

Ministère de l'Enseignement Supérieur Et de la  
Recherche Scientifique  
Université de Carthage  
INSAT Institut National des Sciences Appliquées et de  
Technologie



## Rapport du Projet de Fin d'Année

Département de Génie Informatique et Mathématiques  
Filière : Génie logiciel

---

# Mise en place d'un système de recommandation hôtelière basé sur l'analyse sentimentale

---

### Réalisé par :

Eya RAOUNE  
Fatma HAMADA  
Wiem BEN MLOUKA

### Encadré par :

Mme Sana HAMDI

Année universitaire 2023-2024

## Dédicace

*Nous dédions ce travail, tout d'abord, à notre encadrante académique , Mme Sana HAMDI, qui nous a apporté un soutien et des conseils inestimables tout au long de ce projet. Son expertise et sa passion pour l'enseignement ont illuminé notre chemin vers la connaissance.*

*Ce projet est également le fruit des efforts collectifs de tous les membres de notre équipe, qui ont travaillé sans relâche pour atteindre l'objectif souhaité. Nous tenons également à présenter nos chaleureux remerciements à toutes les personnes qui ont contribué, de près ou de loin, à son bon déroulement.*

*Avec un grand respect, nous n'oublions pas d'exprimer notre profonde gratitude à tous les enseignants de l'INSAT, qui nous ont enseigné les bases de l'informatique au cours de ces quatre années et pour les connaissances qu'ils nous ont transmises.*

# Résumé

Ce rapport présente une approche complète pour la mise en place d'un système de recommandation hôtelière hybride basé sur l'analyse sentimentale des avis sur les hôtels et les profils des utilisateurs, en exploitant leur similarité.

Développé dans le cadre du Projet de Fin d'Année à l'INSAT pour l'année universitaire 2023-2024, le projet exploite des techniques avancées de traitement du langage naturel (NLP) et des modèles de réseaux de neurones pour améliorer la précision de l'analyse sentimentale et proposer des recommandations d'hôtels personnalisées et pertinentes.

**Mots clés :**Hôtels, Recommandation hybride, Avis, Profil utilisateur, Similarité Cosinus, Aspects, Analyse sentimentale, Réseaux de neurones, BiLSTM, Word2Vec, TF-IDF, GloVe, Clustering CAH, NLP

# Table des matières

<b>Introduction générale</b>	<b>1</b>
<b>I Business Understanding</b>	<b>3</b>
Introduction . . . . .	3
I.1 Contexte et objectifs du Projet . . . . .	3
I.1.1 Contexte du Projet . . . . .	3
I.1.2 Problématique . . . . .	3
I.1.3 Objectifs du Projet . . . . .	4
I.2 Système de Recommandation . . . . .	4
I.2.1 Concepts de base . . . . .	4
I.2.2 Classification des systèmes de recommandation . . . . .	5
I.2.3 Comparaison des systèmes de recommandation . . . . .	8
I.3 Analyse Sentimentale . . . . .	9
I.3.1 Approche basée sur les lexiques . . . . .	9
I.3.2 Approche par Apprentissage Automatique . . . . .	10
I.4 Architecture Globale de la solution proposée . . . . .	10
I.4.1 Analyse Sentimentale des Avis . . . . .	11
I.4.2 Extraction des profils des Utilisateurs . . . . .	11
I.4.3 Mise en Correspondance et Recommandation . . . . .	11
I.5 Méthodologie du Projet : CRISP-DM . . . . .	12
Conclusion . . . . .	13
<b>II Compréhension et préparation des données</b>	<b>14</b>
Introduction . . . . .	14
II.1 Data Understanding . . . . .	14
II.1.1 Sources de Données : Profils et Intérêts des Utilisateurs . . . . .	14
II.1.2 Sources de Données : Avis sur les Hôtels . . . . .	14

II.2 Data Preparation . . . . .	15
II.2.1 Sources de Données : Profils et Intérêts des Utilisateurs . . . . .	15
II.2.2 Sources de Données : Avis sur les Hôtels . . . . .	16
Conclusion . . . . .	18
<b>III Modélisation et évaluation</b>	<b>19</b>
Introduction . . . . .	19
III.1 Architecture détaillée de la solution proposée . . . . .	19
III.2 Clustering des similarités entre utilisateurs en fonction de leurs intérêts et légendes .	20
III.2.1 Représentation Vectorielle . . . . .	20
III.2.2 Similarité cosinus . . . . .	20
III.3 Analyse sentimentale des avis sur les hôtels . . . . .	23
III.3.1 VADER . . . . .	23
III.3.2 Utilisation des Réseaux de Neurones . . . . .	25
III.3.3 Prédictions . . . . .	29
III.3.4 Extraction des aspects des utilisateurs . . . . .	30
III.4 Matching des utilisateurs et recommandation . . . . .	31
III.4.1 Matching des utilisateurs . . . . .	31
III.4.2 Recommandation . . . . .	34
Conclusion . . . . .	34
<b>IV Déploiement</b>	<b>35</b>
Introduction . . . . .	35
IV.1 Technologies utilisés . . . . .	35
IV.1.1 Framework Django Pour Le Développement . . . . .	35
IV.1.2 Intégration de Flask pour les Requêtes de Recommandation . . . . .	35
IV.2 Visualisation . . . . .	36
IV.2.1 Interface Principale des Hôtels . . . . .	36
IV.2.2 Authentification des Utilisateurs . . . . .	36

IV.2.3 Page de Recommandations Personnalisées . . . . .	37
IV.2.4 Évaluation des résultats . . . . .	38
Conclusion . . . . .	38
<b>Conclusion et perspectives</b>	<b>39</b>
<b>Bibliographie</b>	<b>41</b>

## Table des figures

1	Notation par étoiles . . . . .	1
2	Notation par classification des avis en positive , negative . . . . .	2
3	Filtrage basé sur le contenu [1] . . . . .	5
4	Filtrage collaboratif [1] . . . . .	6
5	Conception d'hybridation monolithique . . . . .	7
6	Approches de l'analyse sentimentales . . . . .	9
7	Architecture globale du système de recommandation . . . . .	11
8	Processus CRISP-DM . . . . .	12
9	Préparation des intérêts . . . . .	16
10	Préférences globales des utilisateurs . . . . .	16
11	Avis avant préparation . . . . .	17
12	Nombre de valeurs manquantes par colonne essentielle . . . . .	17
13	Architecture détaillée de la solution proposée . . . . .	19
14	Score de Silhouette . . . . .	21
15	Répartition du nombre d'utilisateurs par cluster . . . . .	21
16	Matrice de similiarité pour les utilisateurs du cluster 5 . . . . .	22
17	Les utilisateurs appartenant au cluster 5 . . . . .	22
18	Répartition des avis positifs et négatifs avec VADER . . . . .	23
19	Matrice de Confusion - VADER . . . . .	24
20	Rapport de Classification - VADER . . . . .	24
21	Évolution du modèle biLSTM avec overfitting . . . . .	27
22	Évolution du modèle BiLSTM + GloVE . . . . .	28
23	Matrice de Confusion du modèle . . . . .	29
24	Affectation des scores de prédictions . . . . .	29
25	Prédictions sur des critiques inconnues . . . . .	30
26	Échantillon d'extraction des aspects par utilisateurs . . . . .	31

27	Agrégation des intérêts de chaque cluster . . . . .	32
28	Expansion des synonymes des aspects . . . . .	33
29	Expansion des synonymes des intérêts . . . . .	33
30	Fusion des DataFrames et Agrégation des données par cluster . . . . .	34
31	Interface Principale des Hôtels . . . . .	36
32	Interface Authentification des Utilisateurs . . . . .	36
33	Page de Recommandations Personnalisée . . . . .	37

## Liste des tableaux

1	Comparaison des systèmes de recommandation . . . . .	8
---	--	---

# Introduction générale

L'avènement des technologies numériques a profondément transformé la manière dont nous planifions nos voyages. En 2015, 71 % des Français ont utilisé Internet pour rechercher des informations, planifier et réserver leurs séjours.

La diversité et l'abondance d'informations dématérialisées permettent aux voyageurs de mieux préparer leurs séjours. Les “infomédiaires”, ces sites web qui agrègent les avis et les informations des voyageurs sur les établissements et les activités touristiques, jouent un rôle crucial dans le choix des touristes. Ces dernières années, de nombreux sites ont intégré des fonctionnalités autrefois réservées aux réseaux sociaux : les utilisateurs peuvent désormais renseigner un profil détaillé, incluant un “profil voyageur” et sont encouragés à contribuer en partageant des avis, des recommandations, des photos ou d'autres contenus liés à leurs expériences de voyage.

A commencer par les plate-formes de réservation hôtelière qui utilisaient principalement des systèmes de notation par étoiles pour évaluer les établissements, offrant ainsi une évaluation simplifiée de la qualité globale.

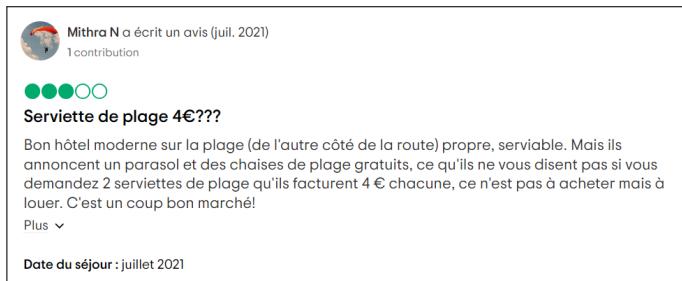


FIGURE 1 – Notation par étoiles

Cependant, ces systèmes ont rapidement révélé leurs limites. Les utilisateurs pouvaient attribuer des notes faibles à des hôtels qu'ils avaient appréciés, ou inversement, créant ainsi une distorsion dans les évaluations.

Pour remédier à ces problèmes, les plateformes de recommandation ont introduit la classification des avis en opinions positives, négatives et neutres. Cela permettait aux voyageurs d'exprimer plus en détail leurs sentiments et leurs expériences. Cependant, cette approche s'est également heurtée à des défis majeurs. Les opinions des voyageurs sont souvent complexes et nuancées : un utilisateur peut adorer les chambres d'un hôtel mais critiquer sévèrement la qualité de la restauration.

	<b>Veronique</b>	Commentaire envoyé le 20 mars 2024
	 France	<b>Très bien</b>
	<a href="#">Chambre Lits Jumeaux</a>	 Propreté, très bon petit déjeuner, proximité transports en commun
	1 nuit · mars 2024	 Chambre très petite mais fonctionnelle La salle du petit déjeuner est au sous sol, il n'y a donc pas de fenêtre
	Famille	

FIGURE 2 – Notation par classification des avis en positive , negative

Face à ces difficultés, les systèmes de recommandation hôtelière ont évolué vers une approche dite "aspect-based" (basée sur les aspects). Dans ce modèle, les avis des utilisateurs sont analysés en fonction de différents aspects de leur expérience, tels que les chambres, la restauration, le service, etc. Cette approche permet une compréhension plus fine des préférences des voyageurs et vise à fournir des recommandations plus précises et pertinentes.

Notre projet s'inscrit dans cette évolution en utilisant ces informations détaillées pour prédire plus efficacement la satisfaction des voyageurs. Il représente une première étape vers la création d'un assistant intelligent capable de recommander des établissements en fonction des critères spécifiques qui comptent le plus pour les utilisateurs.

# I Business Understanding

## Introduction

Durant ce premier chapitre, nous allons présenter la phase de compréhension du projet. Nous y exposerons nos objectifs et les contraintes à équilibrer pour les atteindre. Nous aborderons les différentes approches de recommandation d'hôtels, telles que le filtrage basé sur le contenu, le filtrage collaboratif et les méthodes hybrides, ainsi que l'importance de l'analyse sentimentale. Nous expliquerons aussi la méthodologie CRISP-DM qui guidera notre démarche, de la compréhension des besoins jusqu'au déploiement du système.

### I.1 Contexte et objectifs du Projet

#### I.1.1 Contexte du Projet

Dans un monde où les choix d'hôtels sont de plus en plus nombreux, il est essentiel de fournir aux utilisateurs des recommandations personnalisées pour améliorer leur expérience de réservation. Les systèmes de recommandation basés sur l'analyse sentimentale permettent de comprendre les préférences et les avis des utilisateurs pour offrir des suggestions d'hôtels plus précises et satisfaisantes. Ce projet vise à développer un système de recommandation hôtelière utilisant l'analyse des sentiments des commentaires des utilisateurs et les intérêts extraits a partir des réseaux sociaux ( Publications sur Instagram , etc).

#### I.1.2 Problématique

Dans un contexte où l'offre hôtelière est de plus en plus vaste et diversifiée , comment peut-on exploiter de manière efficace les données provenant des réseaux sociaux et des plateformes de réservation en ligne afin de créer un système de recommandation personnalisé qui améliore l'expérience . En particulier, comment prendre en compte les intérêts personnels des utilisateurs et les commentaires d'autres clients afin de suggérer des hôtels pertinents et satisfaisants ?

Cette problématique met en évidence les challenges associées à la collecte, à l'analyse et à l'utilisation de grandes quantités de données pour formuler des recommandations précises, tout en tenant compte de la variété des préférences et des opinions des utilisateurs.

### I.1.3 Objectifs du Projet

Les principaux objectifs de ce projet sont les suivants :

**1.Extraction et Analyse des Intérêts des Utilisateurs** : Recueillir et analyser les intérêts des utilisateurs à partir des légendes de leurs publications Instagram.

**2.Clustering des Utilisateurs par Intérêts** : Classer les utilisateurs en groupes basés sur la similarité de leurs intérêts.

**3.Analyse Sentimentale des Commentaires d'Hôtels** : Analyser les commentaires d'hôtels de Booking.com pour déterminer s'ils sont positifs ou négatifs.

**4.Matching des Utilisateurs** : Associer les utilisateurs de la première étape avec ceux de la deuxième étape en fonction de la similarité entre les aspects extraits des commentaires et les intérêts des utilisateurs.

**5.Recommandation d'Hôtels** : Recommander des hôtels aux utilisateurs en se basant sur les préférences des utilisateurs du même cluster.

## I.2 Système de Recommandation

### I.2.1 Concepts de base

Les systèmes de recommandation ont été définis de diverses manières. Une définition courante et générale, attribuée à Robin Burke (Burke, 2002), décrit un système de recommandation comme étant :

”capable de fournir des recommandations personnalisées ou permettant de guider l’utilisateur vers des ressources intéressantes ou utiles au sein d’un espace de données important” [Burke [2002]]

Au cœur de tout système de recommandations se trouvent des données sur les préférences exprimées par les utilisateurs pour les différents items. Cette préférence est généralement représentée par un triplet (utilisateur ; item ; note), où chaque élément a un rôle spécifique :

**1. Utilisateur** : Il s’agit de l’entité pour laquelle nous collectons des informations et à qui nous souhaitons fournir des recommandations personnalisées. Chaque utilisateur est identifié de manière unique dans le système.

**2. Item** : Il s’agit de l’objet, produit, service ou toute autre entité pour laquelle nous souhaitons faire des recommandations. Chaque item est également identifié de manière unique dans le système.

**3. Note** : La note représente la préférence exprimée par un utilisateur pour un item spécifique.

Cette note peut prendre différentes formes selon le contexte du système de recommandation. Par exemple, elle peut être une évaluation numérique (comme une étoile sur 5, une note sur 10, etc.), une indication binaire (par exemple, "positive" ou "negative"), ou même un score de probabilité.

### I.2.2 Classification des systèmes de recommandation

La classification des systèmes de recommandation, selon les travaux d'Adomavicius & Tuzhilin (2005) et Burke (2002), se divise principalement en trois types principaux : le filtrage basé sur le contenu, le filtrage collaboratif, et le filtrage hybride.

#### a. Filtrage basé sur le contenu

Le content-based analysis se base sur un ensemble de contenus sans prendre en compte les utilisateurs et identifie les similitudes entre ces contenus pour proposer des recommandations. Dans le contexte de la recommandation hôtelière, l'analyse de contenu consisterait par exemple à identifier les caractéristiques clés d'un hôtel en analysant ses attributs tels que la localisation, le type d'hébergement, les équipements disponibles, etc. Plus des hôtels partagent un grand nombre de caractéristiques similaires plus ils seront considérés comme "proches" les uns des autres. Cela permet de détecter des hôtels offrant des expériences similaires ou ayant des caractéristiques identiques et d'en déduire des recommandations pour l'utilisateur. Un avantage clé des systèmes basés contenu - absent dans les systèmes collaboratifs - est que l'utilisateur peut recevoir des recommandations même s'il est le seul utilisateur du système.

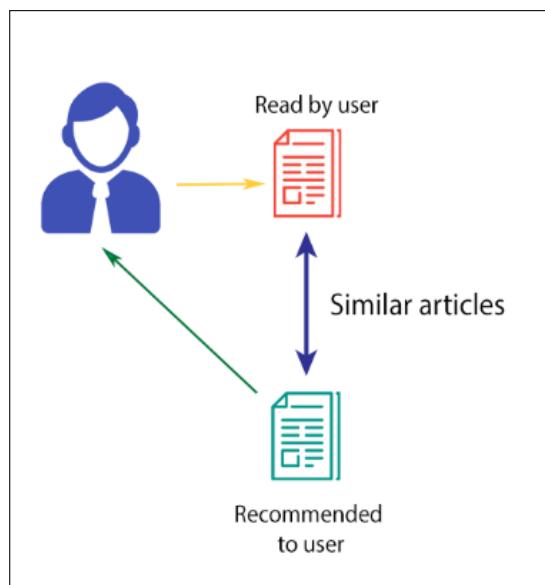


FIGURE 3 – Filtrage basé sur le contenu [1]

Le profil de l'utilisateur est exprimé sous forme d'une liste d'intérêts basée sur les mêmes

caractéristiques. La coïncidence entre les caractéristiques des éléments et le profil de l'utilisateur peut être mesurée de différentes manières :

- l'indice de Dice ou d'autres mesures de similarité
- le TF-IDF (Term Frequency-Inverse Document Frequency)
- les techniques basées sur la similarité des espaces vectoriels (les approches bayésiennes , les arbres de décision, etc.)

### b. Filtrage collaboratif

Le filtrage collaboratif repose sur l'adage : Si deux personnes ont aimé des contenus identiques par le passé, elles ont une probabilité élevée d'aimer les mêmes choses dans le futur.

Les recommandations personnalisées issues du filtrage collaboratif peuvent être calculées de diverses manières. Notamment en se basant sur le profil des lecteurs (User-based), ou en utilisant les profils de contenus (Item-based) .

Les systèmes de recommandation basés sur les items se distinguent des systèmes basés sur les utilisateurs principalement par la manière dont ils calculent les recommandations. Alors que les systèmes basés sur les utilisateurs se concentrent sur l'identification d'utilisateurs similaires en fonction de leurs comportements et préférences passés, les systèmes basés sur les items se concentrent sur l'identification de similarités entre les items en fonction des interactions des utilisateurs. Autrement dit un profil d'article est construit à partir de la liste des utilisateurs ayant lu ou aimé cet article.

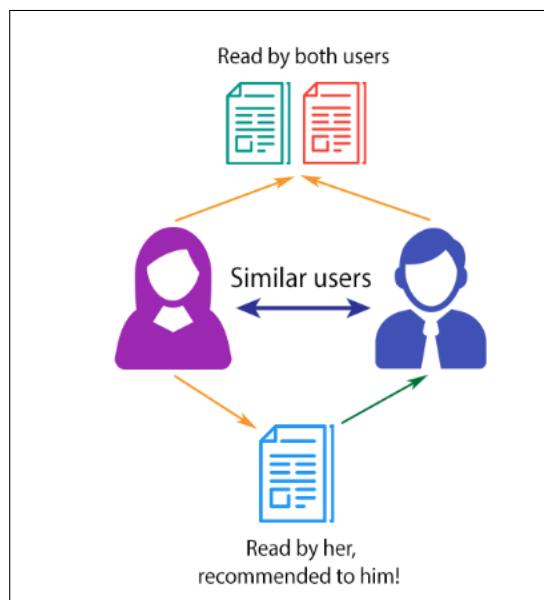


FIGURE 4 – Filtrage collaboratif [1]

Les systèmes de recommandation collaboratifs, par leur diversité, s'appuient donc sur de nombreuses techniques, qu'il s'agisse de :

- similarité entre utilisateurs (coefficient de corrélation de Pearson , etc.) ou de sélection de voisinage (les algorithmes basés sur la recherche de voisinage) pour les approches User-to-User
- similarité entre éléments (la mesure de similarité cosinus , etc.) pour les approches Item-to-Item
- techniques de prédition de scores (analyse en composantes principales ou ACP, factorisation de matrices, analyse sémantique latente, règles d'association, approches bayésiennes, etc.) pour les autres approches.

### c. Filtrage hybride

Cette approche combine les techniques de filtrage basé sur le contenu et de filtrage collaboratif pour tirer parti des avantages de chacune.

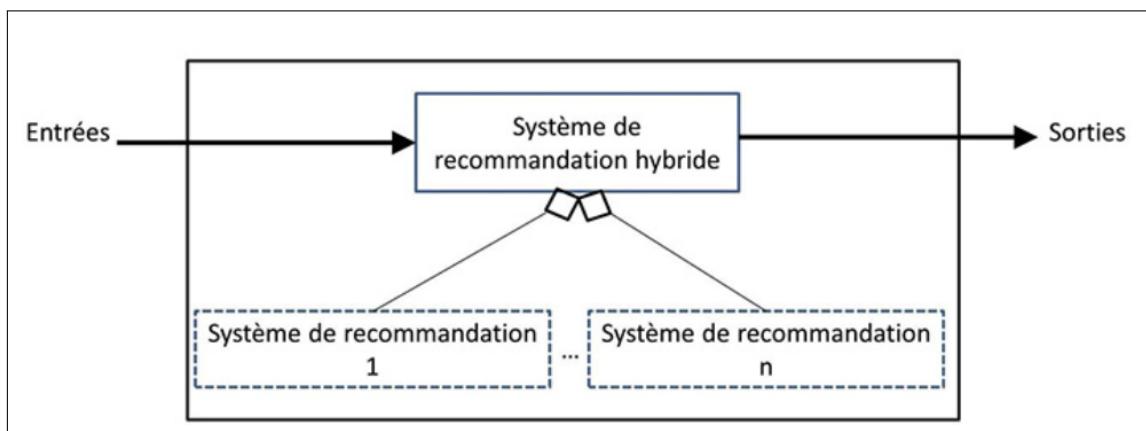


FIGURE 5 – Conception d'hybridation monolithique

La figure présentée illustre de manière captivante le concept fondamental des systèmes de recommandation basés sur le filtrage hybride. Voici une analyse plus détaillée :

• **Entrées (Inputs)** : À gauche de la figure, nous avons deux entrées, qui peuvent être des informations sur les utilisateurs, leurs préférences, ou des éléments à recommander (par exemple, des hotels).

• **Système de Recommandation Hybride** : Le bloc central de la figure représente le système de recommandation hybride. Celui-ci intègre plusieurs systèmes de recommandation individuels pour générer des recommandations globales. Ces systèmes individuels peuvent être basés sur différentes techniques, telles que la filtrage collaboratif, la filtrage basé sur le contenu, ou d'autres algorithmes.

• **Sorties (Outputs)** : À droite de la figure, nous avons les sorties du système de recommandation hybride. Ces sorties sont les recommandations finales qui sont présentées à l'utilisateur.

En résumé, le système de recommandation hybride combine les forces de différents systèmes pour offrir des recommandations plus précises et diversifiées aux utilisateurs. Cela permet d'améliorer l'expérience de l'utilisateur en lui proposant des suggestions pertinentes.

### I.2.3 Comparaison des systèmes de recommandation

Voici un tableau de comparaison basé sur les avantages et inconvénients des systèmes de recommandation basés sur le contenu et collaboratifs :

Caractéristique	Recommandation basée sur le contenu	Recommandation collaborative
Recommande des éléments similaires à ceux aimés dans le passé par les utilisateurs	✓	✓
Prend en compte le profil des utilisateurs pour des recommandations pertinentes	✓	✓
Fonctionne avec de nombreux types de données (textuelles, numériques, etc.)	✓	✗
Les données relatives aux autres utilisateurs sont inutiles	✓	✗
Peut recommander de nouveaux éléments ou des éléments non populaires	✓	✓
Peut distinguer des éléments représentés par le même ensemble de mots-clés	✗	✓
Possède un historique pour les nouveaux utilisateurs	✗	✓
Nécessite le feedback utilisateur pour des recommandations précises	✗	✓
Analyse les préférences et les comportements d'un utilisateur pour recommander des éléments appréciés par d'autres utilisateurs ayant des goûts similaires	✗	✓

TABLE 1 – Comparaison des systèmes de recommandation

Pour tirer parti des caractéristiques spécifiques de chacune des recommandations à savoir collaborative et basée sur le contenu, nous avons choisi d'adopter une approche basée sur la recommandation hybride, qui permet de fournir des recommandations pertinentes en se basant à la fois sur le profil de l'utilisateur lui-même et sur les préférences des utilisateurs qui lui sont similaires .

### I.3 Analyse Sentimentale

L'analyse sentimentale, également appelée opinion mining ou extraction de sentiment, est une méthode essentielle pour comprendre les opinions et les sentiments exprimés dans les commentaires des clients. Elle joue un rôle crucial dans la prédiction de la satisfaction des utilisateurs et dans la recommandation d'hôtels adaptés à leurs préférences.

Dans notre approche, nous utilisons des techniques de traitement du langage naturel (NLP) et d'apprentissage automatique pour analyser les sentiments exprimés dans les commentaires des utilisateurs. Ces techniques nous permettent d'extraire des informations sur la perception générale des utilisateurs à l'égard des hôtels, en évaluant la tonalité positive ou négative des commentaires. En utilisant des méthodes de classification d'apprentissage automatique, notre objectif est de développer un modèle de prédiction pour l'analyse sentimentale des commentaires d'hôtels.

Il existe différentes façons et techniques disponibles pour l'analyse des sentiments, qui sont principalement regroupées en deux grands groupes : 1) les méthodes basées sur les lexiques et 2) les méthodes d'apprentissage automatique .

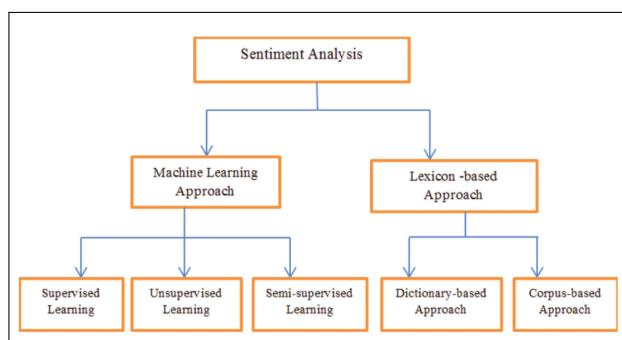


FIGURE 6 – Approches de l'analyse sentimentales

#### I.3.1 Approche basée sur les lexiques

L'approche basée sur les lexiques repose sur l'utilisation de dictionnaires ou de lexiques de mots associés à des émotions spécifiques. Cette méthode consiste à attribuer des scores aux mots en

fonction de leur valence émotionnelle (positive, négative ou neutre) et à agréger ces scores pour évaluer le sentiment global d'un texte.

Par exemple, le lexique VADER (Valence Aware Dictionary and sEntiment Reasoner) est une méthode basée sur un dictionnaire largement utilisée pour l'analyse des sentiments. Il attribue des scores de polarité (positifs, négatifs ou neutres) aux mots en fonction de leur contexte et agrège ces scores pour évaluer le sentiment global d'un texte.

### I.3.2 Approche par Apprentissage Automatique

En revanche, l'approche par apprentissage automatique utilise des algorithmes d'apprentissage supervisé ou non supervisé pour classer automatiquement les textes en fonction de leur sentiment. Les méthodes courantes incluent les classificateurs probabilistes tels que Naïve Bayes, les réseaux de neurones artificiels, les machines à vecteurs de support (SVM) et les arbres de décision. Ces méthodes apprennent à partir de données étiquetées pour prédire le sentiment d'un nouveau texte.

En conclusion, Notre étude vise à fournir aux utilisateurs un outil de recommandation précis et personnalisé en exploitant les données textuelles disponibles sur les plateformes de réservation d'hôtels. En utilisant l'analyse sentimentale, nous cherchons à améliorer la pertinence des recommandations en prenant en compte les préférences subjectives des utilisateurs exprimées dans leurs commentaires. Cette approche permettra d'enrichir l'expérience des utilisateurs en leur proposant des suggestions d'hôtels qui correspondent à leurs attentes et à leurs sentiments

## I.4 Architecture Globale de la solution proposée

L'architecture globale de notre système de recommandation d'hôtels est représentée dans la figure 7. Elle se compose de plusieurs étapes clés, chacune jouant un rôle crucial dans le processus global de recommandation. Cette architecture est divisée en deux parties indépendantes : l'analyse des avis des utilisateurs pour chaque hôtel et l'extraction des profils des utilisateurs à partir des réseaux sociaux, suivies d'une étape de mise en correspondance qui relie les deux parties.

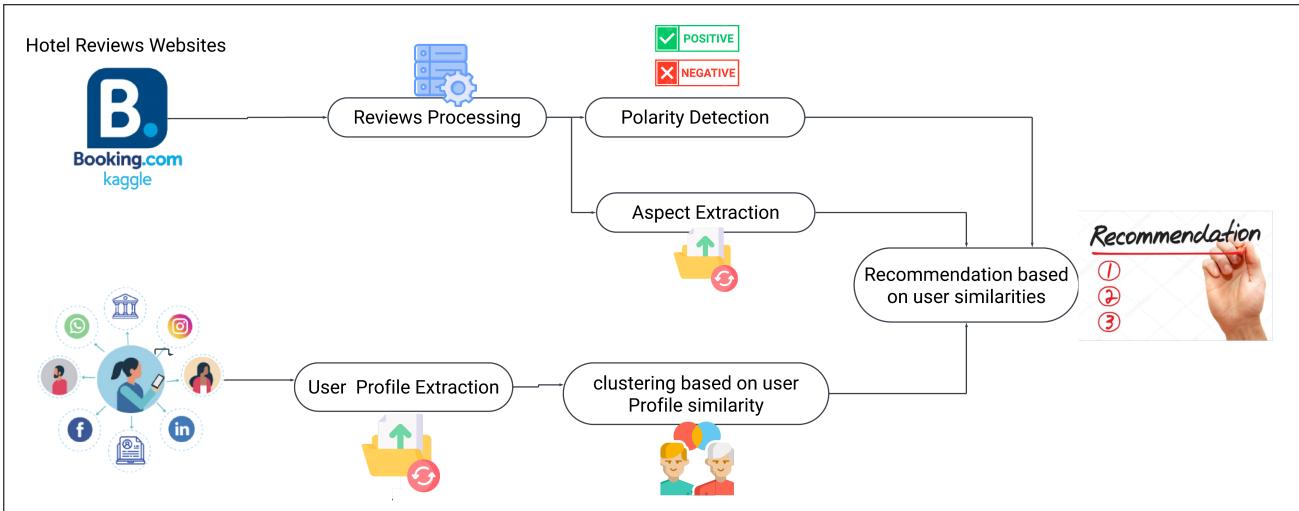


FIGURE 7 – Architecture globale du système de recommandation

#### I.4.1 Analyse Sentimentale des Avis

Nous commençons par collecter des données d'avis sur les hôtels à partir de plateformes telles que Booking.com et Kaggle. Ces données incluent des commentaires et des évaluations des utilisateurs sur divers hôtels. Les avis collectés sont ensuite traités pour extraire les informations pertinentes. Cette étape comprend le nettoyage des données et la préparation des commentaires pour une analyse ultérieure. Nous procédons ensuite à une analyse de sentiment pour déterminer la polarité des commentaires, identifiant ainsi les avis positifs et négatifs. Enfin, l'extraction des aspects permet d'identifier les éléments spécifiques des hôtels mentionnés dans les avis, tels que la propreté, le service, et l'emplacement.

#### I.4.2 Extraction des profils des Utilisateurs

Parallèlement, nous extrayons les profils des utilisateurs à partir de leurs activités sur les réseaux sociaux (comme Instagram, etc.). Cela nous aide à comprendre leurs préférences et centres d'intérêt. Les utilisateurs sont ensuite regroupés en clusters en fonction de la similarité de leurs intérêts. Cela facilite la segmentation des utilisateurs en groupes homogènes ayant des préférences similaires.

#### I.4.3 Mise en Correspondance et Recommandation

Enfin, notre système fournit des recommandations d'hôtels en se basant sur les clusters formés et les analyses de sentiment. Lorsqu'un utilisateur saisit son nom, le système recommande des hôtels appréciés par d'autres utilisateurs du même cluster, assurant ainsi des suggestions personnalisées et pertinentes.

## I.5 Méthodologie du Projet : CRISP-DM

Afin de garantir le bon déroulement de notre projet, il est primordial de mettre en œuvre une méthodologie claire et bien organisée qui correspond à la nature de notre projet. La méthodologie CRISP-DM, largement répandue et parfaitement adaptée à nos besoins, a été adoptée par notre équipe. Les six étapes principales de la méthodologie CRISP-DM nous accompagnent tout au long du projet, depuis la compréhension du problème métier jusqu'au déploiement.

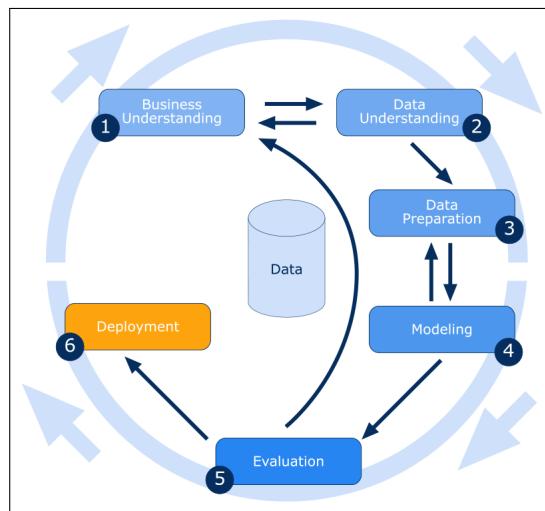


FIGURE 8 – Processus CRISP-DM

**1.Compréhension métier :** Comprendre les objectifs commerciaux et les besoins du projet. Définir le problème à résoudre avec les données.

**2.Compréhension des données :** Explorer les données disponibles, évaluer leur qualité et comprendre leur structure.

**3.Préparation des données :** Nettoyer, transformer et préparer les données pour l'analyse. Cela inclut la sélection des variables, l'imputation des valeurs manquantes, etc.

**4.Modélisation :** Construire des modèles prédictifs ou des algorithmes d'apprentissage automatique en utilisant les données préparées.

**5.Évaluation :** Évaluer la performance des modèles et les ajuster si nécessaire. Vérifier qu'ils répondent aux objectifs commerciaux.

**6.Déploiement :** Mettre en œuvre les modèles dans un environnement opérationnel et surveiller leur performance.

## **Conclusion**

Ce chapitre présente une vue d'ensemble de notre projet, incluant le contexte, la problématique, les objectifs, l'architecture globale du système de recommandation et la méthodologie adoptée. Le chapitre suivant établit les sources de données, telles que les profils des utilisateurs, ainsi que les avis sur les hôtels, et les étapes de préparation des données pour l'analyse.

## II Compréhension et préparation des données

### Introduction

Ce chapitre se concentre sur la compréhension des données utilisées et leur préparation pour notre système de recommandation d'hôtels. Nous détaillerons les sources de données, à savoir les profils des utilisateurs avec leurs centres d'intérêt et les avis d'hôtels de Booking.com. La préparation des données, incluant le nettoyage et le filtrage, est essentielle pour garantir la précision et l'efficacité de notre modèle de recommandation.

#### II.1 Data Understanding

Dans cette section, nous fournirons une explication détaillée des sources de données utilisées dans notre système de recommandation d'hôtels, suivie des étapes de préparation de ces données pour l'analyse. Notre système s'appuie principalement sur deux sources de données : les profils des utilisateurs avec leurs centres d'intérêt et les avis d'hôtels de Booking.com. La préparation de ces données est cruciale pour garantir la précision et l'efficacité de notre modèle de recommandation.

##### II.1.1 Sources de Données : Profils et Intérêts des Utilisateurs

Notre première source de données consiste en des profils d'utilisateurs, comprenant des informations détaillées sur les utilisateurs, leurs abonnés, leurs publications Instagram et leurs centres d'intérêt. Ce jeu de données comprend 884 instances (profils d'utilisateurs) et 6 attributs principaux :

- **username** : Nom de l'utilisateur
- **user\_id** : Identifiant unique de l'utilisateur
- **followers\_count** : Nombre de followers de l'utilisateur
- **posts\_count** : Nombre de publications de l'utilisateur
- **posts** : Détails des publications de l'utilisateur
- **interests** : Centres d'intérêt de l'utilisateur

##### II.1.2 Sources de Données : Avis sur les Hôtels

Cette source regroupe les avis des utilisateurs sur différents hôtels, provient exclusivement de Booking.com, l'un des principaux sites de réservation d'hôtels en ligne et a été collecté à partir de Kaggle, une plateforme populaire pour les ensembles de données et les projets de data science.

Ce jeu de données comporte 26675 instances et les 16 attributs suivants :

- **review\_title** : Titre de l'avis
- **reviewed\_at** : Date de l'avis
- **reviewed\_by** : Auteur de l'avis
- **images** : Images incluses dans l'avis
- **crawled\_at** : Date de collecte des données
- **url** : URL de l'avis
- **hotel\_name** : Nom de l'hôtel
- **hotel\_url** : URL de l'hôtel
- **avg\_rating** : Note moyenne de l'hôtel
- **nationality** : Nationalité de l'auteur de l'avis
- **rating** : Note de l'avis
- **review\_text** : Texte de l'avis
- **raw\_review\_text** : Texte brut de l'avis
- **tags** : Tags associés à l'avis
- **meta** : Métadonnées

## II.2 Data Preparation

### II.2.1 Sources de Données : Profils et Intérêts des Utilisateurs

Avant d'entreprendre des analyses plus approfondies, les données ont été préparées pour garantir leur qualité et leur pertinence. L'extraction des centres d'intérêt des utilisateurs à partir des données brutes avait pour objectif de regrouper les intérêts des utilisateurs dans une structure facilement analysable. La fonction développée pour cette tâche a été conçue spécifiquement pour parcourir les données brutes et extraire les sous-catégories d'intérêt de chaque catégorie d'intérêt. L'objectif principal de cette fonction était de simplifier et de standardiser la représentation des centres d'intérêt des utilisateurs. Voici un exemple de données d'entrée et de sortie de la fonction d'extraction des centres d'intérêt.

Input	Output
Business and Industry: [social media, Business], family and relationships: [family, parenting], fitness and wellness: [Physical exercise], shopping and fashion: [clothing, fashion accessories], sports: [football/soccer]	social media , Business family, parenting, Physical exercise ,clothing, fashion accessories, football/soccer

FIGURE 9 – Préparation des intérêts

Une analyse des centres d'intérêt a ensuite été réalisée pour comprendre les préférences globales des utilisateurs. Cette analyse a révélé les intérêts les plus courants parmi les utilisateurs, ainsi que leur répartition. Par exemple, il a été observé que les intérêts liés aux médias sociaux, à la mode et au divertissement étaient parmi les plus répandus.

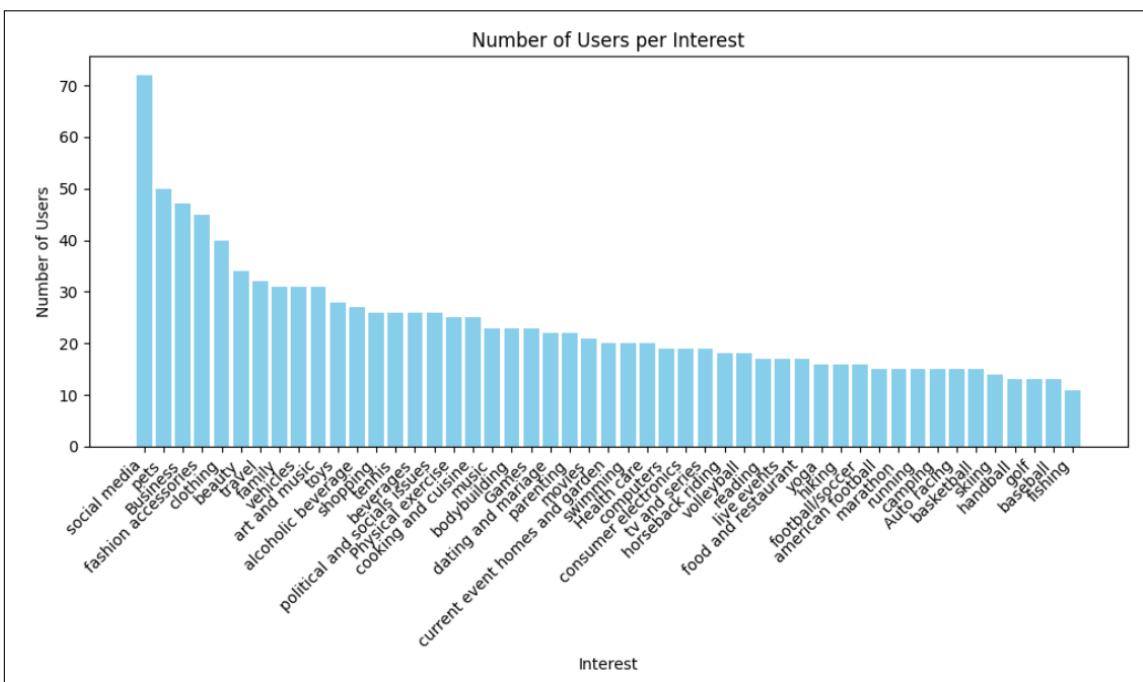


FIGURE 10 – Préférences globales des utilisateurs

L'analyse et la visualisation des centres d'intérêt des utilisateurs nous permettent de comprendre les préférences dominantes au sein de notre jeu de données.

## II.2.2 Sources de Données : Avis sur les Hôtels

Dans cette phase de préparation des données, plusieurs étapes ont été entreprises pour garantir la qualité et la cohérence des avis sur les hôtels. Voici les principales actions réalisées :

**Filtrage des Colonnes Requises :** Seules les colonnes pertinentes pour l'analyse ont été conservées, notamment 'reviewed\_by' (l'auteur de l'avis), 'hotel\_name' (nom de l'hôtel), 'hotel\_url'

(URL de l'hôtel), 'review\_title' (titre de l'avis), 'review\_text' (texte de l'avis), 'rating' (note attribuée à l'hôtel) et 'tags' (tags associés à l'avis).

### Traitement des Avis et Normalisation du Texte :



FIGURE 11 – Avis avant préparation

Les avis ne contenant pas de texte mais seulement le message "There are no comments available for this review" ont été remplacés par une chaîne vide.

Le texte des avis a été normalisé pour assurer une cohérence dans l'analyse :

- Conversion en minuscules pour éviter les doublons dus à la casse.
- Suppression des espaces supplémentaires entre les mots.
- Suppression des mots "everything" et "nothing", jugés non informatifs dans ce contexte.

**Combinaison du Titre et du Texte de l'Avis :** Étant donné qu'il y a beaucoup d'avis avec un titre mais sans texte, les deux variables ont été combinées en une seule, nommée 'final\_review', pour simplifier l'analyse.

### Suppression des Valeurs Manquantes :

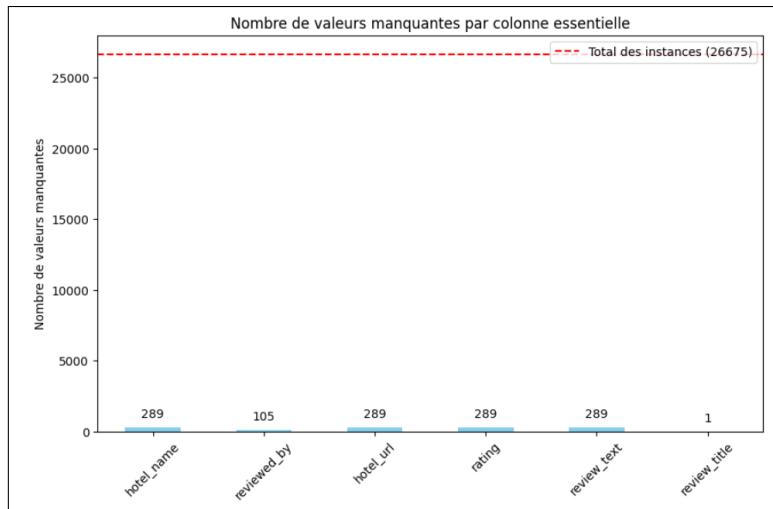


FIGURE 12 – Nombre de valeurs manquantes par colonne essentielle

Toutes les lignes contenant des valeurs manquantes dans les colonnes essentielles telles que 'hotel\_name', 'reviewed\_by', 'hotel\_url', 'rating' et 'final\_review' ont été supprimées.

## Conclusion

Au cours de ce chapitre, nous avons minutieusement préparé les deux sources de données principales pour notre système de recommandation hôtelière . Afin de faciliter la création de clusters d'utilisateurs, nous avons extrait et organisé les sous-catégories d'intérêts pour les utilisateurs d'Instagram. Afin de garantir la qualité des données, nous avons procédé à un filtrage, à un nettoyage et à une standardisation des avis sur les hôtels de Booking.com. Ces préparatifs assurent une fondation solide pour les analyses émotionnelles et la correspondance des utilisateurs, préparant ainsi le terrain pour des recommandations hôtelières personnalisées performantes.

# III Modélisation et évaluation

## Introduction

Dans ce chapitre, nous présenterons d'abord l'architecture détaillée ensuite nous abordons les étapes et les méthodes utilisées pour créer notre système de recommandation . La modélisation se divise en trois phases distinctes :L'extraction des intérêts à partir des publications Instagram (clustering des similarités entre utilisateurs), l'analyse sentimentale des avis sur les hôtels de chaque utilisateur et finalement le matching des utilisateurs des deux datasets et la génération de recommandations.

### III.1 Architecture détaillée de la solution proposée

Dans la figure 13, nous présenterons l'architecture détaillée de la solution proposée pour notre système de recommandation. Cette architecture comprendra les composants essentiels ainsi que les étapes et différentes méthodes utilisées

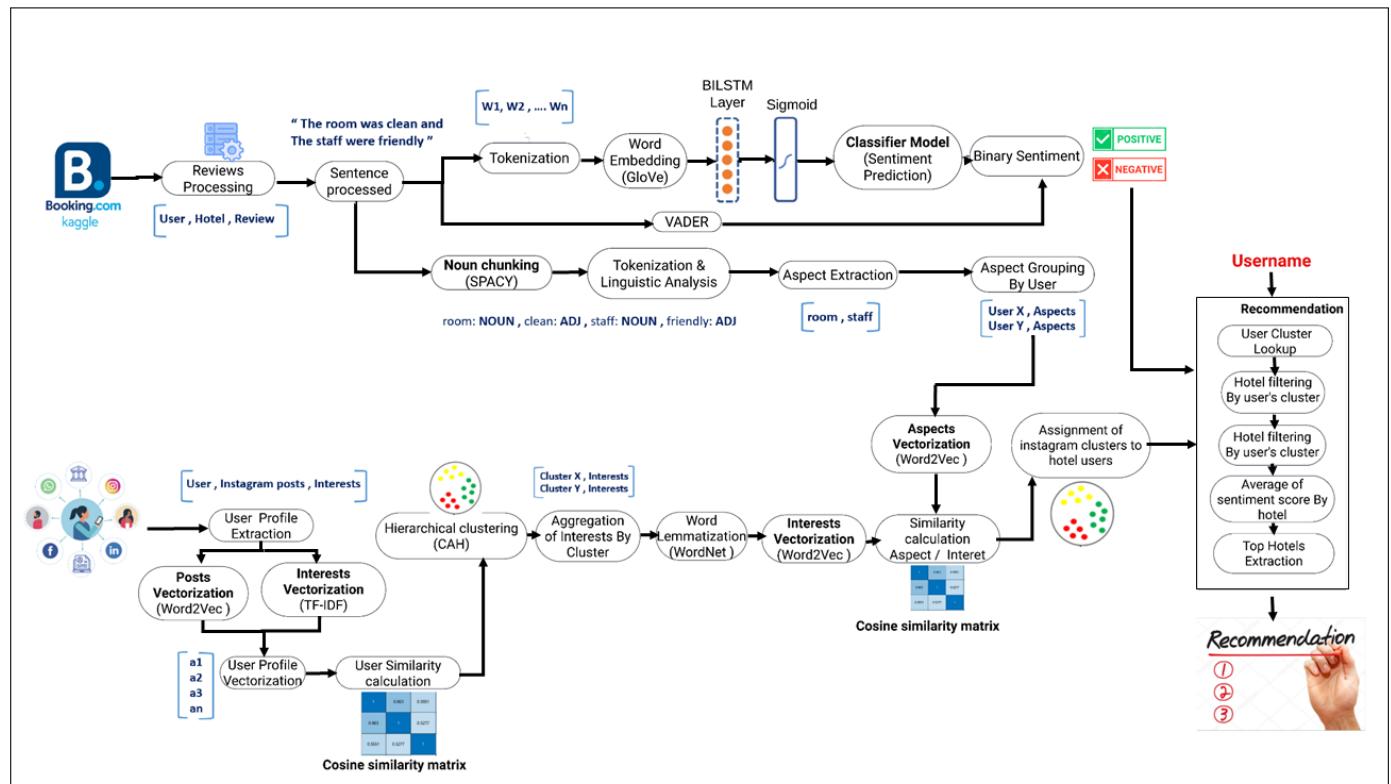


FIGURE 13 – Architecture détaillée de la solution proposée

## III.2 Clustering des similarités entre utilisateurs en fonction de leurs intérêts et légendes

### III.2.1 Représentation Vectorielle

Dans cette partie, nous allons nous concentrer sur la représentation vectorielle des centres d'intérêt des utilisateurs et des légendes de leurs publications afin de les transformer en données numériques exploitables.

**Représentation vectorielle des intérêts :** Nous avons utilisé la technique **TfidfVectorizer**[2] pour représenter les centres d'intérêt des utilisateurs. Cette méthode génère une représentation pondérée des termes d'intérêt, en tenant compte de leur importance relative.

**Représentation vectorielle des légendes de publications :** Pour capturer les informations sémantiques des légendes de publications des utilisateurs, nous avons utilisé **Word2Vec**[6]. Cette méthode produit des vecteurs sémantiques pour chaque mot dans les légendes de publications.

### III.2.2 Similarité cosinus

Nous avons mesuré la similarité entre les utilisateurs en calculant la similarité cosinus entre leurs vecteurs d'intérêts et de légendes de publications combinés. Cette technique est largement utilisée dans le contexte des systèmes de recommandation et catégorisation car elle permet de capturer la similitude sémantique entre les vecteurs.

#### Identification des utilisateurs les plus similaires :

Nous avons développé une fonctionnalité pour identifier les utilisateurs les plus similaires. Elle récupère les scores de similarité entre un utilisateur donné et tous les autres utilisateurs, exclut l'utilisateur donné de la liste des recommandations, trie les utilisateurs par score de similarité et sélectionne les 10 plus similaires. Enfin, elle renvoie les noms et les centres d'intérêt des utilisateurs recommandés.

#### Clustering hiérarchique CAH :

Nous avons utilisé un algorithme de clustering hiérarchique pour regrouper les utilisateurs. Cette méthode permet de créer une hiérarchie de clusters en fusionnant progressivement les plus similaires entre eux, formant ainsi un dendrogramme.

Grâce à cette méthode, nous pouvons identifier des groupes d'intérêt à différentes granularités.

#### Choix optimal du nombre de clusters :

La figure 14 montre l'évolution du score de silhouette en fonction du nombre de clusters.

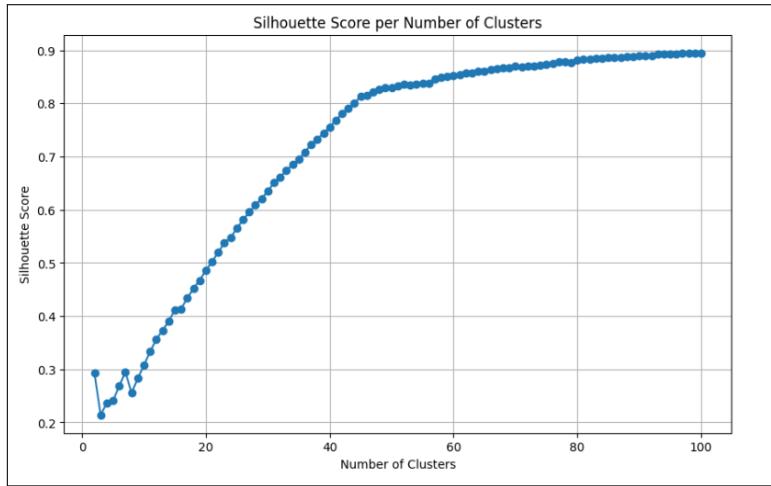


FIGURE 14 – Score de Silhouette

On observe que le score de silhouette commence à s'élever significativement à partir de 45 clusters. Cela suggère que le clustering avec 45 clusters est susceptible d'être optimal pour les données en question.

#### Analyse et visualisation du résultat du clustering obtenu :

La figure 15 présente un graphique illustrant la répartition du nombre d'utilisateurs par cluster. L'analyse du graphique révèle une variation significative du nombre d'utilisateurs par cluster. Les clusters représentent des segments de clientèle distincts avec des intérêts différents.

On peut observer que le cluster le plus peuplé rassemble environ 40 utilisateurs, tandis que le plus petit compte 11 utilisateurs, ce qui est cohérent par rapport au nombre d'utilisateurs total qui est de 884 utilisateurs.

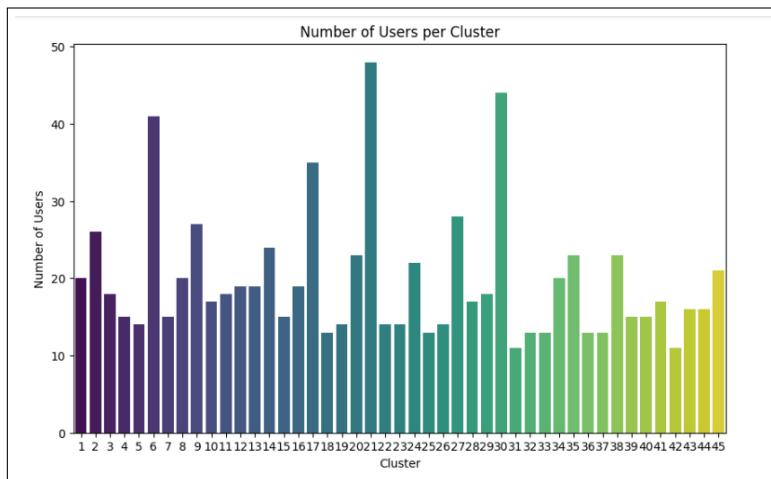


FIGURE 15 – Répartition du nombre d'utilisateurs par cluster

## Evaluation du résultat de clustering obtenu :

Pour s'assurer de la cohérence des résultats de clustering nous avons affiché la matrice de similarité de quelques clusters. A titre d'exemple, on prend le cluster numéro 5 comme le montre la figure 16 de la matrice , les valeurs de similarité sont toutes proches 1, ce qui confirme que les utilisateurs au sein du même cluster sont similaires entre eux.

```
Similarity matrix for users in Cluster 5:  
[[1.          0.76408163 0.72225644 0.72225745 0.70182366 0.79438801  
  0.79453449 0.52721545 0.51791056 0.51837517 0.51818135 0.51796366  
  0.5175587  0.5177957 ]]  
[[0.76408163 1.          0.93409471 0.93409145 0.56096075 0.75248047  
  0.75206737 0.49923552 0.49068846 0.49113879 0.49089153 0.49084786  
  0.49029531 0.49065216 ]]  
[[0.72225644 0.93409471 1.          0.9999997 0.6072678 0.80689387  
  0.80659826 0.39611223 0.52779327 0.52812008 0.52784191 0.52794922  
  0.52744598 0.52785824 ]]  
[[0.72225745 0.93409145 0.9999997 1.          0.60728324 0.80689188  
  0.80659947 0.39611455 0.52779657 0.52812028 0.52784324 0.52794714  
  0.52744991 0.52786145 ]]  
[[0.70182366 0.56096075 0.6072678 0.60728324 1.          0.61062162  
  0.61103414 0.56345702 0.73299879 0.73222612 0.73201405 0.73232907  
  0.73258915 0.7327344 ]]  
[[0.79438801 0.75248047 0.80689387 0.80689188 0.61062162 1.  
  0.9995753 0.50261255 0.65543719 0.65609201 0.65596038 0.65573757  
  0.65526875 0.65564446 ]]  
[[0.79453449 0.75206737 0.80659826 0.80659947 0.61103414 0.9995753  
  1.          0.50269718 0.65554861 0.65599661 0.65584265 0.65571371  
  0.65539955 0.65573155 ]]  
[[0.52721545 0.49923552 0.39611223 0.39611455 0.56345702 0.50261255  
  0.50269718 1.          0.78890491 0.78903304 0.78889141 0.78923091  
  0.78911437 0.78916808 ]]
```

FIGURE 16 – Matrice de similiarité pour les utilisateurs du cluster 5

Nous pouvons aussi vérifier la cohérence sémantique en affichant les noms réels d'utilisateurs au sein du cluster 5. Comme le montre la figure 17, Cristiano Ronaldo est l'utilisateur le plus similaire à lui-même, suivi de Beckham, Morata, Messi et d'autres joueurs de football célèbres. Les centres d'intérêt des utilisateurs recommandés sont cohérents avec ceux de Cristiano Ronaldo, ce qui indique un regroupement efficace des utilisateurs en fonction de leurs préférences.

```
Users in the Cluster 5:  
['Cristiano Ronaldo', 'David Beckham', 'Álvaro Morata', 'Álvaro  
Morata', 'Leo Messi', 'Jack Grealish', 'Rúben Gato Dias PT',  
'Nike', 'Real Madrid C.F.', 'FC Bayern', 'UEFA Champions League',  
'Kyle Walker', 'Premier League', 'LIVE HERE WE GO 🚨 ']
```

FIGURE 17 – Les utilisateurs appartenant au cluster 5

Le clustering des utilisateurs basé sur leurs centres d'intérêt et les légendes de leurs publications Instagram permet de regrouper les utilisateurs ayant des intérêts similaires, ce qui permet de proposer des recommandations plus personnalisées et pertinentes. La combinaison de différentes

techniques de vectorisation et de l'algorithme de clustering hiérarchique offre une approche flexible et efficace pour identifier les utilisateurs partageant des similarités.

### III.3 Analyse sentimentale des avis sur les hôtels

#### III.3.1 VADER

Pour analyser les sentiments exprimés dans les avis d'hôtels, nous avons utilisé VADER (Valence Aware Dictionary and sEntiment Reasoner) [3], un outil de traitement du langage naturel (NLP) spécialisé dans l'analyse des sentiments exprimés dans les textes.

##### Calcul du Score de Sentiment

Nous avons appliqué VADER pour chaque avis afin de calculer un score de sentiment 'compound'. Ce score est une valeur comprise entre -1 (extrêmement négatif) et 1 (extrêmement positif).

##### Classification des Sentiments

En se basant sur le score de sentiment, les avis ont été classifiés en deux catégories : positifs et négatifs. Les avis avec un score supérieur à 0.25 ont été considérés comme positifs, tandis que les autres ont été classifiés comme négatifs.

##### Visualisation et Statistiques

La distribution des sentiments a été visualisée à l'aide d'un diagramme en barres, ce qui a permis de voir clairement la répartition des avis positifs et négatifs dans notre dataset. La répartition des sentiments est la suivante :

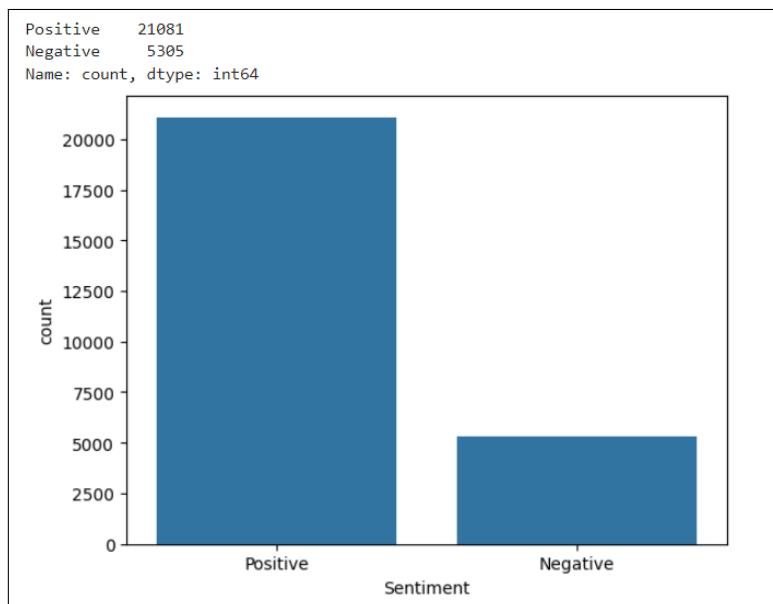


FIGURE 18 – Répartition des avis positifs et négatifs avec VADER

## Évaluation du Modèle

Pour évaluer la performance de notre analyse sentimentale, nous avons comparé les résultats obtenus avec des labels de référence (ground truth). Une matrice de confusion et un rapport de classification ont été utilisés pour cette évaluation. Les résultats montrent que le modèle d'analyse sentimentale basé sur VADER a une précision élevée pour les avis positifs mais des performances médiocres pour les avis négatifs.

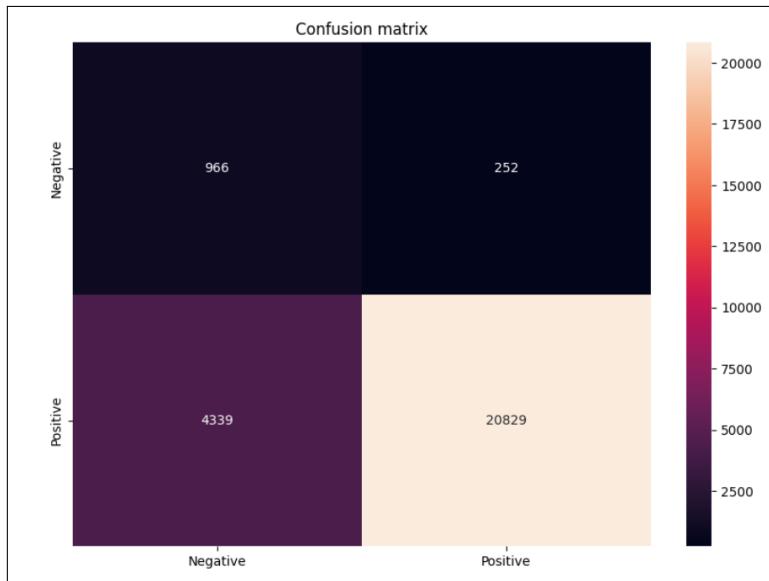


FIGURE 19 – Matrice de Confusion - VADER

	precision	recall	f1-score	support
Negative	0.18	0.79	0.30	1218
Positive	0.99	0.83	0.90	25168
accuracy			0.83	26386
macro avg	0.59	0.81	0.60	26386
weighted avg	0.95	0.83	0.87	26386

FIGURE 20 – Rapport de Classification - VADER

Ces résultats montrent que notre modèle est très performant pour identifier les avis positifs, mais il a du mal à détecter correctement les avis négatifs. Cela peut être dû à l'asymétrie dans le nombre d'avis positifs et négatifs, ainsi qu'à la nature des textes analysés.

En conclusion, l'analyse sentimentale des avis sur les hôtels est une étape cruciale dans notre système de recommandation hôtelière. Bien que notre modèle actuel basé sur VADER soit efficace pour détecter les avis positifs, il nécessite des améliorations pour mieux identifier les avis négatifs.

### III.3.2 Utilisation des Réseaux de Neurones

Pour améliorer la performance de l'analyse sentimentale, nous avons exploré l'utilisation de réseaux de neurones LSTM [4] avec des embeddings GloVe [5] et Word2Vec [6]. Voici un résumé des différentes approches :

#### Équilibrage des Données

Pour chaque modèle, nous avons équilibré le dataset en prenant un nombre égal de critiques positives et négatives.

#### Préparation des Données

Les textes ont été tokenisés et convertis en séquences. Les séquences ont été remplies (padding) pour avoir une longueur uniforme.

#### Modèles Utilisés :

Voici un tableau de comparaison des scores de chaque modèle :

Modèle	Paramètres	Train Accuracy	Test Accuracy
Simple LSTM	<ul style="list-style-type: none"> <li>— Couche Embedding</li> <li>— Couche LSTM (64)</li> <li>— Couche Dense (Activation function : sigmoid)</li> </ul>	54.88%	56.76%
LSTM + Word2Vec	<ul style="list-style-type: none"> <li>— Couche Embedding</li> <li>— Couche LSTM</li> <li>— Couche Dense (Activation function : sigmoid)</li> </ul>	56.03%	75.78%
BiLSTM + GloVe	<ul style="list-style-type: none"> <li>— Couche Embedding</li> <li>— Couche BiLSTM (64)</li> <li>— Couche Dense (Activation function : sigmoid)</li> </ul>	93.20%	93.44%
BiLSTM	<ul style="list-style-type: none"> <li>— Couche Embedding</li> <li>— Couche BiLSTM</li> <li>— Couche Dense (Activation function : sigmoid)</li> </ul>	99.17%	93.23%

Le tableau présente les performances de différents modèles utilisés pour l'analyse des sentiments à partir des avis sur les hôtels. Tous les modèles ont été compilés avec la fonction de perte binaire (binary crossentropy) et l'optimiseur Adam, et ont été entraînés sur un ensemble de données avec un ratio de 80% pour les données d'entraînement et 20% pour les données de test.

Parmi les modèles que nous avons évalués, deux ont présenté des performances élevées : le BiLSTM simple et le BiLSTM avec GloVe. Cependant, le modèle sans GloVe a montré un surajustement (overfitting). Pour illustrer cela, la figure 21 montre l'évolution de l'exactitude et de la perte au fil des époques pour ce modèle.

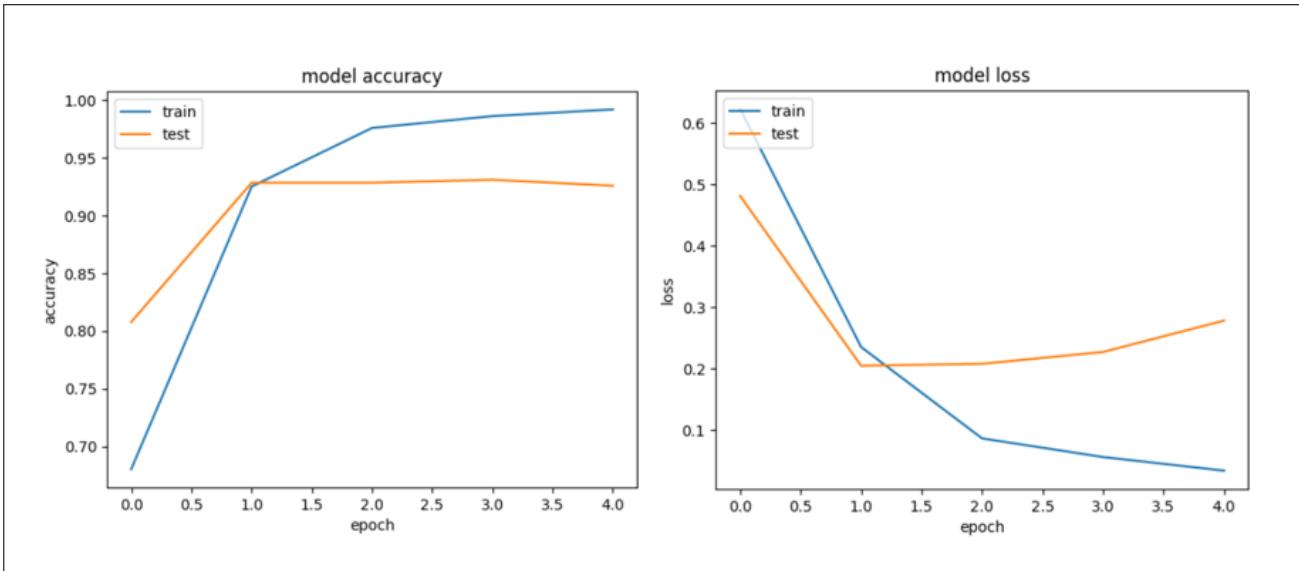


FIGURE 21 – Évolution du modèle biLSTM avec overfitting

la figure ci dessus montre que : L'exactitude (accuracy) sur l'ensemble d'entraînement augmente progressivement à mesure que le modèle apprend. Cela signifie que le modèle s'adapte bien aux données d'entraînement et parvient à réduire l'erreur. Tandis que l'exactitude sur l'ensemble de test atteint un certain seuil, puis commence à diminuer. Cela indique que le modèle ne généralise pas bien aux données qu'il n'a pas vues pendant l'entraînement. En d'autres termes, il surajuste les données d'entraînement et ne parvient pas à bien se comporter sur de nouvelles données.

Après analyse, nous avons choisi d'utiliser l'architecture BiLSTM avec l'incorporation pré-entraînée GloVe (BiLSTM + GloVe), qui a affiché les meilleures performances en termes de précision lors des tests, avec un score d'exactitude de 93,44%.

Cette architecture a été choisie pour plusieurs raisons :

l'utilisation d'une couche BiLSTM permet de capturer efficacement les informations contextuelles des séquences de mots, en tenant compte à la fois des mots précédents et suivants dans une séquence. Cela permet au modèle de mieux comprendre le contexte global des avis sur les hôtels.

l'incorporation pré-entraînée GloVe (Global Vectors for Word Representation) permet au modèle d'utiliser des représentations vectorielles de mots pré-entraînées. Ces embeddings sont basés sur de vastes corpus de texte, ce qui permet au modèle d'avoir une meilleure compréhension des mots et de leur signification .

une visualisation de l'évolution de l'exactitude et de la perte au fil des époques a été effectuée. Voici les graphiques obtenus :

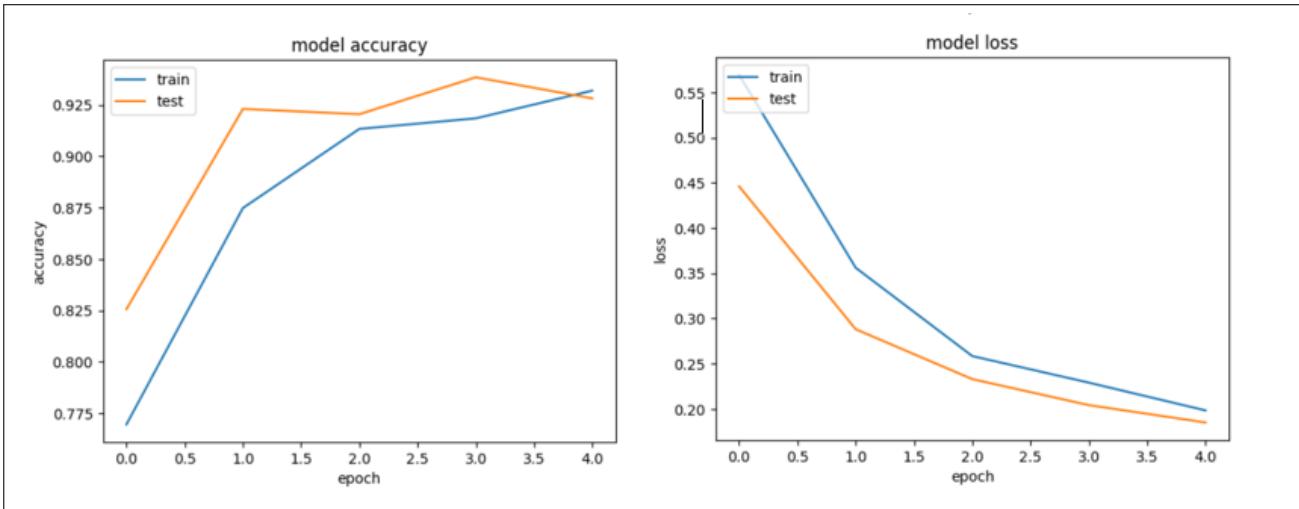


FIGURE 22 – Évolution du modèle BiLSTM + GloVE

Le premier graphique représente l'évolution de l'exactitude du modèle sur les ensembles d'entraînement et de test au fil des époques. On peut observer une augmentation progressive de l'exactitude sur l'ensemble de test, ce qui indique que le modèle apprend de manière efficace les caractéristiques des données et généralise bien sur des données qu'il n'a pas vues lors de l'entraînement. Toutefois, il convient de souligner que l'exactitude sur l'ensemble de test atteint un niveau maximal après quelques époques, ce qui laisse entendre que l'ajout d'époques supplémentaires pourrait ne pas améliorer davantage les performances du modèle.

Le deuxième graphique présente l'évolution de la perte du modèle sur les ensembles d'entraînement et de test au fil des époques. On peut remarquer une diminution progressive de la perte sur l'ensemble de test, ce qui indique que le modèle apprend à prédire avec précision les étiquettes de classe.

En ce qui concerne les performances spécifiques du modèle, une matrice de confusion a été calculée pour évaluer la capacité du modèle à prédire les étiquettes de classe. La matrice de confusion montre que le modèle a correctement classé la plupart des échantillons, avec 1203 vrais négatifs et 22461 vrais positifs.

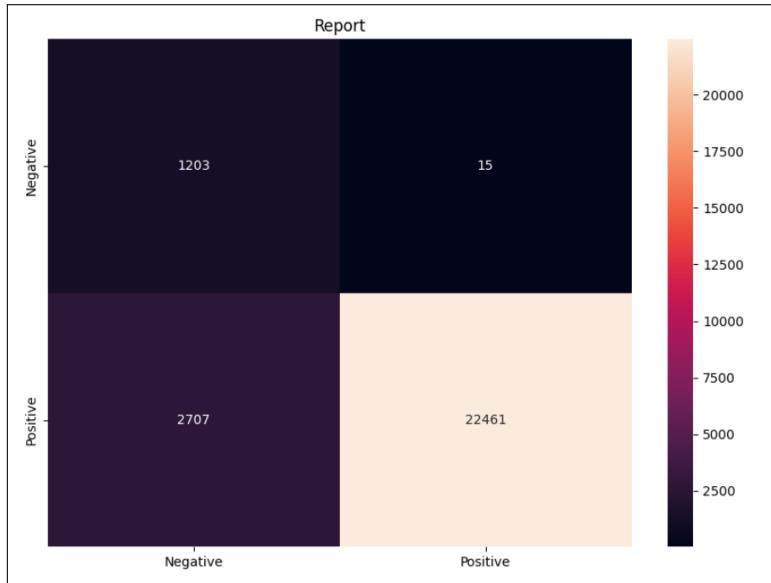


FIGURE 23 – Matrice de Confusion du modèle

### III.3.3 Prédictions

Voici les résultats de quelques prédictions, incluant les détails des utilisateurs, des hôtels, et les sentiments prédits.

reviewed_by	hotel_name	hotel_url	rating	tags	final_review	label_sentiment	predicted_score	predicted_sentiment
0	Kyrylo	Villa Pura Vida	https://www.booking.com/hotel/be/villa-pura-vi... 10.0	Business trip~Solo traveller~Junior Suite~Stay...	exceptional. was perfect! quite, cozy place t...	Positive	0.999794	Positive
1	Dimitri	Villa Pura Vida	https://www.booking.com/hotel/be/villa-pura-vi... 9.0	Leisure trip~Couple~Deluxe Suite~Stayed 1 nigh...	i highly recommend this b&b we enjoyed it a l...	Positive	0.995007	Positive
2	Virginia	Hydro Palace Apartment	https://www.booking.com/hotel/be/hydro-palace... 10.0	Leisure trip~Couple~Apartment with Sea View~St...	exceptional. it was just what we wanted for a ...	Positive	0.999318	Positive

FIGURE 24 – Affectation des scores de prédictions

Chaque critique a été classée comme positive ou négative en fonction de la probabilité prédite par le modèle. Les résultats sont interprétés en fonction de la probabilité : une probabilité supérieure ou égale à 0.5 indique un sentiment positif, tandis qu'une probabilité inférieure à 0.5 indique un sentiment négatif.

```

1/1 [=====] - 0s 50ms/step
it smelt like raw sewerage and was so disappointing
Predicted sentiment : Negative
passable
Predicted sentiment : Negative
the hotel is exceptional
Predicted sentiment : Positive
disappointing
Predicted sentiment : Negative
bathroom floor dirty
Predicted sentiment : Negative
i love it
Predicted sentiment : Positive
the bed wasn't clean and not comfortable
Predicted sentiment : Positive
i didn't like the food
Predicted sentiment : Negative

```

FIGURE 25 – Prédictions sur des critiques inconnues

Les prédictions réalisées sur les critiques inconnues montrent que notre modèle BiLSTM est capable de généraliser efficacement et de fournir des prédictions précises sur des nouvelles données. La plupart des prédictions alignent bien avec l'intuition humaine concernant les sentiments exprimés dans les critiques, démontrant ainsi la robustesse et la fiabilité de notre modèle de prédiction des sentiments.

En conclusion, le modèle BiLSTM + GloVe présente de bonnes performances dans la prédiction des sentiments à partir des avis sur les hôtels, avec une précision globale de 90

### III.3.4 Extraction des aspects des utilisateurs

L'extraction des aspects est une tâche essentielle pour comprendre les préoccupations principales des utilisateurs dans leurs critiques d'hôtels. Pour accomplir cela, nous avons utilisé la bibliothèque spaCy qui nous a permis de traiter efficacement le langage naturel et d'extraire des aspects pertinents des critiques.

#### Chargement et Préparation

Nous avons utilisé le modèle en\_core\_web\_sm de spaCy pour analyser les phrases des critiques. Ce modèle est bien adapté pour identifier les structures grammaticales et les dépendances syntaxiques dans le texte.

#### Extraction des Aspects

Les aspects ont été extraits en identifiant les sujets nominaux dans les critiques. Les sujets identifiés ont été considérés comme des aspects potentiels, par exemple, "service", "chambre", "emplacement".

#### Nettoyage des Données :

Les aspects extraits ont été nettoyés pour s'assurer que chaque aspect soit mentionné de manière unique et sans redondance.

## Groupement par Utilisateur

Les aspects ont été regroupés par utilisateur pour fournir une vue d'ensemble des aspects mentionnés par chaque utilisateur dans toutes ses critiques.

## Résultats

Les résultats de cette extraction montrent que nous avons réussi à identifier et regrouper les aspects clés mentionnés dans les critiques d'hôtels. Par exemple, pour un utilisateur donné, nous avons pu identifier des aspects tels que "service", "chambre", et "emplacement".

2	Virginia	room, ,windows,rooms , bathrooms ,shower,roo...
3	Kannan	house,stay
4	Sue	apartment, reception,water,beds,area , recepti...
...	...	...
26369	Aynur	room

FIGURE 26 – Échantillon d'extraction des aspects par utilisateurs

Cette extraction d'aspects permet de mieux comprendre les éléments spécifiques des hôtels qui sont importants pour les utilisateurs. Ces informations sont particulièrement utiles pour l'affectation des utilisateurs aux clusters de l'étape précédente .

L'utilisation de VADER pour l'analyse sentimentale a fourni une première estimation rapide et efficace des sentiments dans les avis d'hôtels. Cependant, l'application de réseaux de neurones, en particulier les modèles Bidirectional LSTM avec des embeddings GloVe, a considérablement amélioré la précision de la classification des sentiments. Ces résultats démontrent l'importance d'utiliser des techniques avancées de NLP pour obtenir des recommandations plus précises et pertinentes dans notre système de recommandation hôtelière.

## III.4 Matching des utilisateurs et recommandation

### III.4.1 Matching des utilisateurs

#### Agrégation des intérêts de chaque cluster

Afin de pouvoir faire la correspondance entre les utilisateurs Instagram et les utilisateurs qui ont séjourné dans des hôtels, et attribuer à ces derniers un cluster parmi les clusters formés à partir de l'analyse des intérêts des utilisateurs instagram, on procède à l'agrégation des intérêts de chaque cluster.

cluster		interests_text
0	1	current event homes and garden
1	2	political and social issues
2	3	running, Physical exercise, social media, pare...
3	4	family, american football
4	5	basketball, dating and mariage, beauty, Physic...
5	6	shopping, social media, fashion accessories, c...
6	7	vehicles, Physical exercise, social media, Aut...
7	8	bodybuilding, Health care, parenting, social m...
8	9	alcoholic beverage, beverages
9	10	food and restaurant

FIGURE 27 – Agrégation des intérêts de chaque cluster

On convertit ensuite les intérêts agrégés de chaque cluster en vecteurs ?TF-IDF, puis on normalise ces vecteurs pour avoir des analyses plus précises.

#### Calcul de la similarité entre les aspects d'hôtels de chaque utilisateur et les intérêts agrégés des clusters :

Pour réaliser le matching des utilisateurs avec les clusters d'intérêt, nous avons créé une matrice de similarité entre les aspects des utilisateurs et les intérêts des clusters. Voici les étapes détaillées de ce processus :

- **Lemmatisation et expansion des synonymes des intérêts et aspects :** Nous avons utilisé la bibliothèque NLTK[9] pour convertir les mots en leurs formes de base et étendre les synonymes des aspects et intérêts, étant donné que les deux ne partagent pas exactement le même corpus. Pour cela, nous avons utilisé WordNet qui est une base de données lexicale de l'anglais qui organise les mots en “synsets” et fournit des informations sur leurs relations sémantiques. En effet, l'utilisation des synsets de WordNet via NLTK, nous permet d'explorer les relations conceptuelles entre les mots, telles que les hyperonymes (mots plus généraux), les hyponymes (mots plus spécifiques), les synonymes, les antonymes, etc. Cela permet une analyse plus approfondie du sens des mots dans un contexte donné.

	reviewed_by	aspects	expanded_aspects
0	Kyrylo	Solo traveller,Stayed 1 night,,Business trip,J...	{Jnr, persist, stayed, „ stick, I, take, stay...
1	Dimitri	Deluxe Suite,Executive Double Room with Sofa B...	{residue, persist, ii, jazz, state, get_laid, ...
2	Virginia	problem,Couple,room Business trip,location Bus...	{Little_Joe, persist, ii, placement, state, bi...
3	Kannan	Solo traveller,stay,Business trip,Stayed 4 nig...	{Little_Joe, Jnr, persist, stick, „ stayed, t...
4	Sue	Couple,owner Leisure trip,Standard Queen Room,...	{Little_Joe, persist, ii, avail, drinking, pla...

FIGURE 28 – Expansion des synonymes des aspects

	reviewed_by	aspects	expanded_aspects
0	Kyrylo	Solo traveller,Stayed 1 night,,Business trip,J...	{Jnr, persist, stayed, „ stick, I, take, stay...
1	Dimitri	Deluxe Suite,Executive Double Room with Sofa B...	{residue, persist, ii, jazz, state, get_laid, ...
2	Virginia	Couple,room Business trip,location Bus...	{Little_Joe, persist, ii, placement, state, bi...
3	Kannan	Solo traveller,stay,Business trip,Stayed 4 nig...	{Little_Joe, Jnr, persist, stick, „ stayed, t...
4	Sue	Couple,owner Leisure trip,Standard Queen Room,...	{Little_Joe, persist, ii, avail, drinking, pla...

FIGURE 29 – Expansion des synonymes des intérêts

- **Calcul de la similarité cosinus :** Nous avons utilisé le modèle Word2Vec pré-entraîné pour obtenir les vecteurs des mots et calculer la similarité cosinus entre les vecteurs moyens des aspects et des intérêts.
- **Calcul de la matrice de similarité :** Lors de la construction de la matrice de similarité, nous avons pris en compte les aspects mentionnés par les utilisateurs, les descriptions des hôtels, et les évaluations des autres utilisateurs. Chaque entrée de cette matrice représente la similarité cosinus entre les aspects d'un utilisateur et les centres d'intérêt d'un cluster.

#### Affectation des clusters aux utilisateurs :

Nous avons déterminé les clusters d'utilisateurs partageant des intérêts similaires, en attribuant à chaque utilisateur un cluster en fonction de la similarité de leurs intérêts avec ceux des autres utilisateurs du même cluster. Chaque utilisateur est associé au cluster présentant la plus grande similarité avec ses intérêts.

#### Fusion des DataFrames et Agrégation des données par cluster :

La fusion des DataFrames consiste à intégrer les clusters assignés avec les commentaires d'hôtels, permettant ainsi de créer un ensemble de données complet. Ensuite, ce nouvel ensemble de données est soumis à une agrégation par cluster, où les données sont regroupées par cluster et par hôtel. Cette agrégation permet d'obtenir des statistiques agrégées telles que le nombre d'occurrences de l'hôtel par cluster et les moyennes des scores de sentiments prédits des commentaires.

username	aspects	assigned_cluster	hotel_name	hotel_url	predicted_score
0 Kyrylo	Solo traveller.Stayed 1 night.Business trip.J...	25	Villa Pura Vida	<a href="https://www.booking.com/hotel/be/villa-pura-v...">https://www.booking.com/hotel/be/villa-pura-v...</a>	0.999794
1 Dimitri	Deluxe Suite,Executive Double Room with Sofa B...	19	Villa Pura Vida	<a href="https://www.booking.com/hotel/be/villa-pura-v...">https://www.booking.com/hotel/be/villa-pura-v...</a>	0.995007
2 Dimitri	Deluxe Suite,Executive Double Room with Sofa B...	19	Novotel Gent Centrum	<a href="https://www.booking.com/hotel/be/novotelgenten...">https://www.booking.com/hotel/be/novotelgenten...</a>	0.993026
3 Dimitri	Deluxe Suite,Executive Double Room with Sofa B...	19	Hotel Dolce La Hulpe Brussels	<a href="https://www.booking.com/hotel/be/dolce-la-hulp...">https://www.booking.com/hotel/be/dolce-la-hulp...</a>	0.493406
4 Dimitri	Deluxe Suite,Executive Double Room with Sofa B...	19	B&B Le flaneur	<a href="https://www.booking.com/hotel/be/le-flaneur.en...">https://www.booking.com/hotel/be/le-flaneur.en...</a>	0.984048
...	...	...	...	...	...
26381 Aymar	Stayed 1 night.Couple room.Double Room Submitt...	25	Hotel Du Congres	<a href="https://www.booking.com/hotel/be/ducongres.en...">https://www.booking.com/hotel/be/ducongres.en...</a>	0.520505
26382 Chrizter	Studio (2 Adults).Stayed 1 night.Family with ...	44	Aparthotel Adagio Access Bruxelles Europe Apart...	<a href="https://www.booking.com/hotel/be/adagio-access...">https://www.booking.com/hotel/be/adagio-access...</a>	0.598312
26383 Olya	Solo traveller.place.Stayed 3 nights.Double Ro...	25	Logies Windsor	<a href="https://www.booking.com/hotel/be/windsor-cast...">https://www.booking.com/hotel/be/windsor-cast...</a>	0.599799
26384 Subodh	Solo traveler.Standard Double Room.Stayed 2 ...	2	Hotel Ter Eet	<a href="https://www.booking.com/hotel/be/best-western...">https://www.booking.com/hotel/be/best-western...</a>	0.566089
26385 Zhi	Couple.Stayed 2 nights.Standard Double or Twin...	2	NH Brussels Carrefour L'Europe	<a href="https://www.booking.com/hotel/be/carrefour.on...">https://www.booking.com/hotel/be/carrefour.on...</a>	0.787512

20346 rows × 6 columns

FIGURE 30 – Fusion des DataFrames et Agrégation des données par cluster

### III.4.2 Recommandation

La fonction de recommandation trouve le cluster assigné à l'utilisateur spécifié en recherchant son nom d'utilisateur dans le DataFrame contenant les différents utilisateurs et leur cluster associé. Une fois le cluster de l'utilisateur identifié, elle filtre le DataFrame contenant les noms d'hôtels pour obtenir les hôtels appartenant au même cluster. Finalement, elle extrait les noms des hôtels des nb meilleurs hôtels possédant la meilleure moyenne de score correspondant à ce cluster, et les retourne dans une liste.

Cette approche permet de recommander des hôtels en se basant sur les préférences groupées par cluster des utilisateurs. Grâce à cette fonction, pour chaque utilisateur, nous avons identifié les hôtels les mieux notés et les plus pertinents au sein de son cluster en prenant en compte les évaluations des utilisateurs du même cluster et les aspects mentionnés dans leurs avis.

## Conclusion

Ce chapitre a détaillé les étapes de modélisation et d'évaluation de notre système de recommandation d'hôtels. Le processus de matching et d'assignation des utilisateurs au cluster le plus approprié a permis de combiner deux analyses clés : l'analyse sentimentale des avis des utilisateurs ayant séjourné dans des hôtels et le regroupement des profils des utilisateurs Instagram les plus similaires . Cette intégration a abouti à la formation de deux datasets finaux :Un dataset contenant les noms d'hôtels, leurs URLs, les clusters assignés, et les moyennes de scores attribuées par cluster et par hôtel et un dataset répertoriant les utilisateurs et leurs clusters respectifs. Ces datasets ont permis de développer une recommandation d'hôtels pertinente et personnalisée, basée sur les intérêts spécifiques de chaque utilisateur. Dans le cinquième chapitre et dernier chapitre, nous aborderons le déploiement de notre système de recommandation. .

## IV Déploiement

### Introduction

Le long de ce chapitre nous allons aborder le déploiement de notre système de recommandation. Tout d'abord, nous présenterons les technologies utilisées. Ensuite, nous explorerons la visualisation avec l'interface principale des hôtels, l'authentification des utilisateurs et la page de recommandations personnalisées. Enfin, nous évaluerons les résultats obtenus, démontrant l'efficacité de notre système à répondre aux préférences des utilisateurs. Ce chapitre illustre notre démarche pour rendre notre modèle accessible et convivial aux utilisateurs finaux.

### IV.1 Technologies utilisés

#### IV.1.1 Framework Django Pour Le Développement

Afin de mettre à disposition des utilisateurs notre système de recommandation et d'analyse des sentiments, nous avons créé un site web en utilisant Django [7], un framework web basé sur Python. Le site propose diverses options pour améliorer l'expérience des utilisateurs en offrant des recommandations sur mesure d'hôtels.

#### IV.1.2 Intégration de Flask pour les Requêtes de Recommandation

Pour la génération des recommandations, nous avons utilisé Flask [8], un autre framework web en Python, pour faire le lien avec notre modèle de recommandation hébergé sur Google Colab. Voici comment cela fonctionne :

- Flask API : Nous avons développé une API Flask qui déploie le modèle de recommandation hébergé sur Google Colab.
- Appel de l'API depuis Django : Django envoie des requêtes à cette API Flask pour obtenir les recommandations en temps réel.
- Retour des Recommandations : Les résultats des recommandations sont renvoyés à Django, qui les affiche ensuite à l'utilisateur.

## IV.2 Visualisation

### IV.2.1 Interface Principale des Hôtels

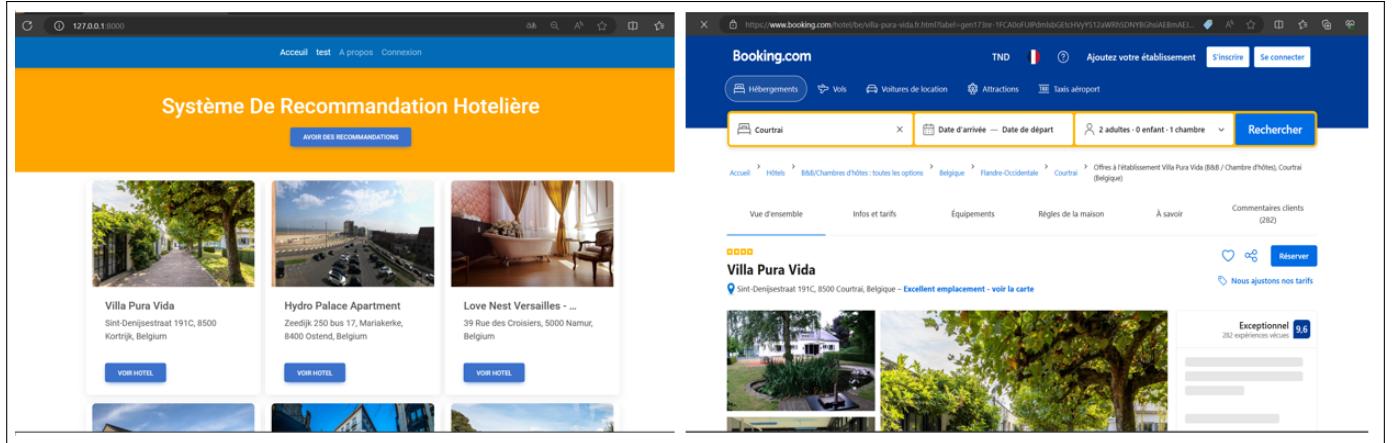


FIGURE 31 – Interface Principale des Hôtels

- L'interface initiale affiche une liste d'hôtels à disposition.
- Chaque hôtel offre un lien direct vers sa page sur Booking.com, offrant ainsi aux utilisateurs la possibilité d'obtenir davantage d'informations ou de réserver en un seul clic.

### IV.2.2 Authentification des Utilisateurs

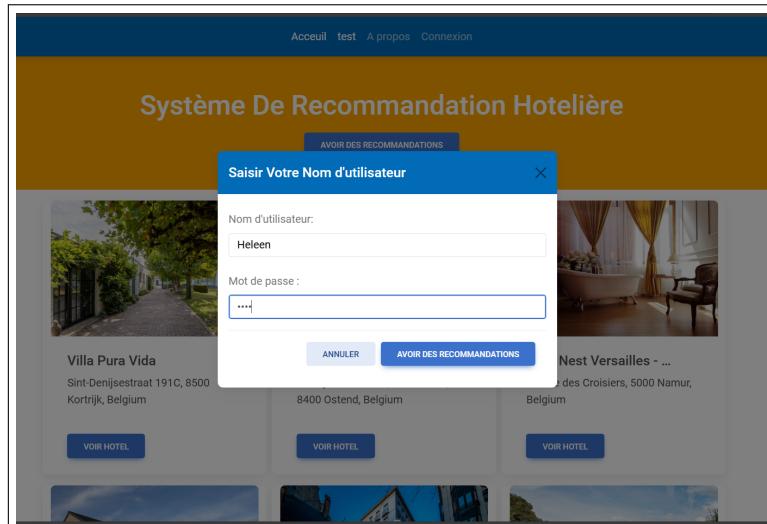


FIGURE 32 – Interface Authentification des Utilisateurs

- Pour accéder aux recommandations, les utilisateurs doivent se connecter en utilisant leur nom d'utilisateur et leur mot de passe.

#### IV.2.3 Page de Recommandations Personnalisées

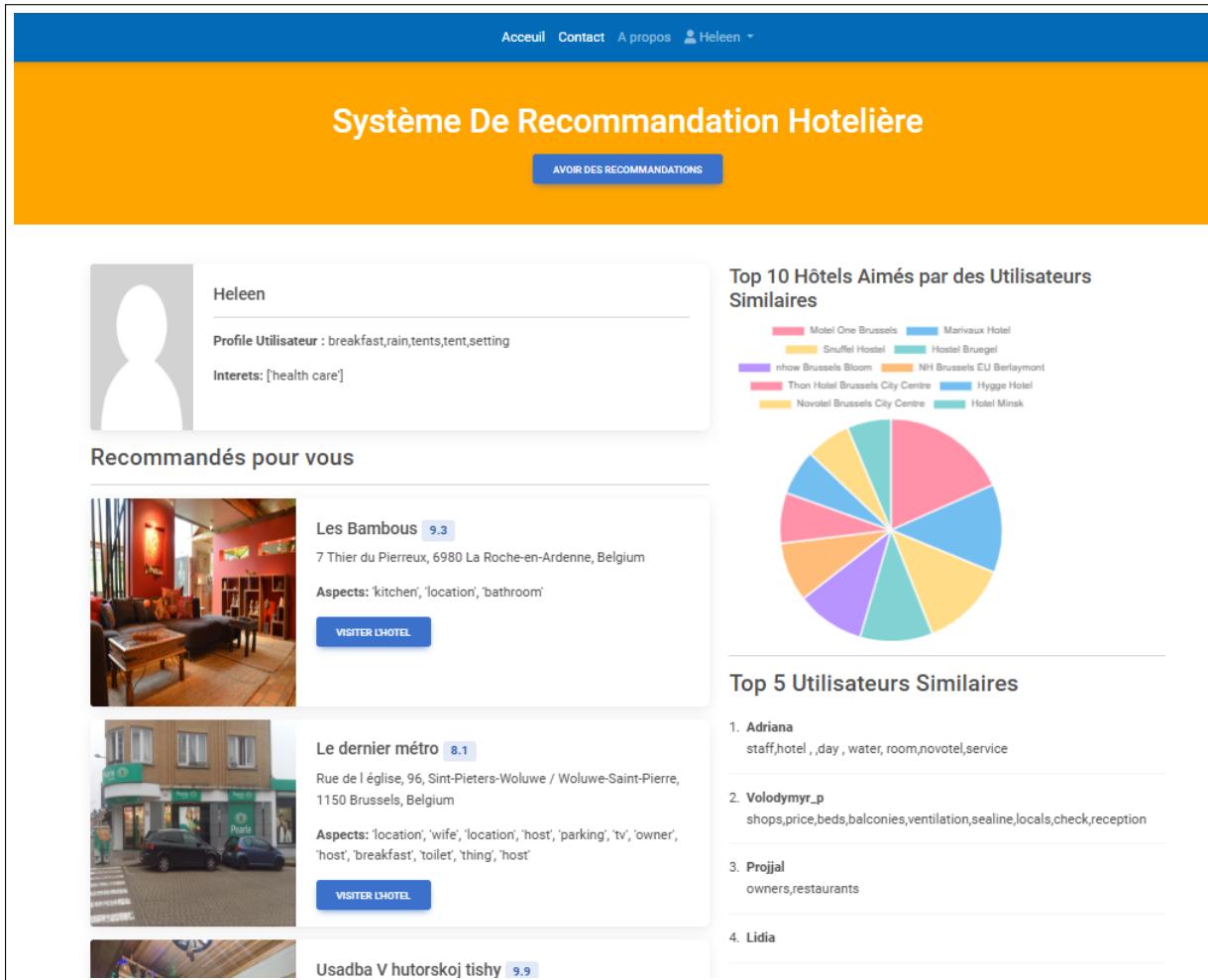


FIGURE 33 – Page de Recommandations Personnalisée

- Cette page affiche une liste d'hôtels recommandés spécifiquement pour l'utilisateur authentifié.
- Le système utilise des algorithmes de clustering pour regrouper les utilisateurs ayant des intérêts similaires et fournir des recommandations plus précises.
- Outre les recommandations, un graphique en pie chart (camembert) présente les hôtels les plus populaires parmi des utilisateurs similaires, offrant ainsi une vision d'ensemble des tendances parmi des groupes d'utilisateurs partageant des préférences similaires.
- On affiche également les 5 utilisateurs les plus semblables au sein du même cluster, ce qui permet à l'utilisateur de voir avec qui il partage des préférences.

#### **IV.2.4 Évaluation des résultats**

Notre système de recommandation d'hôtels semble fonctionner de manière satisfaisante, avec des résultats qui correspondent aux préférences générales de l'utilisateur "Heleen" de la figure 33. En examinant les aspects importants pour Heleen tels que le petit-déjeuner, la localisation et l'ambiance générale, nous avons constaté que nos recommandations s'alignent étroitement avec ces critères. Par exemple, dans le cas de "Les Bambous", bien que le terme "kitchen" ne corresponde pas directement à "breakfast", nous avons pu établir une correspondance en raison de l'association de la cuisine avec la préparation des repas du matin. De même, "Le dernier métro" a été une recommandation forte, en offrant non seulement un petit-déjeuner, mais aussi d'autres aspects contribuant à une expérience globale agréable, tels que la localisation et l'hospitalité.

De plus, en répondant à l'intérêt spécifique de Heleen pour le "health care", ces hôtels offrent une expérience holistique qui prend en compte non seulement son confort et sa satisfaction, mais aussi sa santé et son bien-être général pendant son séjour.

## **Conclusion**

Dans ce dernier chapitre, nous avons évalué dans quelle mesure notre modèle de recommandation et d'analyse des sentiments répond aux objectifs fixés. Nous avons ensuite présenté les principales interfaces et les choix techniques utilisés pour développer notre site web et déployer notre modèle permettant de rendre ainsi l'application accessible et conviviale pour les utilisateurs finaux .

# Conclusion et perspectives

Durant ce projet, nous avons présenté une approche exhaustive pour l'analyse des sentiments dans les avis sur les hôtels et la recommandation personnalisée d'hôtels aux utilisateurs. En exploitant des techniques avancées de traitement du langage naturel (NLP) et des modèles de réseaux de neurones, nous avons considérablement amélioré la précision de l'analyse sentimentale par rapport aux méthodes traditionnelles telles que VADER. Notamment, le modèle BiLSTM avec les embeddings GloVe s'est démarqué avec une précision globale de 93,44

En parallèle, la mise en place d'un clustering CAH des utilisateurs basé sur leurs profils et intérêts et la recommandation d'hôtels en fonction de ces clusters ont permis de fournir des recommandations plus précises et pertinentes.

En combinant l'analyse des sentiments avec le clustering des utilisateurs, nous avons pu proposer des recommandations pertinentes en prenant en compte à la fois leurs préférences spécifiques et la similarité entre les utilisateurs appartenant au même cluster.

Le déploiement de notre système de recommandation dans une interface conviviale grâce à Django et Flask offre une expérience utilisateur améliorée, permettant aux utilisateurs d'accéder facilement aux recommandations personnalisées.

Pour améliorer les performances de notre système de recommandation et recommander des hôtels plus pertinents et personnalisés. Plusieurs perspectives peuvent être envisagées :

**Améliorer l'analyse sentimentale et l'extraction d'aspects :** L'utilisation de l'approche d'apprentissage multitâche permettrait d'apprendre simultanément plusieurs tâches, ce qui peut améliorer la performance des deux tâches.

**Exploiter des techniques d'apprentissage multitâche :** L'apprentissage multitâche permettrait d'apprendre simultanément plusieurs tâches, telles que l'analyse sentimentale et l'extraction d'aspects, ce qui pourrait améliorer la performance des deux tâches.

**Enrichir le système de recommandation :** On peut prendre en compte d'autres critères comme les photos, les descriptions d'hôtels, les avis des experts, etc.

**Évaluer le système à grande échelle :** Des tests utilisateurs permettraient d'évaluer l'efficacité du système dans un contexte réel et d'identifier d'éventuels problèmes d'utilisabilité.

**Personnalisation dynamique :** Par l'apprentissage en continu, le système peut apprendre en permanence des interactions des utilisateurs avec les recommandations et affiner ses modèles en conséquence.

**Exploration d'autres techniques de modélisation :** On peut recourir à la méthode d'Apprentissage par renforcement. Cette approche permet de modéliser l'interaction entre le système et les utilisateurs et d'optimiser les recommandations en fonction des actions et des réactions des utilisateurs.

# Bibliographie

## Références

- [1]. URL <https://towardsdatascience.com/brief-on-recommender-systems-b86a1068a4dd>.
- [2]. URL [https://scikit-learn.org/stable/modules/generated/sklearn.feature\\_extraction.text.TfidfVectorizer](https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer).
- [3]. URL <https://towardsdatascience.com/an-short-introduction-to-vader-3f3860208d53>.
- [4]. URL <https://www.datasciencetoday.net/index.php/fr/machine-learning/148-reseaux-neuronaux-recurrents-et-lstm#:~:text=Les%20r%C3%A9seaux%20de%20longue%20m%C3%A9moire,tr%C3%A8s%20longs%20entre%20les%20deux>.
- [5]. URL <https://nlp.stanford.edu/projects/glove/>.
- [6]. URL <https://www.tensorflow.org/text/tutorials/word2vec>.
- [7]. URL <https://docs.djangoproject.com/en/5.0/>.
- [8]. URL <https://flask.palletsprojects.com/en/3.0.x/>.
- [9]. URL <https://www.nltk.org/howto/wordnet>.

Robin Burke. Hybrid recommender systems : Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4) :331–370, 2002.

## Résumé

Ce rapport présente une approche complète pour la mise en place d'un système de recommandation hôtelière hybride basé sur l'analyse sentimentale des avis sur les hôtels et les profils des utilisateurs, en exploitant leur similarité.

Développé dans le cadre du Projet de Fin d'Année à l'INSAT pour l'année universitaire 2023-2024, le projet exploite des techniques avancées de traitement du langage naturel (NLP) et des modèles de réseaux de neurones pour améliorer la précision de l'analyse sentimentale et proposer des recommandations d'hôtels personnalisées et pertinentes.

**Mots clés :** Hôtels, Recommandation hybride, Avis, Profil utilisateur, Similarité Cosinus, Aspects, Analyse sentimentale, Réseaux de neurones, BiLSTM, Word2Vec, TF-IDF, GloVE, Clustering CAH, NLP

## Abstract

This report presents a comprehensive approach to implementing a hybrid hotel recommendation system based on sentimental analysis of hotel reviews and user profiles, leveraging their similarity.

Developed as part of the End-of-Year Project at INSAT for the 2023-2024 academic year, the project leverages advanced natural language processing (NLP) techniques and neural network models to improve the accuracy of sentiment analysis and provide personalized and relevant hotel recommendations.

**Keywords:** Hotels, Hybrid Recommendation, Reviews, User Profile, Cosine Similarity, Aspects, Sentiment Analysis, Neural Networks, BiLSTM, Word2Vec, TF-IDF, GloVE, CAH Clustering, NLP.

## ملخص

يقدم هذا التقرير نهجاً شاملأً لتنفيذ نظام توصية فندقية هجين يعتمد على تحليل المشاعر في مراجعات الفنادق والملفات الشخصية للمستخدمين، من خلال استغلال تشابههم

تم تطوير هذا المشروع كجزء من مشروع نهاية السنة في المعهد الوطني للعلوم التطبيقية والتكنولوجيا للسنة الجامعية 2023-2024، ويستفيد المشروع من تقنيات متقدمة لمعالجة اللغة الطبيعية ونمذج الشبكات العصبية لتحسين دقة تحليل المشاعر وتقديم توصيات فندقية مخصصة وملائمة

**الكلمات المفتاحية:** فنادق، توصية هجينة، مراجعات، ملف المستخدم، تشابه جيد الناتم، الجوانب، تحليل المشاعر، الشبكات العصبية، معالجة اللغة الطبيعية (NLP)، تجميع CAH ، TF-ID ， Word2Vec ، BiLSTM