



Project: Data Exploration and NLP Modeling

Objectif et contexte

Nous avons récolté des avis sur plusieurs boutiques en ligne aux thématiques diverses à partir du site trustpilot.

L'objectif est d'identifier les caractéristiques de chaque catégorie de boutique à partir de leurs avis et de comprendre le niveau de satisfaction client pour chacune d'entre elles.

Pour cela nous avons réalisé une analyse exploratoire de nos données, identifié les thématiques, observé le niveau de satisfaction globale à partir des avis client et créé des modèles qui permettent de déduire d'un avis son niveau de satisfaction.

Présentation des données du projet

Pour notre étude nous avons utilisé une base de données comportant 38308 avis et 7 colonnes.

Ce dataset se compose d'une ligne par avis scrappé. On y retrouve une colonne "name" qui contient le pseudo de l'auteur de l'avis, une colonne shop qui affiche le nom

de la boutique en ligne notée, une colonne "score" qui donne la note décernée par l'utilisateur, une colonne "review" qui contient l'avis, une colonne "review_date" avec la date de publication de l'avis, et enfin une colonne "experience_date" avec la date à laquelle l'utilisateur a profité des services de la boutique en ligne.

3825 avis sont supprimés de la base de données car ne comportant aucun texte dans la colonne review.

Nous allons ajouter une colonne genre qui déduit le genre de l'auteur de l'avis à partir de son pseudo.

Analyse des n-grams dans les avis

Nous avons ensuite procédé à une analyse des n-grams dans les avis pour déterminer les ensembles de mots les plus couramment utilisées. Cela nous a permis de déceler des tendances dans les commentaires et de mieux comprendre les sentiments des clients. Les avis étant en très grande majorité positifs (5 étoiles) les mots qui ressortent le plus souvent sont positifs.

Création d'un dataset échantillon

Pour faciliter les tâches de résumé et de modélisation qui arrivent ensuite, il a fallu réduire la taille du jeu de donnée. Nous avons choisi de conserver 1600 lignes, pour accélérer la génération de résumé. Nous avons aussi rééquilibré le jeu de donnée pour voir autant d'avis positifs que négatifs dans le dataset, afin d'améliorer les résultats des modèles.

Correction orthographique des avis

Nous avons effectué une correction orthographique des avis grâce à Spell Checker afin d'améliorer la qualité des données. Les résultats n'ont pas l'air si fiable.

Génération de résumé

Pour créer un résumé de chacun de nos avis, nous utiliserons BART, et plus spécifiquement BARTnez qui a l'avantage de mieux traiter le langage français, un transformer régulièrement utilisé pour la génération de résumés. Nos avis étant particulièrement courts, les résumer ne servirait pas à grand chose. Nous allons donc créer un résumé à partir des avis de chaque boutique à partir de notre dataset

échantillon. Pour cela nous allons faire une agrégation sur notre dataset pour obtenir une ligne par shop, concaténer tous les avis de ce shop dans une nouvelle colonne total_review, puis résumer cette longue chaîne de caractères. Nous avons comparé les résumés de BART à ceux de BARTez, les seconds semblent plus précis et fiables.

Traduction

Nous avons traduit nos avis grâce à google traduction et les résultats sont très satisfaisants.

Topic modeling

Pour l'analyse de sujet, nous avons utilisé le modèle LDA (Latent Dirichlet Allocation) pour identifier les principaux thèmes dans notre ensemble de données. Cela nous a permis de regrouper les avis en fonction de leurs thèmes et de mieux comprendre les préoccupations principales des clients. Nous avons créé une fonction qui génère les mots clés de nos avis.

Embedding to identify similar words

Pour identifier les mots similaires, nous avons utilisé des techniques d'incorporation de mots comme Word2Vec. Ces outils nous ont permis de créer des vecteurs de mots, qui capturent le contexte sémantique des mots dans le corpus des avis. Cela nous permet de visualiser la proximité et le degré de similarité de nos mots.

Modèles de classification supervisé

Ici nous allons essayer de prédire à partir des avis si l'expérience a été bonne (3,4 ou 5 étoiles) ou si elle a été mauvaise (1 ou 2 étoiles).

Pour cela, nous avons développé plusieurs modèles de classification supervisée, y compris random forest, des réseaux de neurones avec et sans embedding layer pré-entraînée, BERT et GPT. Ces modèles ont été formés sur des caractéristiques extraites des avis, y compris le TF-IDF des mots et les caractéristiques basées sur des sentiments.

Le modèle qui semble le mieux performer est le réseau de neurones avec embedding layer pré-entraînée. C'est donc ce modèle que nous évaluons par la suite.

Interprétation des résultats

Nous avons visualisé les lignes pour lesquelles le modèle se trompe mais difficile de comprendre pourquoi. Nous avons aussi utilisé un modèle de détection de sentiment sur les avis pour comparer avec le nôtre.

Streamlit

Nous avons créé une application Streamlit pour démontrer notre projet. Cela inclut des visualisations interactives, des exemples de prédictions et une interface utilisateur pour tester notre modèle avec de nouveaux avis. Aussi une interface qui génère des résumés à partir d'un texte en entrée.