Analyse de la réputation d'une franchise de Tim Hortons

Tommy Pou

Juin 2024

1 Introduction

La chaîne de restauration rapide *Tim Hortons* figure parmi les entreprises les plus populaires au Canada. Malgré sa popularité, cette chaîne suscite des avis assez partagés chez les clients. Dans ce papier, j'entamerai une analyse de réputation d'un restaurant *Tim Hortons* en particulier.

2 Méthodologie

Le restaurant dont il est question se trouve sur 183g Boul Hymus. Pour réaliser cette analyse, j'effectuerai d'abord une fouille de données (*data mining*) pour collecter les données des revues soumises sur Google Maps. Par la suite, je montrerai des observations statistiques à l'aide de visualisations et je ferai une analyse de sentiment pour mieux comprendre le contenu de ces revues.

2.1 Fouille de données

Pour obtenir automatiquement les données des revues sur Google Maps, j'ai rédigé un script (*scrape.py*) en Python pour scraper la page Web. Je me suis servi du module **Selenium** qui, grâce au WebDriver, permet d'automatiser des traitements du fureteur. Par exemple, le programme glisse la fenêtre de revues en continu afin de charger les revues antérieures.

Les résultats ont été enregistrés dans un fichier *.csv* intermédiaire. Parmi les données extraites, il y a le texte associé à la revue et le nombre d'étoiles. La date n'a pas été enregistrée, mais les revues sont classées en ordre de publication.

2.2 Visualisation de données

L'analyse des données est contenue dans le fichier *analysis.ipynb*. Les librairies **MatPlotLib** et **seaborn** sont utiles pour générer des graphiques. Je me

suis servi de ces deux librairies pour produire des histogrammes.

Une autre librairie que j'ai trouvé pertinente est **WordCloud**. Elle permet de visualiser les termes les plus fréquents pour un ensemble de textes en faisant une mise à l'échelle.

2.3 Traitement de langage naturel (NLP)

La librairie **NLTK** comprend plusieurs fonctions utiles en traitement de langage. Par exemple, en pré-traitement, la librairie permet de filtrer les *stopwords* tels que les déterminants *the* ou *a*. La librairie **TextBlob** comprend également d'autres fonctions utiles en NLP telles que la fonction *correct()* qui corrige les mots mal orthographiés.

Enfin, pour réaliser l'analyse de sentiment, je me suis servi du modèle VADER, qui est implémenté dans la librairie **NLTK**.

3 Résultats et analyse

3.1 Statistiques

Sur Google Maps, l'établissement est évalué à 3.6 étoiles avec 707 avis. Parmi ces avis, 322 utilisateurs ont rédigé une revue. L'analyse portera sur ces 322 revues.

Les revues sont assez polarisées. Les avis à 1 étoile sont les plus fréquents, suivis des avis à 5 étoiles.

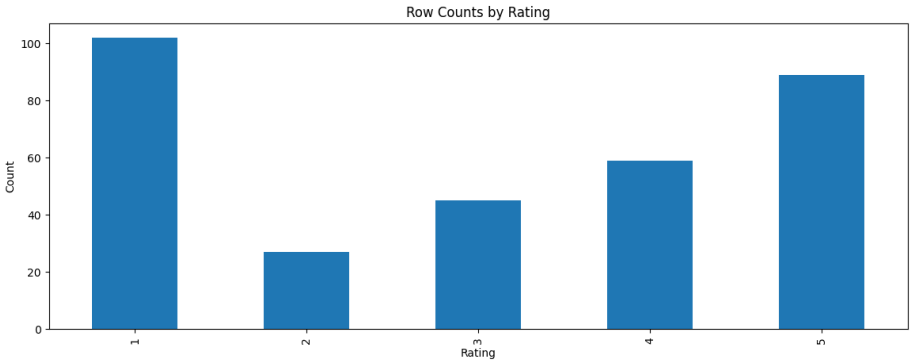

FIGURE 1 – Distribution des revues sur Google Maps.

Parmi les revues négatives (une ou deux étoiles), les termes les plus récurrents sont : *order, coffee, service, time* et *slow*. Parmi les revues positives (quatre ou

cinq étoiles), les termes les plus fréquents sont : *service, good, coffee, staff* et *great*.

FIGURE 2 – Mot-nuages des revues négatives (à gauche) et positives (à droite).

Le manque de dates limite l'analyse des données. Cependant, sachant que les revues sont triées en ordre de publication, nous pouvons étudier l'évolution de la moyenne des revues. Le graphe ci-dessous suggère que les évaluations semblent tendre vers le bas avec les revues les plus récentes.

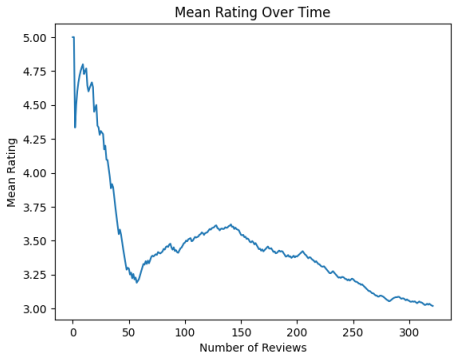

FIGURE 3 – Évolution de la moyenne avec le nombre de revues.

3.2 Analyse de sentiment

Le modèle VADER est un modèle pré-entraîné qui attribue un score au sentiment d'un texte. Ce modèle est basé sur des règles et a été entraîné sur des messages de réseaux sociaux. Le score retourné varie de -1 à 1 selon le sentiment du texte.

Si on applique ce modèle à notre jeu de données, nous devrions observer un score qui accroît en fonction du nombre d'étoiles. Effectivement, les revues avec plus d'étoiles devraient évoquer un sentiment positif et celles avec moins d'étoiles devraient suggérer le sentiment inverse.

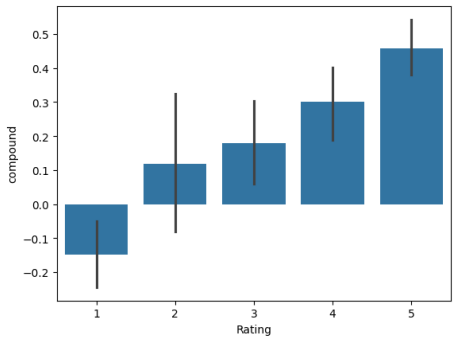

FIGURE 4 – Score VADER moyen selon l'évaluation de la revue.

L'allure du graphique ci-dessus valide le comportement attendu. Le score médiane est de **0.17** et peut être utilisé comme un seuil pour classifier les revues.

Ainsi, nous pouvons procéder à une analyse spécifique selon quatre catégories : la nourriture, le service à la clientèle, l'entretien de l'établissement et le prix. Pour catégoriser les revues, j'ai filtré les textes selon la présence de mot-clés liés à la catégorie. Les résultats sont les suivants.

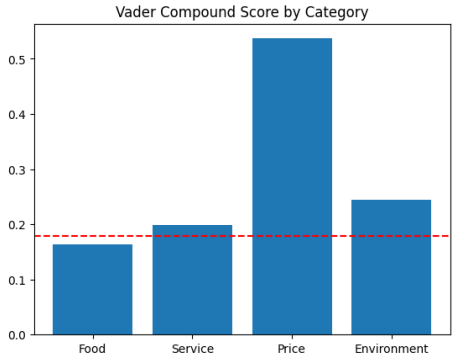

FIGURE 5 – Score VADER moyen par catégorie.

Tel que fait avec les évaluations des revues, nous pouvons examiner la tendance des scores VADER en fonction de l'ancienneté de la revue.

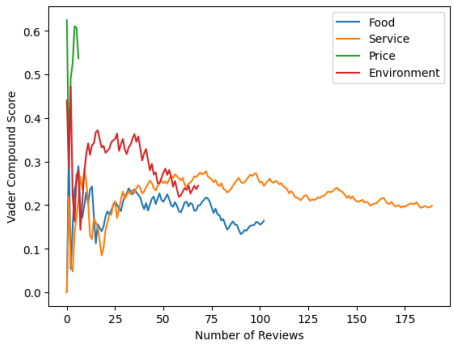

FIGURE 6 – Évolution du score VADER moyen par catégorie.

Les scores, tout comme les évaluations, semblent tendre vers le bas avec les revues les plus récentes.

4 Conclusion

En résumé, voici les points-clés de mon analyse :

1. La distribution des avis est plutôt éparse.
2. Les revues positives et négatives sont majoritairement liées à la nourriture et au service à la clientèle.
3. Les évaluations tendent à être davantage négatives récemment.
4. Soient les quatre catégories suivantes : nourriture, service, environnement et prix. La nourriture évoque le plus de sentiments négatifs et le prix évoque le plus de sentiments positifs. Le service est la catégorie contenant le plus de revues.

Plusieurs améliorations peuvent être apportées à cette analyse. Par exemple, certaines revues ont laissé en pièce jointe des images de l'établissement. À l'aide de méthodes en vision par ordinateur, nous pourrions exploiter ces données supplémentaires pour obtenir des informations supplémentaires sur la qualité du lieu.

Je tiens à rappeler que les revues analysées ne sont qu'une fraction des avis sur Google Maps. Plus de la moitié des avis ne contiennent aucune revue, donc l'échantillon analysée ne représente pas nécessairement l'opinion majoritaire.