Analyse des Sentiments des Commentaires sur les hotels en Anglais

École Nationale des Sciences Appliquées, Al Hoceima
Domaine : Ingénierie des données
Auteur : AGHZAR Otmane
Superviseur : KHAMJANE Aziz

RÉSUMÉ

L'analyse des sentiments, une branche du traitement automatique du langage naturel (NLP), a gagné en popularité ces dernières années en raison de ses nombreuses applications, telles que la surveillance des réseaux sociaux, l'analyse des avis clients et les études de marché. Ce projet se concentre sur l'analyse des sentiments en anglais, en s'appuyant sur des techniques d'apprentissage automatique. L'objectif est de concevoir un modèle performant capable de comprendre et de classifier les sentiments exprimés dans des textes en langue anglaise.

INTRODUCTION

Bien que de nombreuses recherches aient été menées sur l'analyse des sentiments pour des langues principales comme l'anglais, des défis persistent dans la capture précise des nuances émotionnelles, notamment dans des contextes spécifiques tels que les avis de voyageurs. L'anglais, utilisé dans un large éventail de styles et de contextes, présente une diversité linguistique et culturelle qui nécessite des modèles adaptés pour une analyse efficace des sentiments. Ce projet se concentre sur le développement d'un modèle d'analyse des sentiments pour l'anglais, capable de traiter des textes variés avec précision.

Les données ont été collectées sur le site TripAdvisor, fournissant une base riche et variée d'avis de voyageurs. Ces propres à ce contexte. Les données ont été annotées manuellement propres à ce contexte. pour garantir une évaluation rigoureuse des sentiments. Plusieurs algorithmes d'apprentissage automatique ont été testés pour identifier l'approche la plus performante. Ce projet contribue à étendre les techniques d'apprentissage automatique pour analyser les sentiments exprimés dans des contextes réels, tout en offrant des applications pratiques dans des domaines tels que l'industrie du tourisme, évaluation

TRAVAUX RELATIVES

Bien que l'analyse des sentiments ait été largement étudiée pour les langues principales comme l'anglais, les recherches axées sur des contextes spécifiques, tels que les avis de voyageurs, restent relativement limitées. La littérature existante se concentre souvent sur des textes standardisés et formels, en négligeant les défis uniques posés par les contenus générés par les utilisateurs, qui sont souvent informels, dépendants du contexte et variés dans leur ton. Des études notables incluent l'analyse des sentiments sur les plateformes de réseaux sociaux et les avis de produits, mettant en évidence l'efficacité des modèles d'apprentissage automatique pour capturer les nuances émotionnelles.

Les avancées récentes en apprentissage automatique, en particulier les modèles de transfert comme BERT (Représentations Encodeurs Bidirectionnels par Transformeurs), ont permis d'obtenir des résultats de pointe dans l'analyse des sentiments à travers plusieurs langues et domaines. S'appuyant sur ces progrès, notre recherche se concentre spécifiquement sur l'analyse des sentiments des avis de voyageurs en anglais, tout en abordant propres à ce contexte.

DESCRIPTION DE L'ENSEMBLE DE DONNÉES

1. **Collecte des données**

Le projet d'analyse des sentiments des commentaires sur les hôtels en anglais utilise un ensemble de données diversifié compilé à partir de TripAdvisor pour capturer les caractéristiques linguistiques uniques et les expressions de sentiments dans les commentaires des utilisateurs. L'ensemble de données comprend des commentaires collectés à partir de différents hôtels, garantissant un échantillon représentatif des avis des utilisateurs. Les commentaires sont associés à des notes (par exemple, de 1 à 5 étoiles), fournissant des données étiquetées précieuses pour l'analyse des sentiments.

Apify est une plateforme cloud dédiée au web scraping, à l'automatisation de navigateur et à la collecte de données pour l'intelligence artificielle. Elle propose de nombreux outils prêts à l'emploi. Nous avons utilisé un acteur nommé "TripAdvisor Scraper reviews" pour récupérer les avis sur les hôtels postés sur TripAdvisor. Cet outil offre un crédit de 5$ gratuitement qui se regenere chaque mois, Le cout de scraping de 1000 commentaires est de 2$. Bien qu'il existe une API officielle de TripAdvisor, elle présente des limitations en termes de volume de données accessibles par appel. Au total, 2 646 avis ont été collectés.

2. Data Preprocessing

Le prétraitement des données joue un rôle crucial dans le succès des modèles d'analyse des sentiments. Dans cette section, nous décrivons les étapes effectuées pour nettoyer et préparer l'ensemble de données pour un entraînement efficace des modèles.

- Nettoyage de base

La phase initiale de prétraitement des données implique des opérations de nettoyage fondamentales. La fonction basic_cleaning a été utilisée pour garantir l'intégrité et la qualité de l'ensemble de données. Plus précisément, les étapes suivantes ont été effectuées :

✓ Suppression des lignes avec des valeurs manquantes dans la colonne de données textuelles.
✓ Exclusion des lignes avec des chaînes vides dans la colonne de données textuelles.
✓ Élimination des lignes en double pour éviter les biais dans l'entraînement des modèles.
✓ Suppression des caractéristiques.

Certaines colonnes ont été jugées non contributives à la tâche d'analyse des sentiments et ont été supprimées de l'ensemble de données à l'aide de la fonction advanced_clean_text.

Cette étape visait à simplifier l'ensemble de données et à éliminer les informations inutiles qui pourraient potentiellement introduire du bruit pendant l'entraînement des modèles.

- Nettoyage avancé

La fonction advanced_clean_text orchestre un pipeline complet de processus de traitement du texte, y compris :

✓ Suppression de la ponctuation.
✓ Suppression des nombres et des symboles spécifiques.
✓ Suppression des espaces multiples.
✓ Suppression des caractères répétés pour une représentation plus propre.

- Suppression des Stopwords

Une fonction personnalisée remove_stepword_from_text est utilisée pour supprimer les mots vides ont été compilés à partir de diverses sources, y compris une liste manuellement curée, des listes de mots vides thématiques spécifiques, et des mots vides courants en anglais. Les mots vides thématiques ont été sélectionnés pour des catégories spécifiques telles que les nombres, la nourriture, les vêtements, les couleurs, les pronoms et les adverbes. Cela enrichit la liste des mots vides avec des termes contextuellement pertinents qui ne sont pas couverts par les listes de mots vides génériques. L'intégration de ces étapes de nettoyage avancé garantit que l'ensemble de données est soigneusement traité, capturant les subtilités des expressions de sentiments dans les commentaires sur les hôtels tout en abordant des caractéristiques linguistiques et contextuelles spécifiques. Cet ensemble de données raffiné est prêt pour une analyse et un entraînement de modèles ultérieurs dans les étapes suivantes du projet d'analyse des sentiments.

- **Le stemming des données**

Dans le contexte du traitement automatique du langage naturel (NLP), le stemming et la lemmatisation sont des techniques utilisées pour réduire les mots à leur forme de base ou racine,

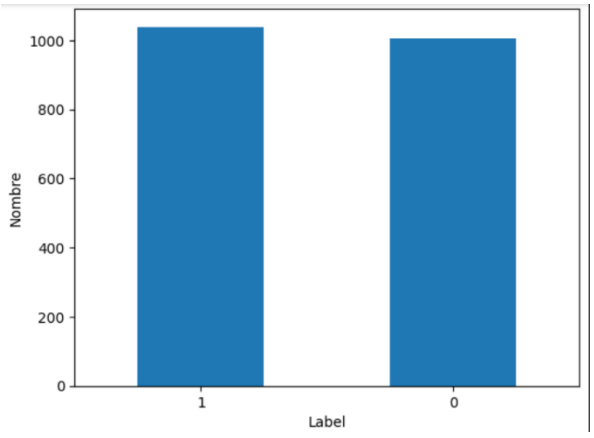
simplifiant ainsi les données textuelles pour l'analyse. Ces techniques sont particulièrement utiles dans les tâches d'analyse des sentiments pour améliorer l'efficacité computationnelle et réduire la dimensionnalité de l'ensemble de données.

Dans ce projet, le processus de stemming et de lemmatisation a été implémenté en utilisant les bibliothèques PorterStemmer et WordNetLemmatizer de NLTK. La fonction lemmatize_and_stem_text a été créée pour appliquer ces deux techniques sur une chaîne de texte complète. Cette fonction divise d'abord le texte en mots, applique la lemmatisation à chaque mot à l'aide de WordNetLemmatizer, puis applique le stemming à l'aide de PorterStemmer. Les mots traités sont ensuite reconstitués en une phrase cohérente.

Cette séquence d'opérations a été appliquée à la colonne text_no_stopwords du DataFrame, et les résultats ont été stockés dans une nouvelle colonne clean_text. Cette approche permet de nettoyer et de simplifier les données textuelles, préparant ainsi l'ensemble de données pour les étapes ultérieures d'analyse et d'entraînement des modèles d'analyse des sentiments.

- **Résumé de l'ensemble de données**

Le DataFrame a ensuite été enregistré dans un fichier CSV ('output.csv') pour référence future. L'ensemble de données final comprend 2646 lignes après avoir supprimé les lignes contenant des valeurs manquantes dans la colonne 'clean_text', offrant ainsi une représentation concise mais linguistiquement informée des expressions de sentiments dans les commentaires sur les hôtels en anglais.



L'ensemble de données se distingue par sa diversité inhérente, capturant une riche variété de variations linguistiques et de nuances thématiques propres aux commentaires sur les hôtels en anglais. L'ensemble de données intègre à la fois des expressions informelles et formelles, ainsi que des expressions spécifiques au secteur hôtelier, et couvre également un large éventail de sujets et de contextes. Cette diversité offre une base robuste et complète au projet d'analyse des sentiments. Elle prend en compte les variations stylistiques et les idiomes spécifiques aux commentaires des utilisateurs, mettant en évidence les subtilités linguistiques souvent observées dans les avis sur les hôtels. Parallèlement, il inclut des exemples d'expressions authentiques, permettant ainsi d'examiner de plus près les opinions et les attitudes des utilisateurs à l'égard des services hôteliers.

	text	label	clean_text
0	clean and quiet hotel, we booked bed and break...	1	clean quiet hotel, book bed breakfast breakfas...
1	on checking in our luggage- it failed to arri...	1	check luggage- fail arriv hour went collect it...
2	excellent staff, always smiling and friendly. ...	1	excel staff, alway smile friendly. outstand ho...
3	great bed and rooms, renovated. beach as well ...	1	great bed rooms, renovated. beach well beati p...
4	stayed for five nights, following a long weeke...	1	stay five nights, follow long weekend break ma...
...	...	...	...
2041	we stayed at this glorified youth hostel this ...	0	stay glorifi youth hostel month supposedli 2 w...
2042	my self and my travelling companion seamus boo...	0	self travel companion seamu book 2 week holida...
2043	the staff was ok till some problems appeared.....	0	staff ok till problem appeared... day came bac...
2044	came back on the 13th july... hotel very expen...	0	came back 13th july... hotel expens £1.50 coke...
2045	agadir is a touristic location, tranquil, safe...	0	agadir tourist location, tranquil, safe pleasa...

3- Models

- Machine Learning Models

Dans la phase initiale, une variété de modèles d'apprentissage automatique traditionnels, y compris Naïve Bayes multinomial, Régression logistique, XGBoost et Machine à vecteurs de support (SVM), sont explorés pour l'analyse des sentiments. Ces modèles utilisent la technique de vectorisation TF-IDF pour convertir les données textuelles en caractéristiques numériques, améliorant ainsi leur efficacité. L'optimisation des hyperparamètres est effectuée à l'aide de GridSearchCV pour affiner les paramètres des modèles. Naïve Bayes prend en compte la plage n-gram, la fréquence maximale du document et le lissage alpha. La Régression logistique se concentre sur la plage n-gram, la fréquence maximale du document, la force de régularisation (C) et le type de pénalité. XGBoost est optimisé pour la plage n-gram, la fréquence maximale du document, le nombre d'estimateurs, la profondeur maximale et le taux d'apprentissage. Le SVM, configuré avec la vectorisation TF-IDF, subit une optimisation des paramètres, y compris la plage n-gram, la fréquence maximale du document, la régularisation (C) et le choix du noyau (linéaire ou fonction de base radiale - RBF).

- Deep Learning Models

Couvrant des variations dans les dimensions d'intégration, les taux de dropout, les filtres convolutifs (GRN), les unités récurrentes (LSTM), et les choix

Architecturaux. Les modèles de réseaux de neurones sont entraînés sur des séquences tokenisées et complétées des données textuelles. Le processus d'entraînement implique plusieurs époques, et les modèles sont évalués sur l'ensemble de test pour évaluer leur performance finale. Les rapports de classification et les scores de précision fournissent une compréhension complète de la capacité des modèles à généraliser sur des données invisibles. La diversité des modèles, Allant des modèles d'apprentissage automatique traditionnels aux architectures de deep learning, permet une exploration approfondie de l'approche la plus appropriée pour l'analyse des sentiments en Darija. Les sections suivantes présentent les résultats et les analyses comparatives, mettant en lumière les forces et les limites de chaque modèle dans le contexte des objectifs du projet.

- Resultats des experiences

MODEL	BEST HYPERPARAMETERS	VALIDATION ACC	FINAL TEST ACC
MultinomialNB	{'alpha' :0.1, 'fit_prior': False, 'max_df': 0.8, 'ngram_range': (1, 2)}	0.973	**0.757**
LogisticReg	{'C': 1.0, 'penalty': 'l2', 'max_df': 0.8, 'ngram_range': (1, 1)}	0.957	0.954
XGBoost	{'learning_rate': 0.2, 'max_depth': 5, 'n_estimators': 50, 'max_df': 0.8, 'ngram_range': (1, 1)}	0.908	0.931
SVM	{'C': 1, 'model__gamma': 'scale', 'model__kernel': 'linear', 'max_df': 0.8, 'ngram_range': (1, 2)}	0.973	0.960
LSTM	{'embedding_dim': 50, 'dropout_rate': 0.2, 'recurrent_units': 64, 'num_layers': 1,'optimizer': 'adam', 'architecture': LSTM}	0.791	0.791
GRU	{'embedding_dim': 100, 'dropout_rate': 0.3, 'recurrent_units': 128, 'num_layers': 2, 'optimizer': 'adam', 'architecture': 'GRU'}	0.798	0.785

Nous constatons que nos classificateurs ont atteint les niveaux de précision souhaités, ce qui est un résultat satisfaisant compte tenu de la complexité de la tâche. En effet, les commentaires sur les hôtels en anglais dépendent directement de l'expérience des utilisateurs, contrairement à d'autres types de données textuelles qui peuvent être relativement indépendantes et contenir suffisamment d'informations pour être classées correctement.

Le tableau 2 présente les résultats du test du classificateur entraîné, qui est le Naïve Bayes multinomial, un algorithme bien connu pour la classification de texte. Les valeurs de précision et de rappel pour chaque classe ont été calculées en créant la matrice de confusion.

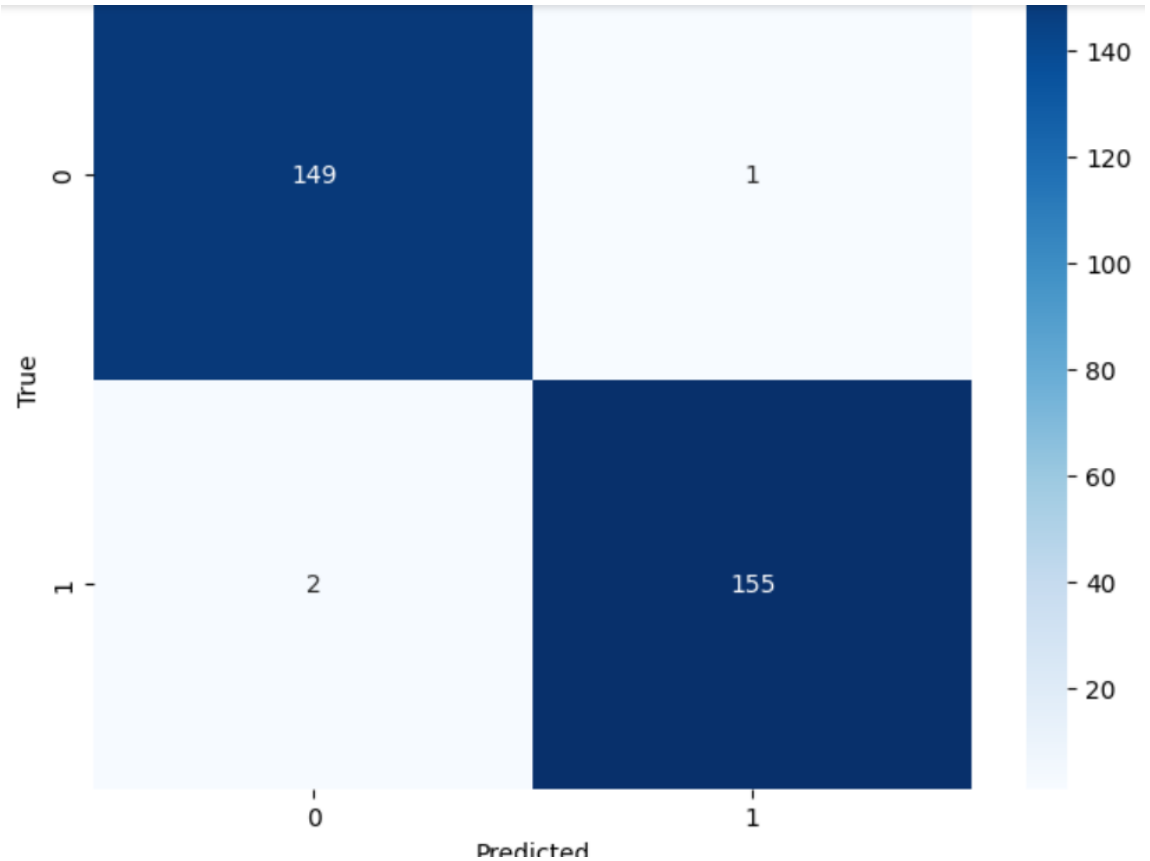


Tableau : Matrice de confusion

5. CONCLUSION:

Dans cette étude, nous avons entrepris une exploration complète de l'analyse des sentiments dans les commentaires sur les hôtels en anglais. La recherche visait à relever les défis linguistiques uniques posés par les commentaires des utilisateurs et à contribuer au domaine plus large du traitement automatique du langage naturel dans des contextes linguistiques informels et spécifiques au secteur hôtelier.

Nos expériences d'analyse des sentiments ont révélé l'efficacité des modèles d'apprentissage automatique traditionnels, tels que le Naïve Bayes multinomial, la Régression logistique, XGBoost et la Machine à vecteurs de support, pour capturer les nuances des sentiments dans les commentaires sur les hôtels. Les modèles, optimisés à l'aide d'une optimisation des hyperparamètres, ont démontré des niveaux de précision prometteurs à la fois sur les ensembles de validation et de test.

- **References**

https://github.com/hamzaae/DCSA/
https://www.tripadvisor.com/Hotels-g293730-Morocco-Hotels.html
https://console.apify.com/actors/Hvp4YfFGyLM635Q2F/input?addFromActorId=Hvp4YfFGyLM635Q2F
https://github.com/abdelghafor-gh/Political-Opinion-Analysis-on-Palestinian-Israeli-Conflict