Rapport – Devoir 6
Canelle Wagner
20232321

Section 1 - Point 3

Nombre de points de données (?)

Il y a 50000 points de données.

Combien de valeurs uniques notre colonne cible contient-elle ?

La colonne « sentiment » qui est la colonne cible, contient 2 valeurs uniques : « positive » et « negative ».

Contient-il des valeurs NaN ?

Il n'y a pas de valeurs NaN.

Y a-t-il des critiques en double ? (Si vous trouvez des lignes en double, combien de doublons avez-vous trouvés ?

Oui il y a 418 critiques en double.

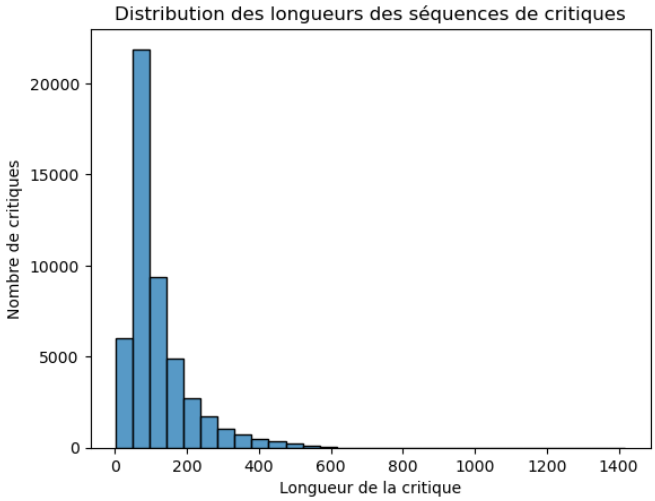
Section 3 - Point 2

Comment les valeurs cibles sont-elles distribuées ? Avons-nous un ensemble de données presque équilibré ?


```
sentiment
positive    24884
negative    24698
Name: count, dtype: int64
```

Les valeurs cibles, qui sont les sentiments, montrent une répartition presque équilibrée avec environ 24 884 critiques positives et 24 698 critiques négatives.

Toutes les critiques ont-elles la même longueur ?



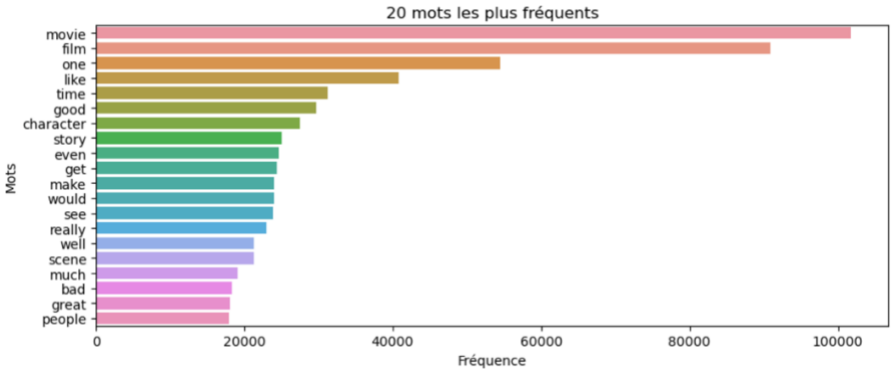
Non, les critiques varient en longueur, comme on peut le voir dans l'histogramme qui montre une large distribution des longueurs des séquences.

Quelle est la longueur moyenne des séquences ?

La longueur moyenne des séquences est d'environ 118 mots.

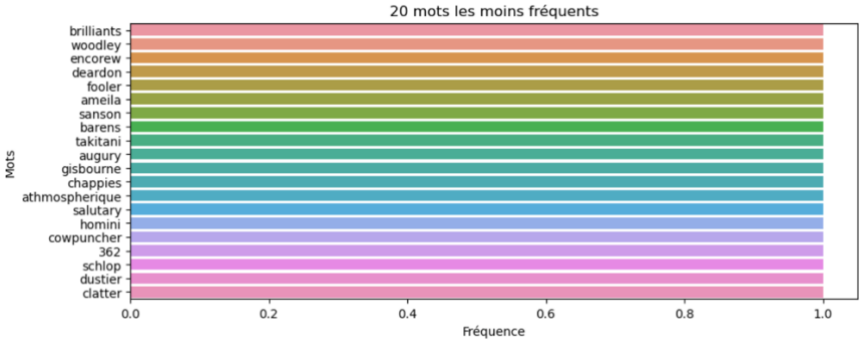
Quels sont les 20 mots les plus fréquents ?

Mots	Fréquence
movie	101691
film	90869
one	54526
like	40787
time	31188
good	29654
character	27544
story	25004
even	24669
get	24412
make	24038
would	24036
see	23841
really	22922
well	21276
scene	21251
much	19151
bad	18293
great	18098
people	17882



Quels sont les 20 mots les moins fréquents ?

Mots	Fréquence
brilliants	1
woodley	1
encorew	1
deardon	1
fooler	1
ameila	1
sanson	1
barens	1
takitani	1
augury	1
gisbourne	1
chappies	1
athmospherique	1
salutary	1
homini	1
cowpuncher	1
362	1
schlop	1
dustier	1
clatter	1



Après avoir effectué une AED, pensez-vous qu'il sera facile de classifier ces critiques ? Pourquoi oui ? Pourquoi non ?

Je dirais que cela pourrait être modérément difficile de classifier ces critiques.

Pourquoi oui :

- Distribution équilibrée : Les étiquettes cibles sont presque équilibrées entre les critiques positives et négatives, ce qui est bon pour un modèle de classification car il n'apprendra pas à favoriser une classe par rapport à l'autre.
- Fréquence des mots : Les mots les plus fréquents comprennent des termes qui semblent pertinents pour les critiques de films, comme 'movie', 'film', 'bad', 'good' ce qui peut aider à distinguer les avis positifs des avis négatifs.

Pourquoi non :

- Longueur variable des critiques : Il y a une grande variation dans la longueur des critiques, ce qui pourrait signifier que certaines critiques contiennent plus de contexte ou d'opinions que d'autres, rendant la classification plus complexe.
- Présence de mots neutres : Beaucoup de mots fréquents sont neutres et pourraient ne pas être très utiles pour la classification, tels que 'one', ‘people’, 'time'. Ils pourraient nécessiter un traitement supplémentaire pour augmenter la performance du modèle.
- Mots peu fréquents : Les mots les moins fréquents semblent être des erreurs de frappe ' sanson ' ou d’orthographe comme ‘athmospherique’ ou encore des termes très spécifiques comme ‘cowpuncher’ qui peuvent ajouter du bruit aux données et ne pas contribuer significativement à la classification.

Section 4 - Point 1

Expliquez les désavantages d'utiliser cette méthode.

L'utilisation de TF-IDF, présente plusieurs inconvénients. Premièrement, TF-IDF crée une matrice de caractéristiques où chaque terme est indépendant des autres. Cela signifie que la méthode ne tient pas compte du contexte ni de l'ordre des mots dans les phrases. Par exemple, les phrases "Le chat mange la souris" et "La souris mange le chat" auraient la même représentation TF-IDF malgré leurs significations très différentes.

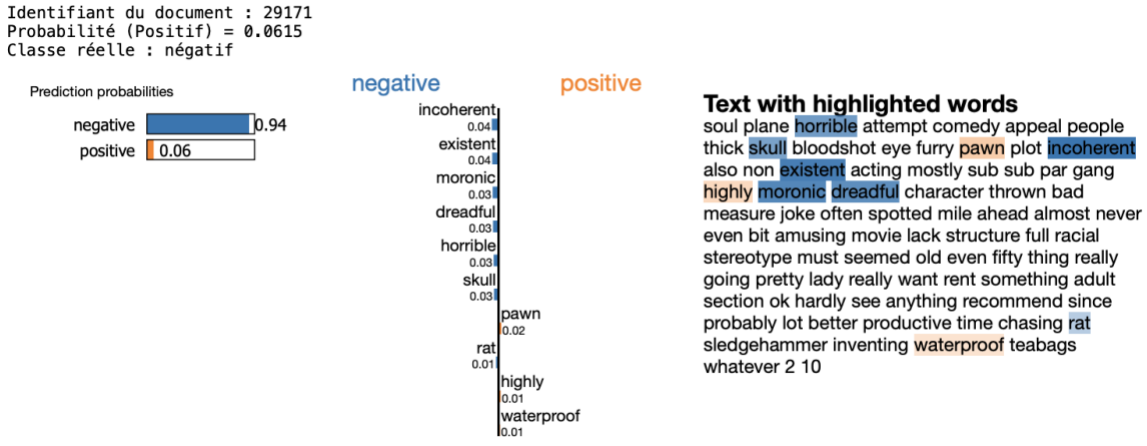
Deuxièmement, TF-IDF ne reconnaît pas les nuances sémantiques des mots. Il traite chaque occurrence d'un mot de manière identique, ignorant les différentes significations que le mot peut avoir dans différents contextes. En outre, les mots rares ou les mots-clés peuvent être surévalués, tandis que les mots courants mais importants pour le sens général d'une phrase peuvent être sous-évalués.

Fournissez une méthode alternative que nous aurions pu utiliser pour trouver une meilleure représentation numérique des mots présents dans notre corpus. (Pourquoi pensez-vous que cela pourrait fonctionner mieux ?)

Une méthode alternative pour représenter numériquement les mots est l'utilisation de Word2Vec. Word2Vec est un modèle de plongement de mots qui représente les mots dans un espace vectoriel de dimensions réduites. Cette méthode capture la sémantique des mots en prenant en compte leur contexte. Par exemple, les mots qui apparaissent dans des contextes similaires auront des vecteurs de plongement similaires, ce qui permet de capturer leur sens sémantique et leur relation avec d'autres mots.

Contrairement à TF-IDF, Word2Vec reconnaît que les mots peuvent avoir des significations différentes en fonction de leur contexte. Il est donc plus efficace pour capturer la sémantique des mots dans de grandes collections de texte. De plus, Word2Vec aide à réduire la dimensionnalité du problème, car il représente les mots dans un espace vectoriel continu et de taille fixe, contrairement à la matrice de caractéristiques de haute dimension générée par TF-IDF.

Utilisation de LIME : Interprétation



Interprétation :

Le modèle a classé la critique comme négative avec une probabilité de 94% (positive à 6%).

Les mots clés contribuant le plus à cette classification négative sont mis en évidence dans le texte. Des termes tels que "incoherent", "moronic" et "horrible", ont un poids significatif dans la décision du modèle, comme le montrent les barres de couleur bleue avec les valeurs associées à chaque terme. La présence de ces termes négatifs suggère que la critique exprime un sentiment défavorable, ce qui est en accord avec la classe prédite par le modèle.

À l'inverse, des termes positifs ou neutres tels que "pawn", "highly" et "waterproof" apparaissent également, mais avec des poids moindres, indiqués par des barres de couleur orange. Il est intéressant de noter que "waterproof" est identifié comme un terme positif, ce qui peut ne pas être pertinent dans le contexte d'une critique de film. Cela indique l'une des limites du modèle en ce qui concerne la prise en compte du contexte spécifique des mots.

En somme, LIME révèle comment certains mots influencent la classification (et avec quel degrés) des critiques par notre modèle, tout en mettant en évidence la nécessité de contextualiser certains termes pour éviter des interprétations erronées.

Section 5

Ma première série d'entraînements a été réalisée avec les hyperparamètres suivants : un taux d'apprentissage de 2e-5, un weight decay de 0.01, batch de 16 et un nombre d'époques fixé à 3. Cela a abouti à des scores de précision approchant 0.91 lors de la deuxième époque, mais une augmentation de la perte de validation entre la deuxième et la troisième époque, tandis que la précision, le score F1 et le rappel sont relativement stables. Cela peut indiquer que le modèle commence à surajuster les données d'entraînement.

J'ai décidé de diminuer le taux d'apprentissage à 1e-5, mais cette modification a entraîné une diminution de la précision. J'ai donc décidé de revenir au taux d'apprentissage original de 2e-5, mais de réduire le nombre d'époques à 2 pour atténuer le risque de surajustement. En raison des restrictions de nos ressources GPU, j'ai dû limiter mon exploration d'ajustements d'hyperparamètres.

La dernière configuration a donné des résultats légèrement inférieurs, avec une précision de 0.9077. Il est intéressant de noter que, malgré un taux d'apprentissage constant, la variation du nombre d'époques a produit des résultats différents.

Learning rate: 2e-5, weight decay: 0.01, batch: 16, époques: 3

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.257100	0.246626	0.903701	0.906290	0.885862	0.927682
2	0.178200	0.267170	0.912474	0.911699	0.923537	0.900161
3	0.113600	0.338157	0.912373	0.913455	0.905787	0.921254

Learning rate: **1e-5**, weight decay: 0.01, batch: 16, époques: **2**

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.071000	0.444207	0.908037	0.909344	0.900039	0.918843
2	0.048400	0.481473	0.909549	0.910879	0.901121	0.920852

Learning rate: **2e-5**, weight decay: 0.01, batch: 16, époques: 2

Epoch	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
1	0.034900	0.550641	0.907028	0.908149	0.900791	0.915629
2	0.030000	0.601994	0.907734	0.909019	0.899980	0.918240

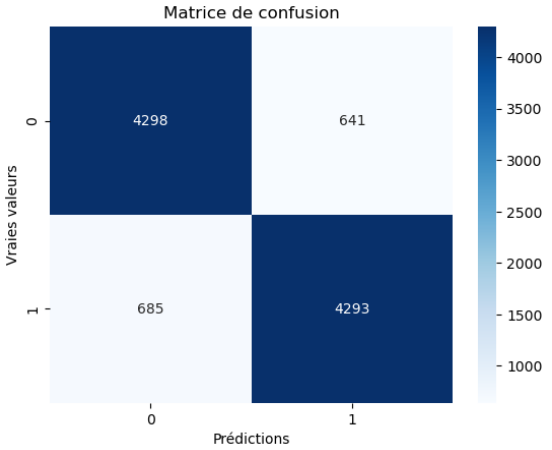
Lors de mon premier entraînement avec un learning rate de 2e-5, un weight decay de 0.01 sur trois époques, j'ai constaté une précision et un score F1 optimaux à la deuxième époque, avec respectivement 0.912474 et 0.911699. Cependant, la troisième époque a montré une légère baisse de ces performances, malgré une diminution de la perte d'entraînement, suggérant un possible surajustement.

Pour contrer ce surajustement, j'ai réduit le learning rate à 1e-5 et limité l'entraînement à deux époques. Cette configuration a légèrement amélioré la précision et le score F1 à la deuxième époque (0.909549 et 0.910879 respectivement), mais à un coût d'une augmentation de la perte de validation.

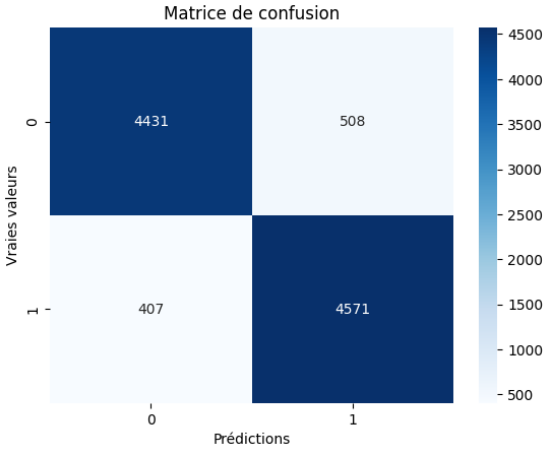
En revenant ensuite à un learning rate de 2e-5 et en réduisant le nombre d'époques à deux, j'ai observé une diminution de la précision et du score F1 lors de la deuxième époque (précision de 0.907734 et F1 de 0.909019). Cette expérience a révélé que, bien que le modèle bénéficie d'un taux d'apprentissage plus élevé pour une convergence rapide, il pourrait nécessiter un peu plus de temps pour affiner ses paramètres et obtenir une meilleure généralisation (mais je suis limitée par mes ressources en GPU de Google Colab).

En résumé, ma recherche indique que le meilleur équilibre entre apprentissage et généralisation pour le modèle DistilBERT a été atteint avec un learning rate de 2e-5 sur trois époques, bien que la meilleure perte de validation ait été obtenue avec un learning rate réduit sur deux époques.

Naïve Bayes



DistilBERT



La matrice de confusion du Naïve Bayes, montre que le modèle a correctement prédit 4298 vrais négatifs et 4293 vrais positifs. Cependant, il y a 641 faux positifs et 685 faux négatifs. Cela suggère que le modèle est relativement équilibré en termes de précision et de rappel, mais présente une certaine tendance à mal classer les instances positives comme négatives et vice versa.

La matrice de DistilBERT, indique une amélioration notable avec 4431 vrais négatifs et 4571 vrais positifs, et moins de faux positifs et faux négatifs (respectivement 508 et 407). Cette amélioration dans toutes les catégories suggère non seulement que DistilBERT a une meilleure précision globale, mais aussi qu'il est plus apte à classifier correctement les cas positifs et négatifs, comme le reflète l'augmentation des scores de rappel et de précision.

En conclusion, les résultats obtenus avec DistilBERT sont meilleurs que ceux obtenus avec le classifieur Naïve Bayes.