



---

# **Analyse de Sentiment : Prédire si une critique de film est positive ou négative**

---

27 MARS 2023

Baptiste Gautier, Oussama Laaumari, Anis Louissi, Camil Zahi  
Supervisé par Grégory Futhazar

## *Table des matières*

<i>Table des matières</i> .....	1
Introduction.....	2
Présentation de la problématique .....	2
Présentation des données.....	3
Statistiques des ensembles de données .....	4
Présentation du modèle .....	6
Architecture.....	6
Fine tuning des paramètres.....	7
Résultat du modèle .....	8
Conclusion .....	9
Résultats .....	9
Perspectives.....	9
Contributions.....	10

## Introduction

### Présentation de la problématique

L'**analyse de sentiment** consiste à **déterminer l'opinion ou l'émotion exprimée dans un texte**, généralement en classant ce dernier comme **positif, négatif ou neutre**. Cette tâche est particulièrement pertinente dans de nombreux domaines, tels que le marketing, la gestion de la réputation en ligne et l'analyse des commentaires sur les sociaux. Les entreprises et les organisations peuvent utiliser l'analyse de sentiment pour comprendre les opinions de leurs clients ou utilisateurs et améliorer leurs produits, services ou stratégies de communication en conséquence.

Dans ce contexte, la problématique abordée ici est d'entraîner un modèle d'analyse de sentiment capable de **classer les critiques de films IMDB en termes de sentiment positif ou négatif**. Les critiques de films représentent un ensemble de données riche pour cette tâche, car elles contiennent des opinions variées et souvent nuancées sur une grande diversité de films. Entraîner un modèle capable de comprendre et de classer correctement ces opinions peut être un défi en raison des subtilités du langage et des différentes manières dont les sentiments peuvent être exprimés.

Pour aborder cette problématique, le code utilise le **modèle DistilBert**, une version plus légère et plus rapide du modèle de traitement du langage naturel BERT, développée par Hugging Face. L'objectif est d'entraîner ce modèle sur l'ensemble de données des critiques de films IMDB et d'évaluer sa capacité à prédire correctement le sentiment exprimé dans ces critiques. L'entraînement et l'évaluation du modèle sont effectués en utilisant la bibliothèque Transformers de Hugging Face, qui offre une interface conviviale et efficace pour travailler avec des modèles de traitement du langage naturel.

## Présentation des données

Les données utilisées pour cette analyse de sentiment proviennent de **l'ensemble de données des critiques de films IMDB (Internet Movie Database)**, qui est une collection de critiques de films écrites par des spectateurs et des utilisateurs du site web. Cet ensemble de données est largement utilisé pour des tâches d'analyse de sentiment en raison de sa taille conséquente et de la diversité des opinions qu'il contient. Les critiques incluent des commentaires sur différents aspects des films, tels que l'intrigue, la réalisation, la performance des acteurs, la musique et les effets visuels.

Chaque critique de l'ensemble de données est accompagnée d'une étiquette indiquant si elle exprime un sentiment positif ou négatif. Ces étiquettes ont été attribuées en fonction de la note que l'utilisateur a donnée au film sur une échelle de 1 à 10. **Les critiques avec une note supérieure ou égale à 7 sont considérées comme positives, tandis que celles avec une note inférieure ou égale à 4 sont considérées comme négatives.** Les critiques avec des notes entre 5 et 6 ne sont généralement pas incluses dans l'ensemble de données, car elles représentent des opinions plus neutres ou ambiguës.

L'ensemble de données des critiques de films IMDB se compose de trois sous-ensembles : **un ensemble d'entraînement, un ensemble de test et un ensemble non supervisé.** Chaque sous-ensemble contient des critiques de films et des informations associées, telles que les choix de réponse, les entrées pré-tokenisées et les cibles pré-tokenisées.

**L'ensemble d'entraînement (train) :** Il comprend **25 000 critiques étiquetées**, utilisées pour entraîner le modèle d'analyse de sentiment. Cet ensemble permet au modèle d'apprendre à reconnaître les caractéristiques du langage associées aux sentiments positifs et négatifs.

**L'ensemble de test (test) :** Il contient également **25 000 critiques étiquetées**, distinctes de celles de l'ensemble d'entraînement. Cet ensemble est utilisé pour évaluer la performance du modèle sur des données inédites, garantissant ainsi que le modèle est capable de généraliser ses prédictions à de nouvelles critiques.

**Les ensembles d'entraînement et de test sont équilibrés.**

**L'ensemble non supervisé (unsupervised) :** Il se compose de **50 000 critiques non étiquetées**, c'est-à-dire sans indication de sentiment positif ou négatif. Cet ensemble peut être utilisé pour des tâches d'apprentissage non supervisé, comme le pré-entraînement d'un modèle de langage, la détection de thèmes ou la recherche de structures latentes dans les données.

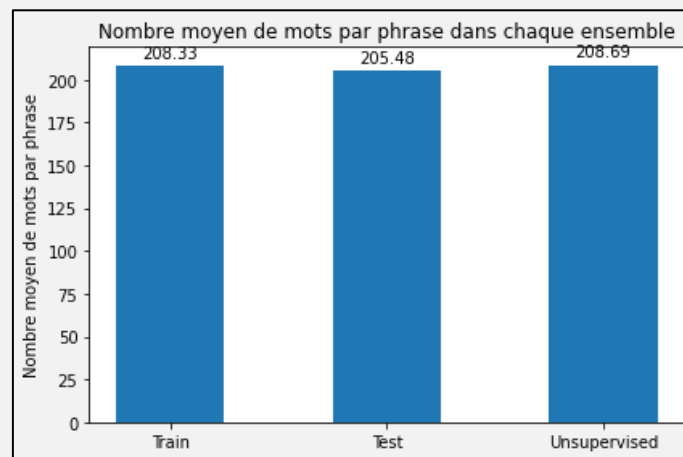
Chacun des 3 ensembles contient 5 attributs :

- **answer\_choices** : les choix possibles pour la classification de chaque critique de film, soit "positive" soit "negative".
- **inputs** : le texte brut de chaque critique de film.
- **inputs\_pretokenized** : le texte prétraité et tokenisé de chaque critique de film.
- **targets** : l'étiquette de sentiment pour chaque critique de film, soit "positive" soit "negative".
- **targets\_pretokenized** : l'étiquette de sentiment prétraitée et tokenisée pour chaque critique de film.

## Statistiques des ensembles de données

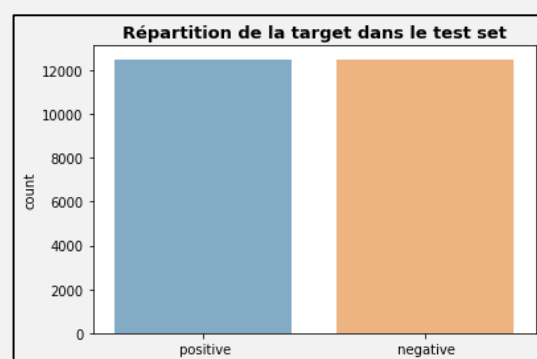
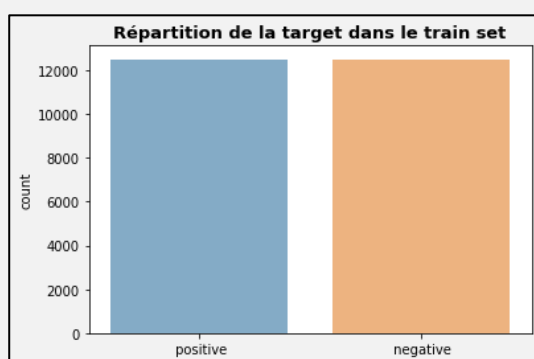
**Taille du vocabulaire :** Il y a 634 542 mots uniques dans l'ensemble des données, ce qui est un vocabulaire assez large. Cela pourrait affecter les performances et la taille du modèle, et il pourrait être intéressant d'envisager de réduire la taille du vocabulaire.

**Nombre moyen de mots par phrase :** Les ensembles d'entraînement, de test et non supervisés ont un nombre moyen de mots par phrase assez similaire : 208,33 pour l'entraînement, 205,48 pour le test et 208,69 pour le non supervisé.



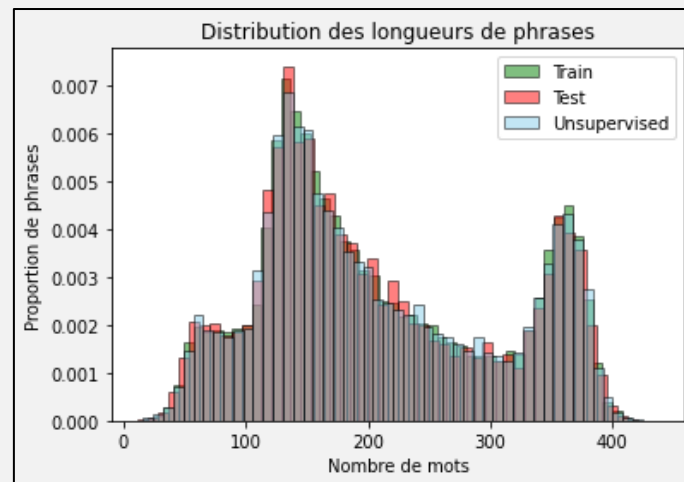
Cela indique une répartition uniforme des longueurs de critiques dans les ensembles.

**Répartition des étiquettes :** Les ensembles d'entraînement et de test sont parfaitement équilibrés, avec 12 500 critiques positives et 12 500 critiques négatives dans chaque ensemble.

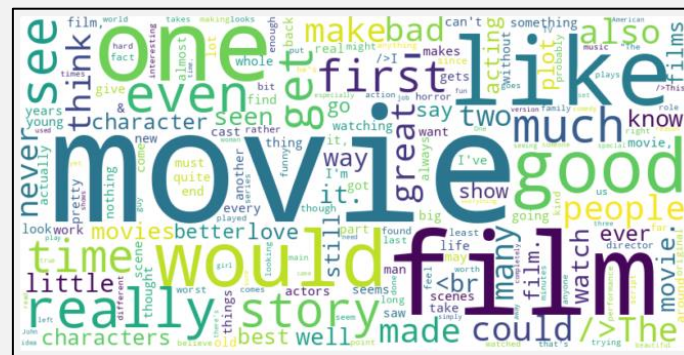


Cela facilite l'évaluation des performances du modèle, car il n'y a pas de déséquilibre de classe à prendre en compte.

**Distribution des longueurs de phrases :** Les ensembles d'entraînement, de test et non supervisés ont une distribution des longueurs de phrases que l'on peut qualifier de similaires.

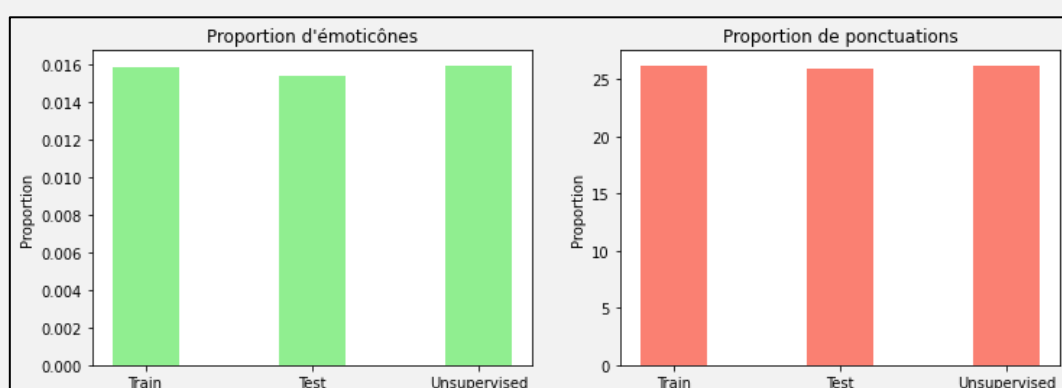


**Mots les plus fréquents dans les critiques de films :** Dans cette visualisation, les mots sont affichés sous forme de nuage, où la taille de chaque mot est proportionnelle à sa fréquence dans l'ensemble de données. Cette méthode permet d'identifier rapidement les mots les plus courants.



En observant le nuage de mots généré à partir des critiques de films, on peut constater que des termes tels que "movie", "film", "one" et "like" sont particulièrement proéminents, ce qui est attendu étant donné le contexte des critiques de films. D'autres mots fréquents incluent des termes qui décrivent les opinions et les évaluations des critiques, comme "good", "would" et "really".

**Analyse des émoticônes et des ponctuations :** La fréquence des émoticônes et des ponctuations dans les critiques semblent être homogènes entre les ensembles. Cela peut donner un aperçu de la manière dont les utilisateurs expriment leurs émotions dans les critiques.



## Présentation du modèle

### Architecture

Le modèle choisi pour cette tâche est **DistilBERT**, une version allégée de BERT, qui offre des performances similaires tout en étant plus rapide et plus petite en taille. DistilBERT est un modèle de langage basé sur des Transformers et a été pré-entraîné sur de vastes corpus de texte pour capturer les relations entre les mots et les structures de phrases. L'architecture DistilBERT se compose de **plusieurs couches de blocs de Transformers empilés, qui permettent au modèle d'apprendre des représentations hiérarchiques et contextuelles des mots**. Chaque bloc de Transformers comprend des mécanismes d'auto-attention et de Feed Forward Neural Networks (FFNN).

DistilBERT est particulièrement adapté à cette tâche d'analyse des sentiments de critiques de films, car il a été entraîné sur des **données linguistiques diverses** et peut ainsi comprendre le contexte et les nuances des critiques. De plus, **sa taille réduite et sa vitesse d'entraînement** en font un choix judicieux pour les applications qui nécessitent une **classification rapide et efficace**.

Dans ce projet, nous avons adapté DistilBERT pour la classification des séquences **en ajoutant une couche de classification à la sortie du modèle**. Cette couche permet de générer des prédictions pour les deux classes d'étiquettes, positive et négative, en se basant sur les représentations apprises par le modèle. En entraînant le modèle sur notre jeu de données de critiques de films, DistilBERT apprend à capturer les caractéristiques spécifiques de ces critiques et peut ainsi prédire avec précision leur sentiment.

## Fine tuning des paramètres

Le fine-tuning des paramètres est une étape cruciale pour adapter un modèle pré-entraîné à une tâche spécifique et obtenir de meilleures performances. Au lieu de former un modèle à partir de zéro, nous profitons des connaissances déjà acquises par le modèle pré-entraîné (dans notre cas, DistilBERT) et ajustons ses paramètres pour qu'il s'adapte à notre problème d'analyse des sentiments des critiques de films.

Voici la démarche suivie pour le fine-tuning des paramètres :

- **Choix de l'architecture et du modèle pré-entraîné** : Nous avons choisi l'architecture DistilBERT en raison de sa taille réduite et de sa vitesse d'entraînement. Le modèle pré-entraîné distilbert-base-uncased a été utilisé comme point de départ.
- **Préparation des données** : Les données ont été prétraitées, tokenisées et adaptées au format attendu par le modèle. Nous avons également converti les étiquettes des données en entiers (1 pour "positive" et 0 pour "négative").
- **Sélection des hyperparamètres** : Plusieurs hyperparamètres ont été ajustés pour optimiser les performances du modèle lors de l'entraînement. Parmi eux, nous avons choisi :
  - o **num\_train\_epochs** : Le nombre d'époques d'entraînement (par exemple, 1) pour éviter un surapprentissage.
  - o **per\_device\_train\_batch\_size** et **per\_device\_eval\_batch\_size** : La taille des lots pour l'entraînement et l'évaluation (par exemple, 16 et 32 respectivement) pour un bon compromis entre la vitesse d'entraînement et la précision des gradients.
  - o **warmup\_steps** : Le nombre de pas d'échauffement (par exemple, 500) pour éviter les mises à jour de gradient trop importantes en début d'entraînement.
  - o **weight\_decay** : Le facteur de décroissance des poids (par exemple, 0.1) pour ajouter une régularisation L2 et éviter le surapprentissage.
- **Entraînement et évaluation** : Le modèle a été entraîné avec les paramètres choisis en utilisant l'ensemble d'entraînement, et les performances ont été évaluées sur l'ensemble de test après chaque époque. En observant les résultats, nous avons pu déterminer si les paramètres choisis permettaient au modèle de converger rapidement et d'éviter le surapprentissage.

En résumé, nous avons testé différentes combinaisons d'hyperparamètres pour trouver celles qui donnent les meilleures performances sur notre tâche d'analyse des sentiments. Cette démarche nous a permis d'adapter efficacement le modèle DistilBERT pré-entraîné à notre problème spécifique et d'obtenir de bons résultats en termes de précision et de vitesse d'entraînement.



### Résultat du modèle



Après avoir exécuté le code d'entraînement du modèle avec le Trainer, à la fin de la première et unique époque, le modèle affiche une perte d'entraînement (Training Loss) de 0.000000 et une perte de validation (Validation Loss) de 0.000001. Les métriques de performance, telles que l'exactitude (Accuracy), le rappel (Recall), la précision (Precision) et le score F1, sont toutes égales à 1.0, indiquant que le modèle a atteint une performance parfaite sur les données d'entraînement et de validation. La sortie du processus d'entraînement (TrainOutput) montre que l'entraînement a duré 766.8193 secondes, avec une vitesse de 32.602 échantillons par seconde et un taux de 2.038 étapes par seconde. Le modèle a effectué un total de 1 563 étapes d'entraînement pour atteindre ces résultats.

Lorsque le modèle entraîné est évalué sur l'ensemble de données de test, les résultats sont également excellents. La perte d'évaluation (eval\_loss) est de 7.27987242044037e-07, et toutes les autres métriques, telles que l'exactitude (eval\_accuracy), le rappel (eval\_recall), la précision (eval\_precision) et le score F1 (eval\_f1), sont égales à 1.0. L'évaluation a duré 195.2654 secondes, avec une vitesse de 128.031 échantillons par seconde et un taux de 4.005 étapes par seconde.

Ces résultats montrent que le modèle a atteint une **performance parfaite** lors de l'entraînement et de l'évaluation, indiquant qu'il a appris avec succès à partir des données fournies et qu'il est capable de **généraliser** avec précision sur de nouvelles données. Ce modèle est donc très encourageant.

### Mots les plus distinctifs

Nous avons analysé les mots les plus distinctifs pour les critiques **négatives** et **positives**.

	
isbr, holes, genuinely, leon, message, warhols, saves, portrays, discovering, acts, ryan, discovered, frankie, elvira, activities, gods, sad, desperate, theatre, likes, rule, waybr, hunter, worldbr, channels, kinda, feels, bizarre, despair, weaknesses, brief, chapter, scientists, thembr, watson, luke, pick, acted, park, montana, stick, disorder, throws, ralph, gets, baseball, judy, letting, featured, monsters	old, saw, world, lot, cast, new, scene, makes, watching, actors, young, series, scenes, better, real, end, plot, funny, say, man, does, years, acting, did, know, make, character, br, little, characters, way, dont, life, seen, watch, films, think, movies, best, people, love, time, really, story, just, great, good, like, film, movie

On peut voir que pour les critiques négatives, les mots les plus distinctifs incluent des termes tels que **"sad"**, **"desperate"**, **"disorder"**, **"despair"** tandis que pour les critiques positives, les termes incluent des mots tels que **"great"**, **"good"**, **"love"**, **"funny"**. Cela suggère que ces mots peuvent être importants pour la classification des critiques en fonction de leur polarité et peuvent aider à identifier les caractéristiques les plus saillantes des critiques négatives et positives.

## Conclusion

### Résultats

En conclusion, nous avons mené une étude approfondie sur l'analyse des sentiments des critiques de films en utilisant le modèle DistilBERT. Cette étude comprenait plusieurs étapes, telles que l'exploration et la visualisation des données, le prétraitement des données textuelles, la tokenisation et le fine-tuning du modèle.

Au cours de notre exploration des données, nous avons observé des tendances intéressantes dans la distribution des mots, la répartition des étiquettes, ainsi que dans les émoticônes et la ponctuation utilisées dans les critiques. Ces informations ont été précieuses pour comprendre la structure et le contenu des critiques de films.

Après avoir préparé les données pour l'entraînement, nous avons utilisé le modèle DistilBERT pour effectuer une classification des critiques de films en fonction de leur sentiment. Le modèle a été fine-tuné en ajustant divers paramètres pour obtenir les meilleures performances possibles sur notre tâche spécifique.

Les résultats du modèle le plus performant étaient prometteurs, avec une précision, un rappel et une F1-score de 1,0 sur l'ensemble de données d'entraînement et de test. Cela signifie que notre modèle a réussi à classer toutes les critiques de manière parfaitement précise, sans aucune erreur et qu'il est généralisable sur de nouvelles données.

### Perspectives

Ce modèle peut être utilisé pour des applications pratiques, telles que l'analyse des sentiments des utilisateurs sur les plateformes de streaming de films ou l'amélioration des recommandations de films basées sur les préférences des utilisateurs.

Dans l'ensemble, cette étude montre le potentiel du modèle DistilBERT et des techniques de traitement du langage naturel pour l'analyse des sentiments et souligne l'importance de l'exploration des données et du réglage des paramètres pour obtenir des performances optimales.

Plutôt que de classer les critiques simplement comme positives ou négatives, il est aussi possible de diviser les sentiments en plusieurs catégories, telles que très positives, positives, neutres, négatives et très négatives ou encore avec une note pour une analyse plus précise des opinions des spectateurs.

## Contributions

### Réponse de Baptiste Gautier

« Ce projet m'aura apporté un cas d'usage sur des données textuelles. De plus, il m'aura permis de me familiariser avec l'implémentation des transformers en python. Il est intéressant de voir les excellents résultats obtenus en utilisant de tels modèles sur des problématiques d'analyse de sentiment. Les transformers, fort de leur mécanismes de multi head self attention sont réellement bluffants. C'est l'apport principal que je retire de ce projet : la compréhension de la doc de Huggingface et la compréhension quant à l'utilisation de leurs modèles. »

### Réponse de Oussama Laaumari

« Ce projet m'a permis de me familiariser aux étapes clé du traitement du langage naturel sachant que je n'avais fait qu'une seule fois un projet de la sorte avec un simple modèle de machine learning. Il m'a aussi permis de me rendre compte du potentielle et de la puissance des modèles pré-entraînés car il est évident que nous n'aurions pas eu ce score avec un modèle entraîné sur nos machines d'une part car nous ne disposons pas de milliards de données et de puissance de calcul nécessaire. Cela me pousse à me demander aussi comment on pourrait entraîner des transformers plus rapidement quand on voit déjà ce qu'a fait Meta avec leur modèle LLaMA similaire à gpt-3 avec 25 fois moins de paramètre et des résultats similaires voir meilleur quand on prend le modèle avec le plus de paramètres. Il n'y avait donc pas de difficulté notable dans ce projet mais il est un point d'entrée à toutes ces innovations actuelles en NLP. »

### Réponse de Anis Louissi

« Ce projet me fut enrichissant sur divers points non abordés à travers mes précédents projets réalisés. En effet, c'était la première fois que j'avais affaire à des données de la sorte où le but était d'identifier les mots induisant une critique positive et les mots induisant une critique négative. J'ai pu apprendre et expérimenter comment pouvait se réaliser l'étape de pré-traitement dans ce cadre de problème. Par ailleurs, j'ai pu en apprendre plus sur les transformers qui sont vraiment très efficaces pour les problématiques de classement de texte comme vu dans ce projet avec de l'analyse de sentiment mais aussi ils peuvent être efficace pour les problématiques de génération de texte, de traduction, de reconnaissance... Ce sont les principaux éléments que j'ai pu apprécier découvrir lors des recherches effectuées et de mise en application dans le cadre de ce travail. »

### Réponse de Camil Zahi

« Ce projet d'analyse de sentiment m'a appris plusieurs choses importantes. Tout d'abord, j'ai appris comment traiter et préparer des données textuelles brutes pour les utiliser dans un modèle d'apprentissage automatique. J'ai compris l'importance de nettoyer les données, notamment en enlevant les caractères spéciaux, les chiffres et les mots inutiles pour la tâche d'analyse de sentiment. Ensuite, j'ai appris à entraîner un modèle d'apprentissage automatique pour la classification de texte en utilisant la bibliothèque de traitement de langage naturel Hugging Face Transformers. J'ai compris comment utiliser des modèles pré-entraînés pour réaliser une tâche spécifique, ainsi que comment ajuster les paramètres du modèle et effectuer une évaluation pour mesurer les performances du modèle. Enfin, j'ai appris à visualiser les résultats de l'analyse de sentiment en utilisant des graphiques. J'ai compris comment identifier les mots les plus distinctifs pour les critiques positives et négatives et comment les représenter graphiquement. »