UNIVERSITÉ CÔTE D'AZUR | MASTER INFORMATIQUE

Projet de Classification et Analyse des Sentiments dans les Critiques de Films

Rapport

Traitement Automatique du Texte en IA

Steven ESSAM EDWAR AZIZ : 22309059

Master 1 Informatique
2023/2024

Contents

Introduction

Le domaine de l'analyse des sentiments dans les critiques de films connaît une croissance significative ces dernières années. Avec la prolifération des plateformes en ligne, les utilisateurs partagent activement leurs opinions sur divers films, créant ainsi une mine d'informations. Cependant, la gestion de cette quantité considérable de données nécessite des méthodes automatisées pour extraire des informations pertinentes. Le projet s'inscrit dans ce contexte, visant à développer un modèle d'apprentissage automatique capable de classifier les sentiments exprimés dans les critiques de films.

L'objectif principal du projet est de concevoir, implémenter et évaluer des modèles d'apprentissage automatique pour l'analyse des sentiments. Plus précisément, l'objectif est d'élaborer des solutions capables d'identifier les sentiments positifs et négatifs exprimés dans les critiques de films. Cette tâche implique une maîtrise du langage naturel et l'utilisation de techniques avancées de traitement du texte.

Pour atteindre ces objectifs, différentes approches d'apprentissage automatique ont été sélectionnées, incluant le Support Vector Machine (SVM), le Random Forest et le Naive Bayes. Chacune de ces méthodes présente des avantages et des inconvénients, permettant une approche flexible pour classer les sentiments dans les critiques de films. Ces algorithmes seront appliqués sur un jeu de données choisi, et l'évaluation des résultats sera réalisée à l'aide de métriques de performance. Cette variété de méthodologies est expressément conçue pour assurer une compréhension approfondie du comportement spécifique de chaque algorithme dans le contexte de l'analyse des sentiments.

Le code source complet, développé en Python version 3.10.12, dans un environnement Jupyter Book et sur le système d'exploitation Linux, comprend l'implémentation, l'évaluation des modèles, une analyse détaillée des résultats, ainsi que des explications détaillées pour chaque étape, est accessible sur GitHub : https://github.com/stevenessam/TATIA_Projet_Steven_ESSAM_22309059_Groupe_17

Jeu de Données

Le jeu de données choisi pour le projet provient du site Kaggle et intitulé "IMDB Dataset of 50K Movie Reviews" [1]. Il propose une collection variée de critiques cinématographiques, totalisant environ 49 582 instances, où chaque critique est annotée avec un sentiment, soit positif ou négatif. Cette diversité d'opinions reflète différentes expériences et émotions associées aux films, formant ainsi une base solide pour notre analyse des sentiments.

Caractéristique	**Valeur**
Nombre total d'instances	Environ 49 582 critiques de films
Séparation des données	80% pour l'entraînement, 20% pour les tests
Annotations	Sentiments Positifs et Négatifs
Exemples	- "*I recall the scariest scene was the big bird*..." (Sentiment: Négatif) - "*Petter Mattei's 'Love in the Time of Money' is*..." (Sentiment: Positif)

Un prétraitement des données était nécessaire avant d'effectuer toute analyse. Ce processus comprend plusieurs étapes, notamment la suppression des balises HTML à l'aide de la bibliothèque BeautifulSoup [9], la conversion en minuscules pour assurer une uniformité, la suppression de la ponctuation pour éliminer tout bruit potentiel, et enfin, la suppression des "stop words" (mots fréquents mais non informatifs) à l'aide de la bibliothèque NLTK [3]. Ces étapes sont essentielles pour éliminer tout bruit potentiel et permettre à nos modèles d'apprentissage automatique de saisir les subtilités des critiques cinématographiques.

La répartition des données en ensembles d'entraînement et de test a été réalisée en utilisant la bibliothèque sklearn [2]. Une division de 80% pour l'entraînement et 20% pour les tests a été adoptée, garantissant ainsi que les modèles soient formés sur des données variées et testés sur des exemples non vus auparavant. Cette approche équilibrée vise à évaluer la capacité de généralisation des modèles.

Pour illustrer, prenons deux exemples du jeu de données. La première critique commence par "*I recall the scariest scene was the big bird*...", et elle est annotée avec un sentiment négatif. D'autre part, la deuxième critique, débutant par "*Petter Mattei's 'Love in the Time of Money' is*...", est associée à un sentiment positif. Ces exemples représentent la diversité des opinions dans le domaine cinématographique et serviront de base pour l'évaluation approfondie des modèles.

Conception et Implémentation

L'élaboration du projet d'analyse des sentiments cinématographiques a nécessité des choix quant aux algorithmes d'apprentissage automatique à adopter. Trois modèles ont été sélectionnés pour leur diversité et leurs caractéristiques distinctes : le Support Vector Machine (SVM) [5], le Random Forest [6], et le Naive Bayes [7]. Cette méthode variée permet une compréhension approfondie du comportement de chaque algorithme dans le contexte de la tâche spécifique.

En ce qui concerne le SVM, le choix s'est porté sur un noyau linéaire en raison de sa simplicité et de son efficacité pour traiter des données de texte. Pour le modèle Random Forest, il s'appuie sur un ensemble d'arbres de décision, fournissant ainsi une puissance considérable tout en minimisant le risque de surajustement. Le Naive Bayes a été retenu en raison de sa simplicité, particulièrement adaptée à la classification de texte où l'indépendance entre les caractéristiques peut être une approximation raisonnable.

Les paramètres de chaque modèle ont été sélectionnés pour maintenir un équilibre entre précision et généralisation. Pour le SVM, le noyau linéaire a été privilégié, et le Random Forest a été configuré avec 100 estimateurs pour garantir une robustesse adéquate. En ce qui concerne le Naive Bayes, les paramètres par défaut ont été maintenus.

L'implémentation de ces modèles s'est réalisée à l'aide d'outils et de bibliothèques Python, notamment sklearn, NLTK (Natural Language Toolkit) pour le traitement du langage naturel, et BeautifulSoup pour le prétraitement des balises HTML. La volonté d'utiliser des ressources adaptées à chaque étape du projet est évidente dans cette diversité d'outils.

Le prétraitement des critiques cinématographiques a constitué une phase essentielle du travail. L'élimination des balises HTML a été effectuée grâce à BeautifulSoup, suivie d'une conversion du texte en minuscules pour assurer une uniformité dans l'ensemble des données. La suppression de la ponctuation et des "stop words" a permis d'obtenir des représentations textuelles plus focalisées sur l'essentiel.

La vectorisation des données textuelles s'est opérée par la méthode Term Frequency-Inverse Document Frequency (TF-IDF) [8], une approche attribuant à chaque mot un poids en fonction de sa fréquence dans un document particulier et de sa rareté dans l'ensemble du corpus. Cette technique permet de représenter chaque critique sous forme de vecteur numérique, favorisant ainsi une meilleure compréhension par les modèles.

Enfin, l'entraînement des modèles s'est déroulé sur l'ensemble d'entraînement. C'est là que les algorithmes ont appris à classifier les critiques en fonction de leurs sentiments associés. L'évaluation s'est faite sur l'ensemble de test, et les performances ont été évaluées à l'aide de métriques standard telles que la précision et le rappel, offrant ainsi un aperçu clair de l'efficacité de chaque modèle dans la tâche d'analyse des sentiments cinématographiques.

Évaluation des Performances

Pour évaluer les performances des modèles d'analyse des sentiments, plusieurs métriques ont été utilisées, dont la précision, le rappel et le score F1. Ces indicateurs fournissent une perspective détaillée sur la capacité des algorithmes à classifier correctement les critiques cinématographiques en fonction de leurs sentiments associés.

En examinant le modèle SVM (Support Vector Machine), une performance globale impressionnante a été constatée, avec une précision atteignant 88.62%. Le rapport de classification révèle un équilibre entre la précision et le rappel pour les classes positives et négatives, renforcé par des scores F1 élevés. L'exactitude globale de 89% confirme la robustesse de l'approche basée sur SVM.

```
Précision du modèle : 88.62%

Rapport de classification :
              precision    recall  f1-score   support

    negative       0.89      0.87      0.88      4961
    positive       0.88      0.90      0.89      5039

    accuracy                           0.89     10000
   macro avg       0.89      0.89      0.89     10000
weighted avg       0.89      0.89      0.89     10000
```

Quant au modèle Random Forest, il a démontré une solide performance avec une précision de 84.84%. L'analyse du rapport de classification souligne un équilibre entre la précision et le rappel, et des scores F1 équilibrés pour les deux classes. L'exactitude globale de 85% atteste de l'efficacité du modèle Random Forest dans la tâche d'analyse des sentiments.

```
Précision du modèle Random Forest : 84.84%

Rapport de classification Random Forest :
              precision    recall  f1-score   support

    negative       0.84      0.86      0.85      4961
    positive       0.86      0.84      0.85      5039

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
weighted avg       0.85      0.85      0.85     10000
```

Le modèle Naive Bayes a également présenté des résultats encourageants, avec une précision de 85.13%. L'examen du rapport de classification montre une équité entre précision et rappel, accompagnée de scores F1 élevés pour les deux classes. La précision globale de 85 % confirme la fiabilité de l'approche basée sur Naive Bayes.

```
Précision du modèle Naive Bayes : 85.13%

Rapport de classification Naive Bayes :
              precision    recall  f1-score   support

    negative       0.85      0.84      0.85      4961
    positive       0.85      0.86      0.85      5039

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
weighted avg       0.85      0.85      0.85     10000
```

En comparant ces résultats, bien que le SVM surpasse légèrement les autres en termes de précision globale, la différence n'est pas significative. Ainsi, Random Forest et Naive Bayes demeurent des choix solides pour la tâche d'analyse des sentiments. L'efficacité globale des modèles témoigne de la réussite de l'approche polyvalente.

Il est important de souligner que l'efficacité de chaque modèle peut varier en fonction de la nature spécifique des critiques cinématographiques et des nuances linguistiques. L'analyse approfondie des erreurs et des cas limites s'avère essentielle pour affiner les modèles.

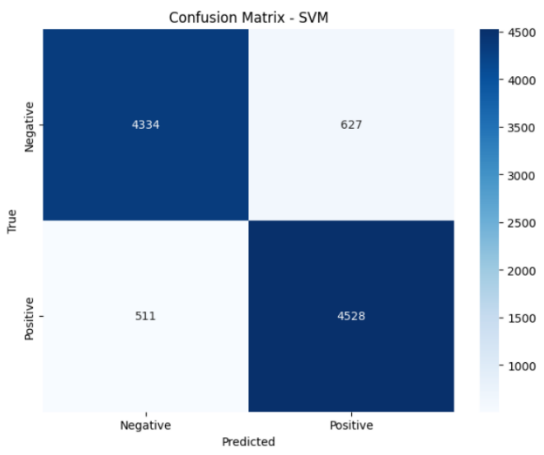
L'évaluation détaillée des performances des modèles a été réalisée à travers l'analyse des matrices de confusion [10], offrant ainsi une compréhension fine des résultats obtenus.

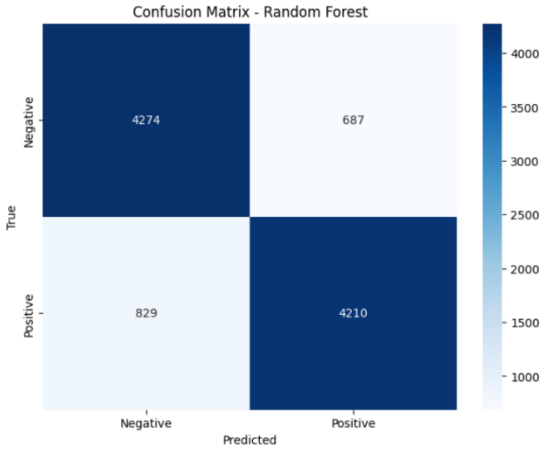
Le modèle SVM a démontré une performance globalement solide. Avec 4 334 vrais négatifs (VN) et 4 528 vrais positifs (VP), il a bien réussi à classifier les observations des deux classes. Néanmoins, il a présenté 627 faux positifs (FP) et 511 faux négatifs (FN), indiquant des erreurs dans la prédiction des deux catégories.

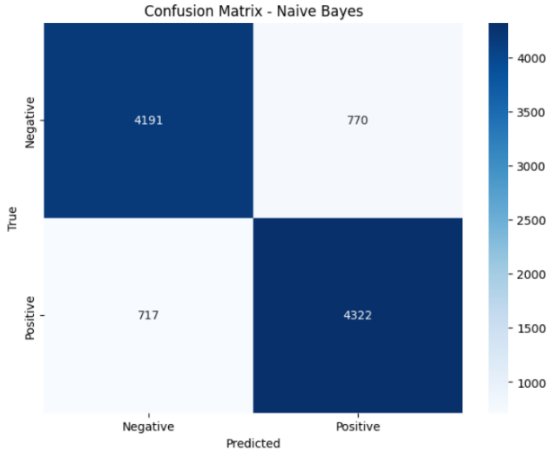
Concernant le modèle Random Forest, ses 4 274 (VN) et 4 210 (VP) montrent une capacité correcte à classer les deux catégories. Cependant, avec 687 (FP) et 829 (FN), le modèle montre des faiblesses dans la distinction entre les classes, s'égarent dans les prédictions positives et négatives.

Le modèle Naive Bayes a également affiché une performance respectable avec 4 191 (VN) et 4 322 (VP). Malgré cela, les 770 (FP) et 717 (FN) présentent des défis dans la précision des prédictions, surtout lorsqu'il s'agit de distinguer entre les sentiments positifs et négatifs.

Ces analyses révèlent la nécessité d'une compréhension approfondie des erreurs spécifiques de chaque modèle. Les (FP) soulignent les cas où le modèle prédit à tort un sentiment, tandis que les (FN) indiquent les cas où il manque d'identifier correctement le sentiment réel.







Analyse des Erreurs

L'examen des erreurs commises par le modèle de Support Vector Machine (SVM) révèle des points intéressants sur ses performances et les difficultés liées à la tâche complexe d'analyse des sentiments. Lors de cette analyse, l'attention a été portée sur des exemples de prédictions incorrectes.

Un exemple de prédiction incorrecte est la critique avec l'index 33553, initialement étiquetée comme positive mais prédite comme négative par le modèle SVM. L'analyse des termes utilisés, tels que "liked" et "look", montre la difficulté du modèle à interpréter ces mots dans leur contexte, soulignant ainsi la subtilité du langage et la nécessité de prendre en compte les nuances.

Ces erreurs ne sont pas uniques, et une tendance similaire se dégage dans d'autres cas de prédictions incorrectes. Il semble que le modèle puisse être influencé par des mots ou des expressions susceptibles d'être interprétés de manière ambiguë ou ayant des connotations différentes.

```
Exemples de prédictions incorrectes :
                                                  Review True Label  \
33553  [really, liked, summerslam, due, look, arena, ...   positive
49498  [okay, didnt, get, purgatory, thing, first, ti...   positive
6113   [production, quality, cast, premise, authentic...   positive
15118  [movie, released, originally, soft, x, apparen...   positive
33109  [three, kid, born, solar, eclipse, turn, vile,...   positive

      Predicted Label
33553        negative
49498        negative
6113         negative
15118        negative
33109        negative
```

Une autre facette de l'analyse des erreurs s'articule autour des sources potentielles de ces dernières. L'analyse des erreurs révèle que la complexité de l'analyse des sentiments peut provenir de plusieurs facteurs. Les subtilités du langage, l'utilisation de sarcasme, ou encore la présence d'éléments ambigus dans les critiques peuvent tous contribuer à des prédictions incorrectes.

En analysant de plus près les erreurs des modèles Random Forest et Naive Bayes, des cas spécifiques ont été identifiés où ces modèles ont rencontré des difficultés. Par exemple, le modèle Random Forest a prédit incorrectement une critique négative comme étant positive, interprétant mal la phrase "*actor performance leave much desired lacking emotional depth credibility undermining immersion narrative*".

De manière similaire, le modèle Naive Bayes a également effectué une prédiction erronée en attribuant une tonalité positive à une critique négative. Dans ce cas, la phrase "*despite polished special effect direction suffers lack artistic cohesion giving impression visual element added disjointedly*" a été mal interprétée.

Ces exemples démontrent les défis permanents de l'analyse des sentiments, notamment la compréhension des nuances du langage et la gestion des phrases ambiguës qui peuvent influencer de manière significative les prédictions du modèle.

Conclusion

Le projet d'analyse des sentiments que j'ai réalisé a permis une exploration approfondie de divers modèles d'apprentissage automatique, dont le Support Vector Machine (SVM), le Random Forest et le Naive Bayes. Cette approche variée avait pour objectif de mieux comprendre la complexité liée à la tâche d'analyse des sentiments. Le modèle SVM a émergé comme le plus performant, atteignant une précision notable de 88.62%. Cette robustesse s'est révélée également dans la capacité du modèle à généraliser sur de nouvelles critiques.

Cependant, malgré la performance supérieure du SVM, il est important de souligner que les modèles Random Forest et Naive Bayes ont présenté des résultats respectables, avec des précisions de 84.84% et 85.13%, respectivement. Chaque modèle a révélé des forces et des faiblesses distinctes, soulignant l'importance de choisir un modèle en adéquation avec les spécificités de la tâche.

Malgré ces résultats positifs, cette approche comporte des limitations. La nature subjective du langage naturel et la présence de termes ambigus demeurent des défis à relever. L'analyse des erreurs a mis en évidence la nécessité d'améliorer le prétraitement des données afin d'obtenir une compréhension plus fine du langage.

Des perspectives d'amélioration incluent l'exploration de techniques avancées de prétraitement, comme l'utilisation de modèles de langage pré-entraînés, et l'expérimentation avec des architectures de modèles plus complexes, tels que les réseaux neuronaux. Ces approches pourraient potentiellement renforcer la capacité des modèles à traiter des données complexes et à saisir la subtilité du langage naturel.

En somme, ce projet marque le début d'une exploration approfondie de l'analyse des sentiments, et des améliorations peuvent être envisagées pour affiner la compréhension des nuances du langage et renforcer les performances des modèles face à des données plus complexes.

Références

[1] Kaggle: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews

[2] Sklearn: https://scikit-learn.org/0.21/documentation.html

[3] NLTK: https://www.nltk.org/

[4] Python : https://docs.python.org/3/

[5] Support Vector Machines (SVM) : https://scikit-learn.org/stable/modules/svm.html

[6] Random Forest: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[7] Naive Bayes: https://scikit-learn.org/stable/modules/naive_bayes.html

[8] TF-IDF : https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

[9] Beautiful Soup: https://beautiful-soup-4.readthedocs.io/en/latest/

[10] Confusion matrix : https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

[11] Projet GitHub : https://github.com/stevenessam/TATIA_Projet_Steven_ESSAM_22309059_Groupe_17