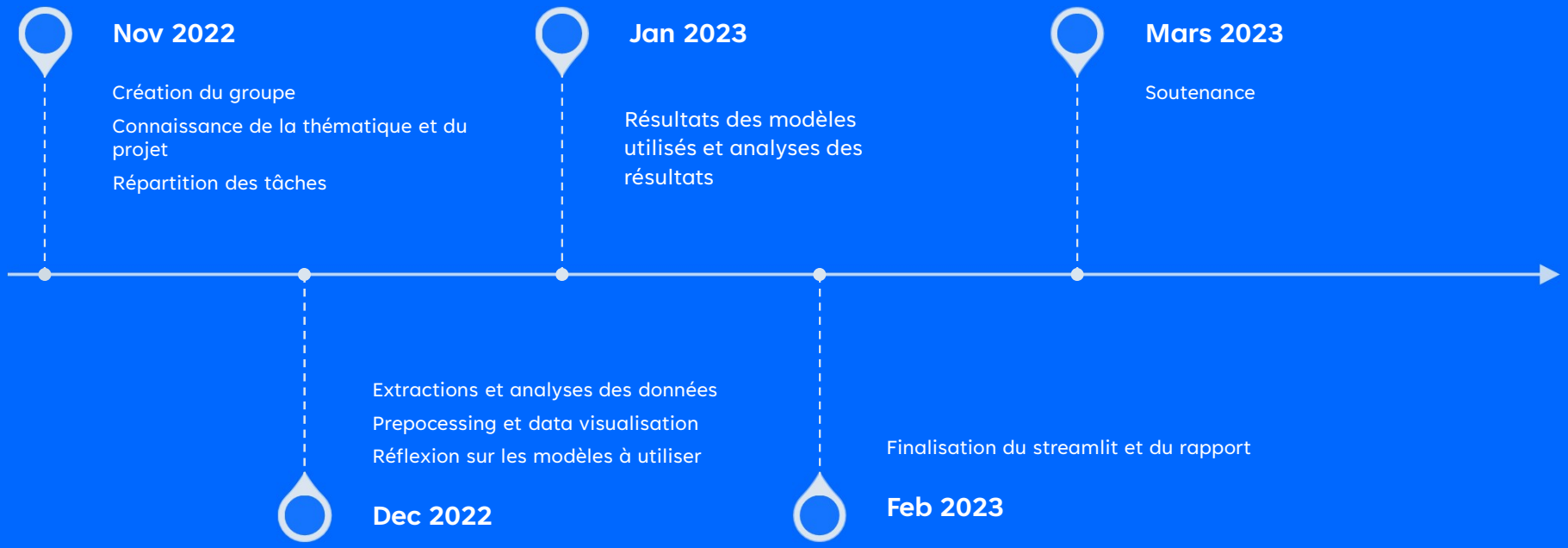
Satisfaction client

Thomas Fourtouill et Yvon-Arnaud Gbe

Chronologie du projet



Organisation du projet Data Science

1

Collecte

Récupération des données par WebScraping

2

Preprocessing

Tokenization et nettoyage des données

Wordcloud

3

Modélisation

Les modèles de machine learning

Les modèles de deep learning

4

Stratégie

Choix des modèles les plus performants

5

Lancement

Déployer le modèle de prédiction des notes sur Streamlit

Ordre du jour

Introduction et contexte

Objectifs

Analyse des données

Métriques et validation du modèle

Conception du Streamlit

Conclusion

Introduction et Contexte

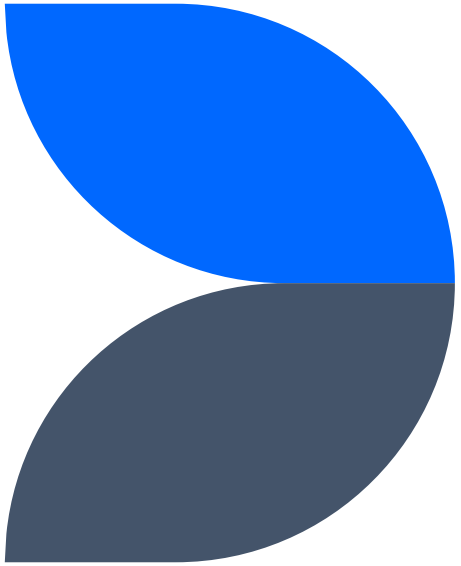
Pour déterminer si elle prend les bonnes décisions et qu'elle se concentre sur la bonne stratégie, une entreprise doit savoir si elle satisfait pleinement ses clients à travers une expérience adaptée à leurs envies. C'est pour cette raison précise que la mesure de la satisfaction client est incontournable pour toute entreprise, et que l'analyse des commentaires clients devient primordiale.

Introduction et Contexte

Dans une société de plus en plus impactée par les avis en ligne et les réseaux sociaux, une entreprise accorde davantage d'importance à leur e-réputation. Les commentaires apportent une certaine notoriété à la marque et permet de booster les ventes.

Objectif principal

Prédire les notes de satisfactions clients des produits à travers l’analyse des commentaires



Prérequis

Création d'un environnement virtuel et installation de toutes les bibliothèques nécessaires

Librairies utilisés :

- Pandas
- Numpy
- Tensorflow version 1.8.1
- Scikit-learn version 1.11.1
- Wordcloud et PIL
- Streamlit version 1.6.0

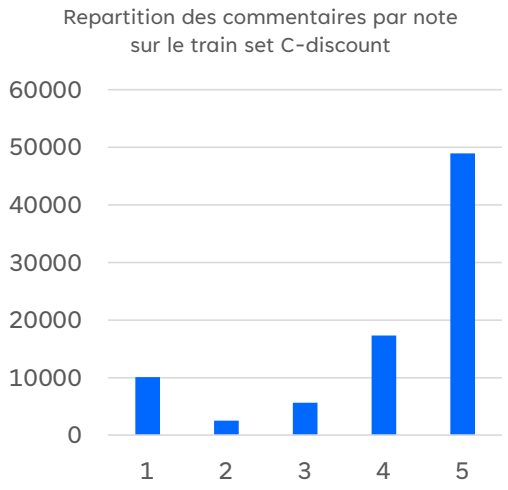
WEB SCRAPING

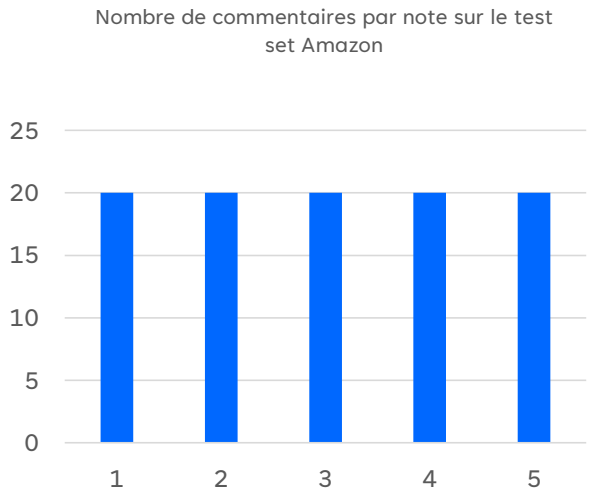
Commentaires
et avis web

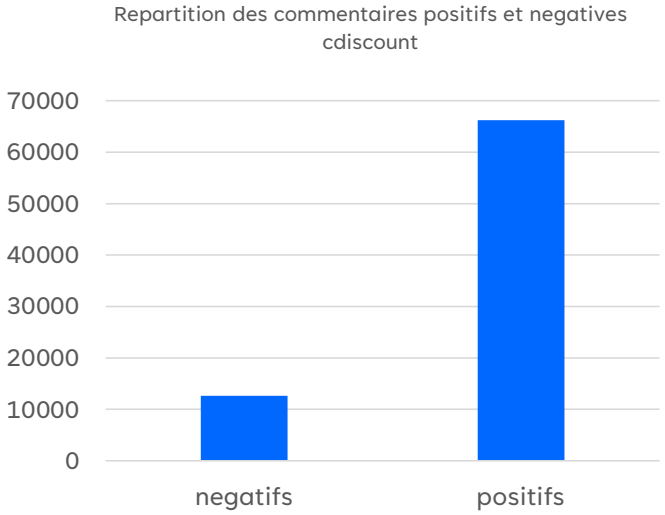
Web Scraping:
Script et
programme

Stockage et
base de
données

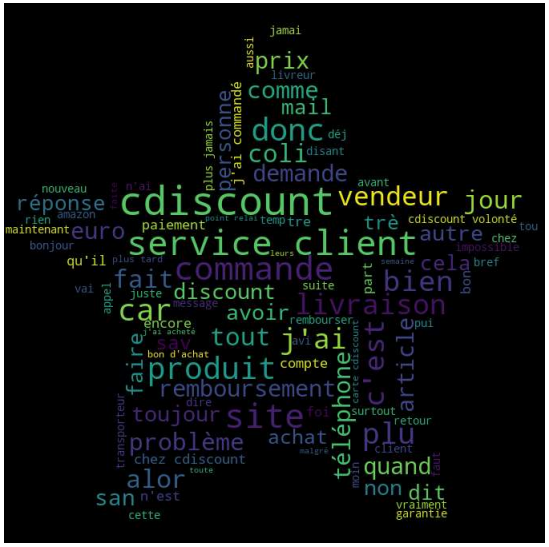
Analyse et répartitions des données



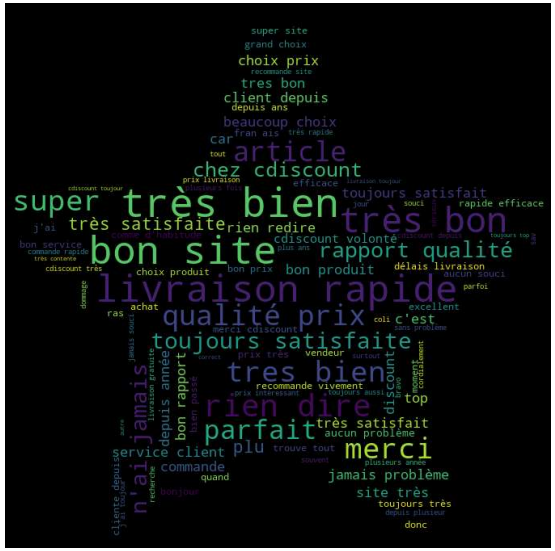




Les mots les plus représentés: Wordcloud



Les 100 mots les plus représentés dans les commentaires négatifs



Les 100 mots les plus représentés dans les commentaires positifs

Métriques et validation des modèles

Métriques des modèles de machine learning avec des groupes de mots

Type	Nom du modèle	Paramètres	Score modèle sur 5 notes	Score négatifs / positifs
Machine Learning	RandomForestClassifier ngram[1, 2]	n_estimators=50	0.68	0.94
Machine Learning	RandomForestClassifier ngram[2, 2]	n_estimators=50	0.64	0.91
Machine Learning	RandomForestClassifier ngram[3, 2]	n_estimators=50	0.63	0.92

Métriques des modèles de Machine Learning

Nom du modèle	Paramètres	Score modèle sur 5 notes	Score négatifs / positifs
DecisionTreeClassifier	max_depth=10	0.69	0.90
GradientBoostingClassifier	n_estimators=25	0.64	0.90
TF_IDF RamdomForestClassifier	n_estimators=50	0.69	0.94

Métriques des modèles de Deep Learning

Type	Nom du modèle	Paramètres	Score 2 du modèle sur 5 notes	Score pour sentiments négatif/positif
Deep Learning ANN	Embedding1	Embedding, globalAveragePooling, Dense(64)	0.71	0.95
Deep Learning ANN	Embedding2	Embedding, globalAveragePooling, Dense(32), Dense(32)	0.70	0.95
Deep Learning ANN	Embedding3	Embedding, globalAveragePooling, Dense(256), Dense(128), Dense(64)	0.71	0.96
Deep Learning ANN	Embedding5	Embedding, globalAveragePooling, Dense(256), Dense(128), Dense(64), Dense(32)	0.71	0.95
Deep Learning ANN	Embedding6	Embedding, globalAveragePooling, Dense(256)	0.70	0.96
Deep Learning ANN	Embedding7	Embedding, globalAveragePooling, Dense(1024)	0.71	0.96
Deep Learning RNN	Embedding4	Embedding, LSTM(200)	0.70	0.95
Deep Learning RNN	Embedding8	Embedding, RNN(GRUCell(128), globalAveragePooling, Dense(256)	0.71	0.96

Résultat des tests sur pour la classification des notes allant de 1 à 5

Résultat des prédictions sur le X_test C-discount

	precision	recall	f1-score	support
1	0.73	0.86	0.79	2012
2	0.00	0.00	0.00	497
3	0.35	0.42	0.38	1121
4	0.50	0.18	0.26	3495
5	0.77	0.94	0.85	9763
accuracy			0.71	16888
macro avg	0.47	0.48	0.46	16888
weighted avg	0.66	0.71	0.66	16888

Classification report sur les prédictions issues du X_test

predictions données réelles	1	3	4	5
1	1730	176	18	88
2	244	169	36	48
3	244	473	164	240
4	100	410	627	2358
5	66	135	397	9165

Confusion matrix sur les prédictions issues du X_test

Résultat des prédictions sur le jeux de test Amazon

	precision	recall	f1-score	support
1	0.33	0.85	0.48	20
2	0.00	0.00	0.00	20
3	0.33	0.30	0.32	20
4	0.43	0.15	0.22	20
5	0.62	0.75	0.68	20
accuracy			0.41	100
macro avg	0.34	0.41	0.34	100
weighted avg	0.34	0.41	0.34	100

Classification report sur les prédictions issues du test_amazon

predictions données réelles	1	3	4	5
1	17	3	0	0
2	17	3	0	0
3	8	6	4	2
4	5	5	3	7
5	4	1	0	15

Confusion matrix sur les prédictions issues du test_amazon

predictions données réelles	1	3	4	5
1	0.85	0.15	0.00	0.00
2	0.85	0.15	0.00	0.00
3	0.40	0.30	0.20	0.10
4	0.25	0.25	0.15	0.35
5	0.20	0.05	0.00	0.75

Confusion matrix sur les prédictions issues du test_amazon ramener en % en ligne

Résultat des tests pour la classification des sentiments négatifs (0) et positifs (1)

Résultat des prédictions sur le jeux de test Amazon

	precision	recall	f1-score	support
0	0.86	0.80	0.83	40
1	0.81	0.88	0.84	40
accuracy			0.84	80
macro avg	0.84	0.84	0.84	80
weighted avg	0.84	0.84	0.84	80

Classification report sur les prédictions issues du test_amazon

predictions	0	1
données réelles		
0	32	8
1	5	35

Confusion matrix sur les prédictions issues du test_amazon

predictions	0	1
données réelles		
0	0.800	0.200
1	0.125	0.875

Confusion matrix sur les prédictions issues du test_amazon ramener en % en ligne

Commentaire sur les modèles

Les résultats nous montrent que les modèles de Deep Learning semble plus performant que les modèles de Machine learning et beaucoup plus efficace à appliquer.

Les résultats sont meilleurs sur les notes positifs et négatifs

Contraintes et Limite du modèle

Contraintes

- Installation d’un environnement virtuel
- Difficulté à choisir les bonnes versions des librairies
- Les modèles doivent s’appliquer sur les mêmes données

Limites

- Prédiction à la note (1à 5) difficile
- Commentaire pas toujours bien rédigé
- Avoir beaucoup plus de données pour améliorer la performance des modèles

Conception interface web : Streamlit

1. Présentation des données
2. Saisir d’un commentaire et calcul des prédictions
3. Documentation du projet

Conclusion

Le projet est largement perfectible notamment avec :

1. L'augmentation des données traitées
2. Les modèles de Deep learning seq2seq avec encodeur et décodeur
3. Demander le retour de l'utilisateur sur la qualité des prédictions pour agréger une base de donnée qui servira à l'amélioration du modèle

Merci à l’équipe Datascientest