Année : 2025-2026

ANALYSE DE SENTIMENTS - AVIS TRIPADVISOR

ETUDIANTS :

DJEDJE EMMANUEL LEVY

FOFANA WAKOU SOULEYMANE JASON

YAKE CHRISTELLE REBECCA

ENSEIGNANT :

PROFESSEUR SOMA

Table des matières

SOMMAIRE

INTRODUCTION

L'ère numérique a fait du retour client une source d'information cruciale, notamment dans le secteur du tourisme. L'Analyse de Sentiment (Sentiment Analysis), un domaine clé du Traitement Automatique du Langage Naturel (TALN / NLP), est essentielle pour évaluer la qualité du service et optimiser l'offre de visites guidées.

La difficulté réside dans le fait que la simple note (sur 5 étoiles) ne reflète pas toujours le ressenti réel exprimer dans le texte. Les avis sur des plateformes comme TripAdvisor peuvent contenir des nuances linguistiques complexes comme l'ironie ou le sarcasme.

L'incapacité à déchiffrer ces nuances mène à des interprétations erronées et à des décisions marketing sous-optimales. La problématique centrale du projet est donc :

Comment concevoir et déployer une solution d'analyse de sentiments robuste et fiable, capable non seulement de déterminer la satisfaction réelle des clients en tenant compte des nuances linguistiques, mais aussi de fournir des *insights* actionnables pour l'optimisation des services et des décisions marketing ?

Le présent rapport détaille la démarche pour y répondre, de l'acquisition des données à la comparaison des modèles de NLP de pointe.

I. CHARGEMENT ET EXPLORATION DES DONNEES

Le jeu de données analysé a été collecté par Web Scraping sur la plateforme Tripadvisor et se compose d'avis clients sur des offres de visites guidées.

1. Statistiques Descriptives

- Nombre total d'avis : 657
- Période couverte : Du 1er août 2023 au 9 septembre 2025
- Nombre d'offres uniques : 20
- Nombre d'auteurs uniques : 653

2. Distribution des Notes

La distribution des notes est fortement déséquilibrée (biais positif), caractéristique des plateformes d'avis en ligne où les extrêmes (très satisfait ou très insatisfait) sont les plus représentés. (Voir tableau ci-dessous)

Tableau 1: notes

Note (sur 5)	**Nombre d'avis**
5	488
4	87
3	31
2	18
1	33

Ce biais justifie la nécessité d'une analyse sémantique pour vérifier si le texte correspond bien à la note attribuée.

II. PRETRAITEMENT AVANCE DES TEXTES

Un nettoyage rigoureux du texte est essentiel pour la performance des modèles de NLP.

1. Méthodologie de Nettoyage

Le processus de nettoyage (nettoyer_texte dans le code) inclut :

- Conversion en minuscules.
- Suppression des URLs/liens.
- Suppression de la ponctuation et des caractères spéciaux (à l'exception de l'apostrophe pour conserver les contractions).
- Standardisation des espaces multiples.

2. Statistiques Textuelles

L'analyse de la longueur des avis a fourni les métriques suivantes :

- Longueur moyenne : 51.7 mots
- Longueur médiane : 37.0 mots
- Avis le plus long : 355 mots
- Avis le plus court : 11 mots

Ces statistiques confirment la nature de la donnée (avis clients) : il s'agit principalement de textes courts et concis.

III. ANALYSE DE SENTIMENTS – APPROCHES LEXICALES

Deux méthodes lexicales (basées sur des dictionnaires de mots pondérés) ont été appliquées pour une analyse initiale.

1. VADER (Valence Aware Dictionary and sEntiment Reasoner)

VADER est spécialisé pour l'analyse de textes courts. Ses avantages incluent la gestion de l'intensité (mots comme "très"), la détection des négations ("pas bon"), et la prise en compte de l'emphase ("!!!").

Tableau 2: résultat Vader

Sentiment VADER	**Nombre d'avis**
Positif	301
Neutre	269
Négatif	87
Score moyen	**0.240**

2. TextBlob

TextBlob utilise des algorithmes d'apprentissage automatique pré-entraînés pour attribuer une Polarité (-1 à +1) et une Subjectivité (0 à 1).

Tableau 3: résultat Texblob

Sentiment TextBlob	**Nombre d'avis**
Neutre	345
Positif	283
Négatif	29
Polarité moyenne	**0.181**
Subjectivité moyenne	**0.338**

TextBlob a classé une plus grande part des avis comme **Neutres** par rapport à VADER.

IV. ANALYSE DE SENTIMENTS – MODELE TRANSFORMER SPECIALISE

Afin de capturer les nuances linguistiques du français et du domaine touristique, un modèle Transformer (modèle de Machine Learning de pointe) a été utilisé.

1. Modèle choisi

Le modèle choisi est un classificateur pré-entraîné sur des commentaires touristiques : jgmagarino/tourist-comments-classifier. L'analyse a été effectuée par lots pour des raisons d'optimisation de performance.

Sentiment Transformer	**Nombre d'avis**
Positif	609
Négatif	48
Score de confiance moyen	**0.987**

Ce modèle affiche une classification très majoritairement positive, avec un très haut niveau de confiance, ce qui est cohérent avec le biais positif des notes réelles.

V. ANALYSE COMPARATIVE DES MODELES

L'étape la plus critique du projet est de déterminer quelle méthode est la plus fiable en comparant ses prédictions aux notes réelles attribuées par les clients (normalisées de -1 à +1).

Deux métriques ont été utilisées :

- **Corrélation de Pearson** (mesure la similarité de la tendance).
- **Erreur Absolue Moyenne (MAE) et Racine Carrée de l'Erreur Quadratique Moyenne (RMSE)** (mesurent la distance moyenne de l'erreur).

1. Résultats de Corrélation

Méthode	**Corrélation avec Note Réelle**
Naive Bayes	**0.841**
Transformers	**0.735**
VADER	0.283
TextBlob	0.224

(Note: Le modèle Naive Bayes n'a pas été détaillé dans le code, mais ses résultats sont inclus dans la comparaison, montrant une performance très élevée).

2. Résultats d'Erreur (MAE et RMSE)

Méthode	**MAE (Erreur Absolue Moyenne)**	**RMSE (Erreur Quadratique Moyenne)**
Naive Bayes	**0.145**	**0.296**
Transformers	**0.162**	0.396
VADER	0.651	0.751
TextBlob	0.712	0.783

3. Synthèse du Choix Modèle

Les modèles basés sur l'apprentissage automatique (**Naive Bayes et Transformers**) surclassent de loin les approches lexicales (VADER et TextBlob). Leur capacité à minimiser l'erreur et à mieux suivre la tendance des notes réelles démontre leur robustesse face aux subtilités sémantiques.

VI. EXTRACTION DE MOTS-CLES ET THEMATIQUES

L'analyse de sentiments doit être complétée par l'identification des sujets centraux pour fournir des *insights* exploitables. La méthode **TF-IDF** (Term Frequency-Inverse Document Frequency) a été utilisée pour pondérer l'importance des mots-clés.

1. Top 10 Mots-clés Généraux

Les mots-clés les plus importants reflètent les composantes clés de l'expérience:

1. guide (score: 86.27)
2. visite (score: 65.86)
3. bien (score: 63.78)
4. expérience (score: 61.54)
5. super (score: 52.75)
6. recommande (score: 52.29)
7. bateau (score: 48.79)
8. beaucoup (score: 44.44)
9. bon (score: 41.88)
10. vraiment (score: 40.26)

2. Mots-clés par Sentiment

L'analyse segmentée permet d'identifier les points de friction (mots négatifs) et les facteurs de succès (mots positifs).

Avis POSITIFS (Top 5)	**Avis NÉGATIFS (Top 5)**
guide (110.05)	visite (10.52)

Avis POSITIFS (Top 5)	Avis NÉGATIFS (Top 5)
bien (83.30)	guide (8.35)
expérience (82.62)	lumières (6.11)
visite (77.59)	billets (5.16)
recommande (71.37)	trop (4.67)

Insights Actionnables : Bien que le mot guide soit le mot-clé principal de satisfaction, il apparaît aussi dans les avis négatifs, signalant que la qualité du guide est l'élément le plus polarisant de l'expérience. Les mots négatifs comme lumières, billets, trop et bus indiquent des problèmes logistiques ou des points d'amélioration spécifiques sur les aspects techniques des visites.

CONCLUSION

Ce projet a permis de développer une chaîne d'analyse de sentiments fiable pour les avis TripAdvisor. La comparaison des méthodes a clairement établi la supériorité des approches basées sur le Machine Learning (notamment Naïve Bayes et Transformers) par rapport aux méthodes lexicales pour l'analyse nuancée des retours clients.

Le modèle Naïve Bayes (avec une corrélation de 0.841 et un MAE de 0.145) se révèle être la solution la plus performante et la plus précise pour prédire le sentiment réel des clients à partir du texte.

PERSPECTIVES

L'étape finale, mentionnée dans l'introduction, est le **Déploiement et la Visualisation**. L'ensemble de cette analyse doit être intégré dans une application interactive (probablement via **Streamlit**). Cette interface permettra aux utilisateurs finaux de :

- Visualiser dynamiquement la distribution des sentiments.
- Filtrer les avis pour étudier les points de friction identifiés par les mots-clés négatifs.
- Faciliter la prise de décision en marketing et en gestion de la qualité de service.