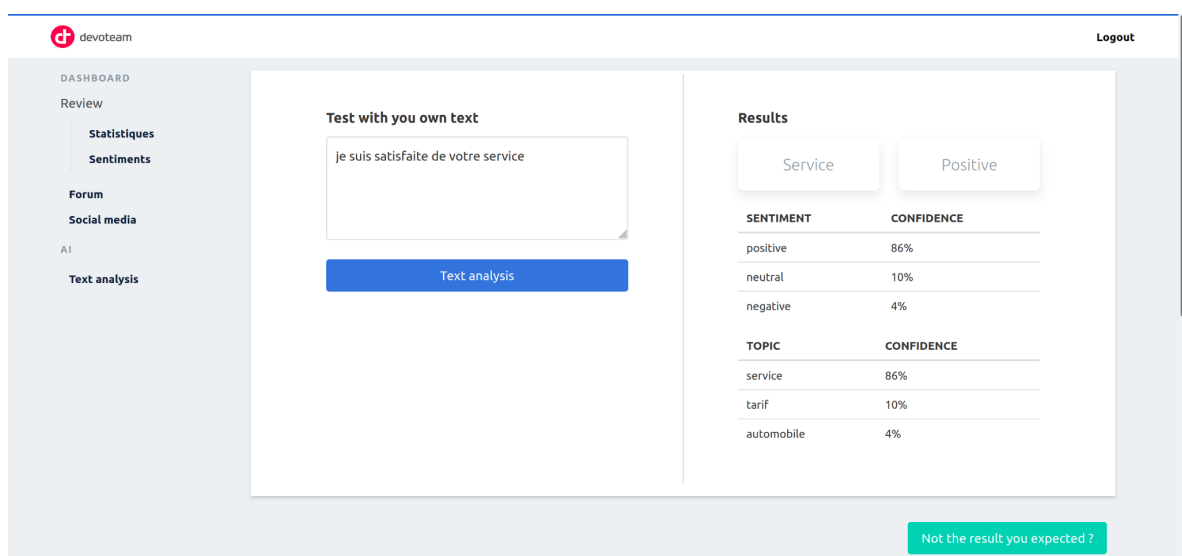


Figure 11, Page d'analyse de texte.

## Monitoring

Nous avons ajouté une fonctionnalité sur la page d'analyse de texte qui permet de maintenir les performances des prédictions de l'IA. Si les prédictions de l'IA sont incorrectes, l'utilisateur peut cliquer sur le bouton situé dans le coin inférieur gauche et saisir l'étiquette que l'utilisateur analyse.

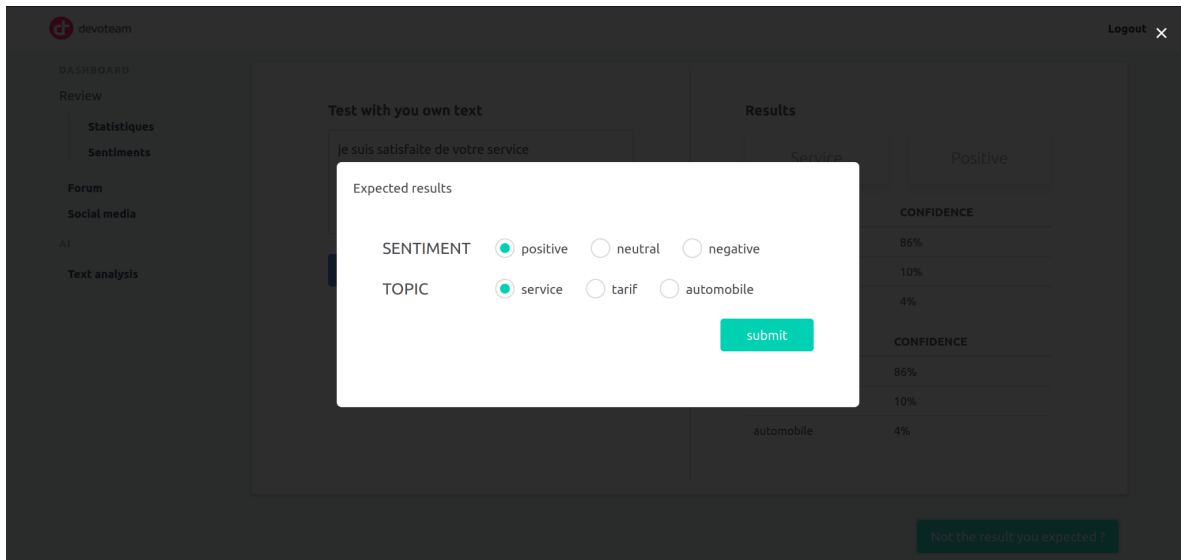


Figure 12, Page de saisie de l'utilisateur pour monitorer la performance de l'IA.

## Mise en oeuvre du projet

### Organisation technique

#### Domaine d'hébergement

L'API sera déployée sur le **serveur de Devoteam** gérés par la structure Devolab (R&D chez Devoteam) mais il sera également possible de la déployer sur le cloud avec une compatibilité Azure, AWS, Google Cloud, etc.

#### Exigence de programmation

Le principale langage de programmation utilisé dans l'API est **Python**, il est employé dans le crawling, la data engineering, le pré-processing, l'implémentation de l'IA, et la programmation de l'application. **HTML** et **CSS** seront utilisés pour la partie front-end de l'application.

#### Accessibilité

##### Compatibilité navigateurs

L'application web sera compatible avec les navigateurs suivants :

- **Google Chrome**
- **Mozilla Firefox**
- **Microsoft Edge**

##### Types d'appareil

L'application web sera conçu de manière dite "responsive" pour qu'il assure une navigation optimale sur tous types d'appareils :

- **Ordinateur portables**
- **Ordinateur de bureau**

#### Sécurité

Nous décidons pour l'authentification d'utiliser la technologie suivante

- **Bcrypt** : Librairie mettant à disposition des fonctions permettant de hacher de l'information (principalement des mots de passe), valider un mot de passe en le comparant à un hash. Cette dépendance est utilisée pour ne pas stocker de mots de passe non-chiffrés dans la base de données, pratique à la fois extrêmement dangereuse et illégale.

#### Logging et monitoring

Nous intégrerons une fonction qui permet de surveiller les erreurs et les bugs afin de maintenir l'application. Les messages de logging peuvent fournir des informations précieuses pour aider à

déterminer la cause des problèmes de performance. De plus, les développeurs analyseront l'application en utilisant la donnée de logging afin d'améliorer l'application.

La fonction permet d'envoyer un message au administrateur sous forme d'alarme si une certaine valeur de précision est dépassée afin d'éviter une dégradation des performances de l'intelligence artificielle.

## Environnement de développement tout au long de la production

L'environnement de développement est sous **Windows** et **Linux**. Le produit fonctionne sur le système d'exploitation utilisé par Devoteam, Windows et pourra être hébergé par des serveurs web ou serveurs d'application.

## Gestion de projet

La **méthode agile (Kanban)** est utilisée pour le développement du projet. Nous avons une **réunion hebdomadaire** avec le chef du projet et une réunion d'avancement avec l'équipe de développement le lundi matin. En outre, nous utilisons les outils suivants :

- **Citilab** : Partage du code, accès au Kanban et aux tâches à réaliser dans l'équipe de développement.
- **Google drive** : Stockage et partage de fichiers, de compte rendu d'avancement, de données, rédaction de documents via l'éditeur **Google Doc** et **Spreadsheet**.
- **Google chat** : Outil principal de communication entre tous les membres du projet.
- **Google meet** : Outil principal de la réunion.

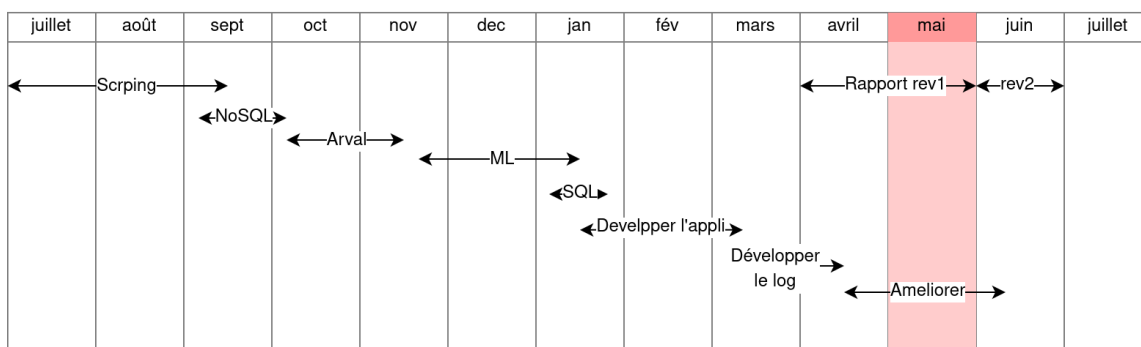


Figure 13. Planning prévisionnel

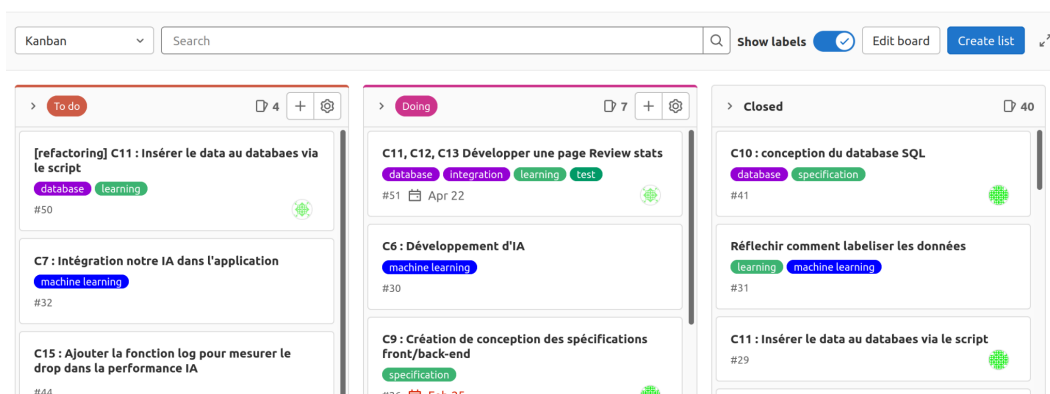


Figure 14. Exemple de notre tableau Kanban

## Composition de l'équipe

L'équipe chargée du projet Customer Intent est constituée de 2 développeurs, 1 tuteur et 1 accompagnement :

- Zied BEN OTHMANE : Accompagnement de l'équipe de développement ( - 03/2022)
- Hamadi CAMARA : Accompagnement de l'équipe de développement (04/2022 -)
- Anthony VILLOT : Tuteur
- Chouaieb NEMRI : Tuteur (- 09/ 2021)
- Oihana PASSICOT : Développeuse junior en IA
- Mirai IIDA : Développeuse junior en IA

## Retours d'expérience sur les outils, techniques et compétences à l'œuvre

- **Crawling** : Il faut plus d'une journée pour extraire (exécuter le code) les données de 7 sites. Cela est dû au grand nombre de pages et au fait que le crawling est exécuté en série. Il est également difficile d'automatiser le crawling pour les sites dont les balises d'HTML changent fréquemment, par exemple Google Review.
- **ETL** : J'ai fait la transformation en Python, mais j'ai codé manuellement les étapes de génération de la clé primaire et de rattachement de la clé d'une autre table. Je ne pensais pas que c'était très pratique de faire la transformation en Python.
- **IA** : J'ai analysé les données d'AXA avec le modèle LDA et j'ai trouvé très utile de pouvoir les classer en groupes de bons sujets.
- **Application** : Trois langages (CSS, JS et HTML) ont été utilisés pour le Frontend, mais en raison des différentes syntaxes, il était difficile pour le développeur en IA.

## Bilan du projet et les améliorations envisageables

Le projet couvrait de nombreuses techniques et m'a beaucoup aidé à améliorer mes compétences. Les améliorations possibles sont les suivantes :

- Automatisation de toutes les étapes (entretien régulier requis).
- Clustering avec toutes les données des compagnies d'assurance, créant un modèle plus global.
- Création de modèles de classification avec des étiquettes entrées par l'utilisateur.
- Considérations commerciales (valoriser les indicateurs).

## Conclusion



## Annex

- [1] G.Deepak and E.Asif, Supervised Machine Learning for Aspect based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation*, pp. 319-323, 23-24 August 2014.
- [2] B.Tomas, K.Michal and S.Josef, Machine Learning Approach to Aspect-Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation*, pp.817-822, 23-24 August 2014.
- [3] B.Caroline, P.Diana and R.Claude, Hybrid Classification for Aspect-based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation*, pp.838-842, 23-24 August 2014.
- [4] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. 2007.
- [5] G.Ian et al., Generative Adversarial Nets. *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2672-2680, 10 June 2014.
- [6] M.Cade, Google Just Open Sourced TensorFlow, Its Artificial Intelligence Engine. <https://www.wired.com/2015/11/google-open-sources-its-artificial-intelligence-engine/> 11 September 2015.
- [7] S.Ilya et al., Sequence to Sequence Learning with Neural Networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 3104-3112, 2 December 2014.
- [8] D.Bahdanau, K.Cho and Y.Bengio, Neural Machine Translation by Jointly Learning to Align and Translate. Paper presented at 3rd International Conference on Learning Representations, 1 Jan 2015.
- [9] V.Ashish, S.Noam et al., Attention Is All You Need. *Proceedings of the 31st Conference on Neural Information Processing System*, 2017.
- [10] T.Iulia, C.Ming-Wei et al., WELL-READ STUDENTS LEARN BETTER: ON THE IMPORTANCE OF PRE-TRAINING COMPACT MODELS. 25 September 2019.
- [11] M.Tomas et al., Efficient Estimation of Word Representations in Vector Space. *Proceedings of the 1st International Conference on Learning Representations*. 7 September 2013.
- [12] X.Hu , L.Bing et al., Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 592-598, 15 - 20 July 2018.
- [13] O.Shinhyeok et al., Deep Context- and Relation-Aware Learning for Aspect-based Sentiment Analysis. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 7 Jun 2021.
- [14] J.Xincheng et al., Joint Multi-modal Aspect-Sentiment Analysis with Auxiliary Cross-modal Relation Detection. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4395-4405, November 2021.