

Analyse des Avis Amazon en anglais

I. Analyse du corpus

i. Description du corpus

Le corpus que nous avons choisi, intitulé “[SetFit/amazon reviews multi en](#)” et disponible sur Hugging Face, est un ensemble de données contenant des avis textuels en anglais publiés par des clients sur Amazon à propos de nombreux articles. Ces avis sont étiquetés selon leur polarité, ce qui permet de réaliser des analyses de sentiments sur des opinions de consommateurs.

La répartition est d'environ 200 000 avis pour l'ensemble de train, 5 000 pour la validation et 5 000 pour le test. Étant donné que notre objectif est d'effectuer une évaluation sur le modèle, nous avons choisi de travailler uniquement avec l'ensemble de données de test. De plus, l'ensemble de train est très volumineux, ce qui a fait planter notre programme.

Le corpus est en format JSONL, ce qui est une variante du format JSON dans laquelle chaque ligne du fichier représente un objet JSON distinct, ce qui facilite le stockage et la lecture par avis.

Il y a 4 colonnes différentes dans le corpus, ce qui représentent respectivement: ID de l'avis, texte de l'avis, label(en format int), label(en format string). Les labels dans le corpus sont représentés par des entiers de 0 à 4, où chaque valeur indique le niveau de satisfaction du client exprimé dans l'avis. En lisant les avis et selon les convention de l'utilisation d'Amazon, nous supposons que les labels de polarité dans le corpus sont notés de la manière suivante:

- 0** : Avis extrêmement négatif
- 1** : Avis négatif
- 2** : Avis neutre ou légèrement négatif
- 3** : Avis positif
- 4** : Avis extrêmement positif

En ce qui concerne le prétraitement, les avis des clients dans ce dataset semblent avoir été conservés dans leur forme originale, incluant toutes les ponctuations. Il n'y a pas de signes indiquant un nettoyage, une lemmatisation ou de suppression des stopwords qui auraient été appliquées.

ii. Description statistique

91,5 % des avis ont une longueur entre 15 et 405(par caractère), 7.2% des avis entre 405-795(par caractère). La longueur moyenne des avis est à 33.9262 mots. Vu que les données n'ont pas été prétraitées, les mots les plus fréquents sont comme suite avec leur fréquence:

Fréquence des mots les plus communs :	
the	7378
I	5147
and	4434
to	4339
a	3941
it	3431
is	2442
of	2209
for	2191
but	1744

iii. Description qualitative

Voici l'exemple d'un avis tiré de notre corpus:

id string · lengths	text string · lengths	label int64	label_text string · classes
10 100%	15→405 91.5%	0 20%	0 20%
en_0656905	Never worked.... no help in getting it fixed or replaced	0	0

II. Présentation de l'outil

i. Description de l'outil

Nous avons utilisé un modèle BERT-base-multilingual-uncased, affiné pour l'analyse de sentiments sur des avis produits dans six langues : anglais, néerlandais, allemand, français, espagnol et italien. Il prédit le sentiment de l'avis sous forme d'un nombre d'étoiles (entre 1 et 5).

Ce modèle est conçu pour être utilisé directement comme modèle d'analyse de sentiments pour les avis produits dans l'une des six langues mentionnées, ou pour un affinement supplémentaire sur des tâches d'analyse de sentiments similaires.

Ce modèle, basé sur BERT-base-multilingual-uncased, a été entraîné sur 17 280 avis Amazon et testé sur un ensemble de 2 400 avis, permettant une reconnaissance relativement précise des sentiments des avis Amazon, avec un taux de précision atteignant 80 %.

BERT se base sur l'Encoder de Transformer (auto-attention et le réseau neuronal feed-forward) et utilise un apprentissage bidirectionnel pour mieux capturer les informations contextuelles. Son entraînement repose sur deux tâches principales :

Modèle de langage masqué (Masked Language Model, MLM) : Pendant l'entraînement, certains mots dans une phrase sont masqués aléatoirement, et le modèle doit prédire ces mots masqués en utilisant le contexte environnant. Grâce à cet apprentissage bidirectionnel, BERT parvient à comprendre chaque mot dans son contexte d'apparition.

Prédiction de la phrase suivante (Next Sentence Prediction, NSP) : Cette tâche permet au modèle d'apprendre les relations entre phrases. On présente au modèle deux phrases consécutives ou non, et il doit déterminer si elles se suivent dans le texte d'origine. Cela améliore les performances de BERT pour des tâches impliquant la compréhension de relations entre phrases, comme les questions-réponses et le raisonnement naturel.

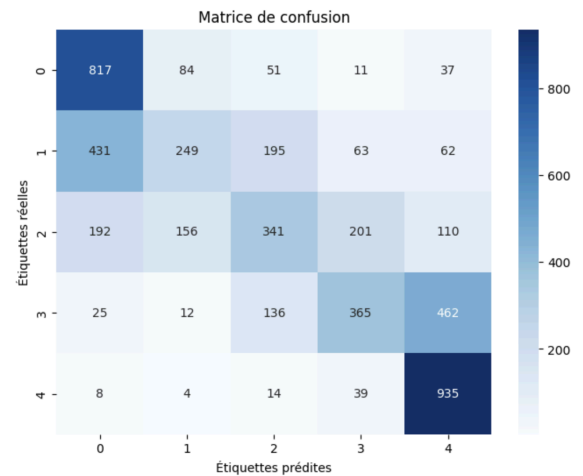
ii. Pourquoi cet outil?

Le modèle BERT est particulièrement bien adapté aux tâches de reconnaissance et de classification des sentiments, car il peut comprendre le sens d'un mot central en fonction du

contexte des mots qui l'entourent, et utiliser l'attention pour analyser l'importance de chaque mot. Dans notre cas, nous avons utilisé une version de BERT enrichie avec des avis Amazon ajoutés après la pré-formation sur le modèle de base. Cela rend le modèle encore plus adapté à la tâche de classification des sentiments des avis Amazon, c'est pourquoi nous avons choisi ce modèle pour notre tâche de classification des sentiments.

III. Mise en œuvre et évaluation

La précision est de 0,52, le rappel de 0,54 et le score F1 de 0,50. Nous avons également visualisé la matrice de confusion avec Seaborn. En ce qui concerne la comparaison avec les modèles disponibles sur Hugging Face, il n'y a qu'un seul modèle pertinent pour notre tâche, nous n'avons donc pas effectué de comparaison.



IV. Discussion et analyse

i. Résultats obtenus

Les résultats obtenus en termes de précision, de rappel et de score F1 ne sont pas satisfaisants. Cependant, d'après la matrice de confusion, les prédictions restent proches des valeurs réelles. Le modèle se montre plus performant pour les classes "0" et "4", ce qui est compréhensible, étant donné que les émotions sont plus marquées dans les commentaires attribuant des notes de "1" et "5" étoiles.

ii. Difficultés rencontrées

Le corpus "train" contient 200 000 commentaires, ce qui rend impossible d'entraîner avec notre machine, donc nous avons décidé d'utiliser "test" pour au moins pouvoir évaluer le modèle.

iii. Pistes d'amélioration

Nous pouvons entraîner un modèle par nous même.