

# **SAE 6 EMS 01- MODELISATION STATISTIQUE POUR LES DONNEES COMPLEXES ET LE BIG DATA**

**REVE KEMY DIAVOU DIAVOU  
GAOUSSOU MAOULOU KONTA**

DEPARTEMENT SCIENCE DES DONNEES

ENCADRÉES PAR : Kamel Mouna

UNIVERSITE DE PERPIGNAN

BUT3 SCIENCE DES DONNEES

ANNEE UNIVERSITAIRE 2023-2024

IUT DE CARCASSONNE



## Sommaire

1. Introduction .....	4
2. Préparation et Analyse des Données.....	4
3. Traitement Linguistique .....	4
4. Application des Fonctions et Enrichissement des Données .....	5
4.1 Préparation pour le Modèle de Machine Learning .....	5
4.2 Construction et Évaluation du Modèle de Machine Learning.....	5
5. Résultat.....	6
5.1 Courbe ROC .....	7
5.2 Log Loss .....	8
5.3 Matrice de Confusion .....	8
6. Rapport de Classification.....	8
7. Métriques Additionnelles.....	8
8. Exactitude .....	9
Conclusion.....	9

## Table des figures

Figure 1. Courbe ROC .....	6
Figure 2. Résultat du code .....	7
Figure 3. Résultat du code classification.....	7
Figure 4. Résultat du code performance du modèle .....	7

## 1. Introduction

Ce rapport présente une analyse des commentaires de films et séries, visant à extraire des caractéristiques linguistiques et sentimentales pour prédire si les commentaires sont positifs ou négatifs. En utilisant des techniques de traitement du langage naturel et d'apprentissage automatique, nous explorons les données textuelles pour en extraire des informations significatives.

Nous commençons par préparer les données en les chargeant à partir d'un fichier CSV, en les explorant pour comprendre leur structure, et en les nettoyant pour éliminer les bruits et les valeurs aberrantes. Ensuite, nous analysons la structure grammaticale des commentaires et évaluons leur sentiment général.

Nous transformons ensuite ces caractéristiques linguistiques et sentimentales en données numériques pour entraîner un modèle de machine Learning. Enfin, nous évaluons la performance du modèle en utilisant diverses métriques.

## 2. Préparation et Analyse des Données

Nous commençons par charger les commentaires sur les films et séries à partir d'un fichier CSV. Ces données constituent la base de notre analyse et comprennent des informations telles que les commentaires et les notes attribuées aux contenus audiovisuels. Cette étape initiale nous permet d'accéder aux données nécessaires pour notre analyse ultérieure.

## 3. Traitement Linguistique

Dans cette phase de traitement linguistique, plusieurs étapes sont effectuées pour analyser en profondeur les commentaires sur les films et séries :

**Chargement du Modèle de Langue :** Le script utilise la bibliothèque `spacy` pour charger un modèle linguistique français. Ce modèle permet une analyse précise de la structure grammaticale des phrases, ce qui est essentiel pour comprendre la syntaxe et la sémantique des commentaires.

**Définition des Fonctions de Traitement :** Plusieurs fonctions sont définies pour extraire des informations linguistiques pertinentes :

**Compter les Émoticônes :** Une fonction est mise en place pour compter le nombre d'émoticônes présentes dans chaque commentaire, offrant ainsi un aperçu de l'expression émotionnelle de l'auteur.

**Compter les Mots :** Une autre fonction est créée pour calculer le nombre total de mots dans chaque commentaire, ce qui permet d'évaluer la longueur et la richesse lexicale des commentaires.

**Analyser la Structure Grammaticale :** Le script analyse les verbes, les entités nommées et les adjectifs présents dans chaque commentaire. Cette analyse offre des insights sur la complexité et le style linguistique utilisé par les commentateurs.

**Déterminer le Temps Dominant des Verbes :** Une fonction est dédiée à l'identification du temps verbal dominant (présent, passé, futur) dans chaque commentaire, ce qui peut fournir des indices sur la perspective temporelle de l'auteur.

**Sentiment des Verbes :** Enfin, le script utilise l'outil TextBlob pour évaluer le caractère positif ou négatif des verbes présents dans les commentaires. Cette analyse de sentiment permet de comprendre les sentiments exprimés par les auteurs à l'égard des films et séries.

Ces étapes de traitement linguistique permettent d'extraire des informations précieuses sur le contenu des commentaires, allant de l'expression émotionnelle à la syntaxe et au sentiment général.

#### 4. Application des Fonctions et Enrichissement des Données

Ces fonctions sont appliquées sur l'ensemble du jeu de données pour créer de nouvelles colonnes représentant ces caractéristiques linguistiques et sentimentales.

##### 4.1 Préparation pour le Modèle de Machine Learning

a. **Création de la Variable Cible :** Le script génère une colonne cible basée sur les notes attribuées aux films ou séries, transformant l'analyse en un problème de classification binaire (positif/négatif).

b. **Encodage et Sélection des Caractéristiques :** Il transforme les caractéristiques catégorielles en numériques et sélectionne des variables pertinentes pour l'apprentissage du modèle.

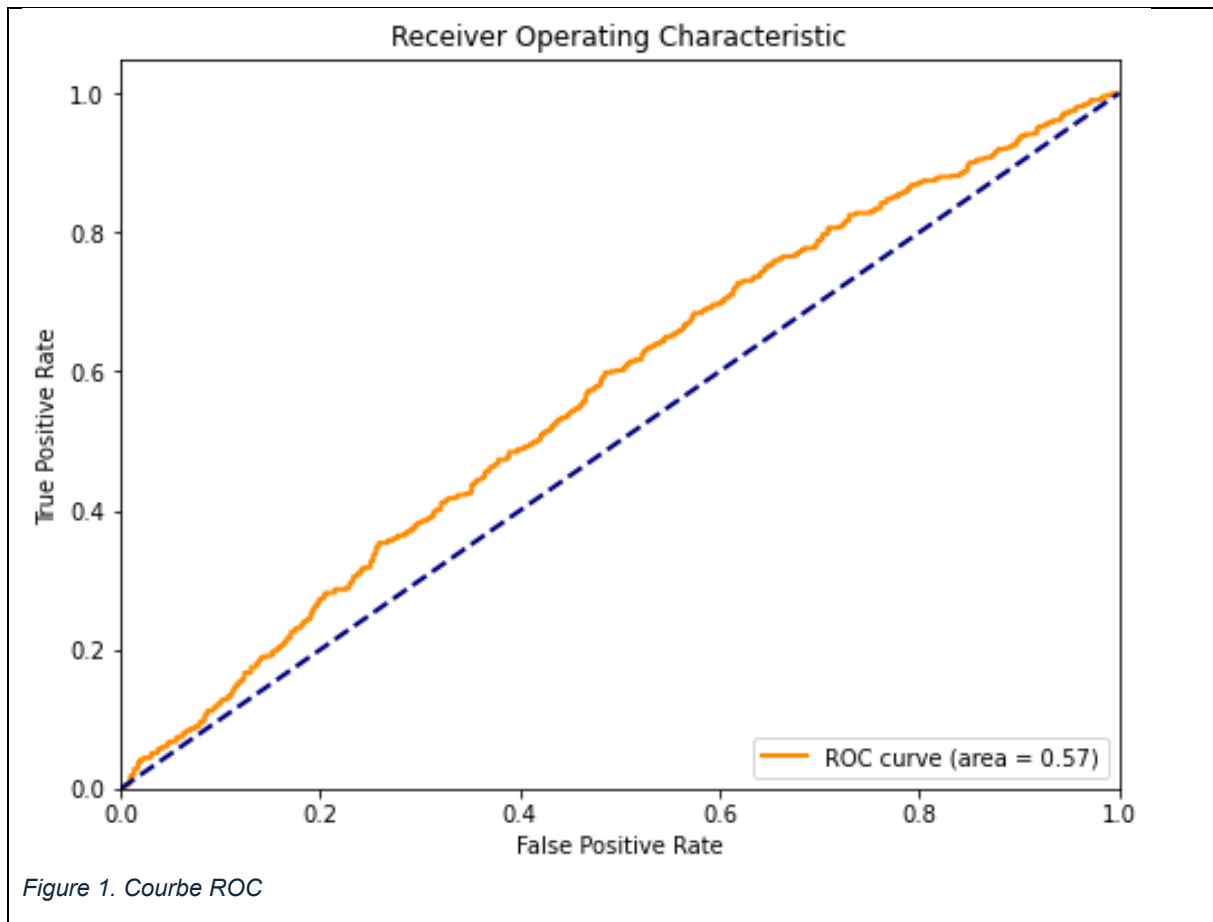
c. **Division des Données :** Les données sont divisées en ensembles d'entraînement et de test pour évaluer la performance du modèle.

##### 4.2 Construction et Évaluation du Modèle de Machine Learning

a. **Entraînement du Modèle :** Un modèle de régression logistique est entraîné sur l'ensemble d'entraînement. Ce type de modèle est bien adapté pour les problèmes de classification binaire.

b. **Évaluation du Modèle :** Le script utilise plusieurs métriques pour évaluer la performance du modèle, telles que l'exactitude, la matrice de confusion, le rapport de classification, la précision, le rappel, le score F1, et l'aire sous la courbe ROC (Receiver Operating Characteristic). Chacune de ces métriques donne un aperçu différent de la performance du modèle.

c. Visualisation des Résultats : La courbe ROC est tracé, fournissant une représentation visuelle de la performance du modèle en termes de taux de vrais positifs par rapport aux faux positifs.



d. Log Loss : Enfin, le log loss est calculé, offrant une mesure de la performance du modèle en termes de probabilités prédites.

## 5. Résultat

```
In [42]:
...: conf_matrix = confusion_matrix(y_test, y_pred)
...: print(f"Matrice de confusion :\n{conf_matrix}")
Matrice de confusion :
[[323 428]
 [247 503]]
```

Figure 2. Résultat du code

```
In [43]:
...: class_report = classification_report(y_test, y_pred)
...: print(f"Rapport de classification :\n{class_report}")
Rapport de classification :
              precision    recall  f1-score   support

     0           0.57       0.43       0.49         751
     1           0.54       0.67       0.60         750

 accuracy          0.55
 macro avg         0.55
 weighted avg      0.55
```

Figure 3. Résultat du code classification

```
In [44]:
...: precision = precision_score(y_test, y_pred)
...: recall = recall_score(y_test, y_pred)
...: f1 = f1_score(y_test, y_pred)
...: print(f"Précision : {precision}")
...: print(f"Rappel : {recall}")
...: print(f"Score F1 : {f1}")
Précision : 0.5402792696025779
Rappel : 0.6706666666666666
Score F1 : 0.5984533016061868
```

Figure 4. Résultat du code performance du modèle

L'analyse de la performance d'un modèle de classification peut être faite à travers plusieurs métriques, et la visualisation de la courbe ROC (Receiver Operating Characteristic) est l'une d'elles. Examinons en détail les résultats obtenus.

### 5.1 Courbe ROC

La courbe ROC illustre la capacité du modèle à différencier les classes. La ligne orange représente la performance du modèle, tandis que la ligne bleue pointillée représente la performance d'un modèle aléatoire. L'aire sous la courbe (AUC) est de 0.57, ce qui est proche de 0.5, l'AUC d'un modèle qui effectuerait des prédictions au hasard. Plus l'AUC est élevée (proche de 1), meilleure est la performance du modèle. Dans ce cas, l'AUC indique une performance à peine meilleure qu'un modèle aléatoire.

## 5.2 Log Loss

Le log loss, ou perte logarithmique, est une mesure de performance pour les modèles de classification où la sortie est une probabilité entre 0 et 1. Une valeur de log loss de 0.687 est assez élevée, ce qui suggère que le modèle fait des prédictions qui sont souvent éloignées de la vérité.

## 5.3 Matrice de Confusion

La matrice de confusion montre les performances du modèle de classification en termes de prédictions correctes et incorrectes réparties comme suit :

- \*\*Vrais négatifs (VN) : \*\* 323
- \*\*Faux positifs (FP) : \*\* 428
- \*\*Faux négatifs (FN) : \*\* 247
- \*\*Vrais positifs (VP):\*\* 503

Idéalement, les nombres les plus élevés devraient être sur la diagonale (VN et VP), ce qui indiquerait un grand nombre de prédictions correctes. Ici, le modèle a tendance à avoir un taux plus élevé de faux positifs que de vrais négatifs, et un nombre légèrement plus élevé de vrais positifs par rapport aux faux négatifs.

## 6. Rapport de Classification

Le rapport de classification offre un résumé des mesures clés :

- La précision est la proportion de prédictions positives qui sont réellement positives. Ici, pour la classe 1, elle est de 0.54, ce qui est modérément bas.
- Le rappel indique la proportion de positifs réels qui ont été correctement identifiés par le modèle. Pour la classe 1, le rappel est de 0.67, ce qui est relativement meilleur que la précision.
- Le score F1 est la moyenne harmonique de la précision et du rappel et tente d'équilibrer les deux. Un score F1 de 0.60 est considéré comme moyen.

## 7. Métriques Additionnelles

- Précision : 0.54 - Indique que lorsque le modèle prédit une classe positive, il est correct 54% du temps.
- Rappel : 0.67 - Signifie que le modèle est capable de détecter 67% des vrais cas positifs.
- Score F1 : 0.60 - Offre un équilibre entre la précision et le rappel, suggérant que le modèle est relativement équilibré.

## 8. Exactitude

L'exactitude (accuracy) globale du modèle est de 0.55, ce qui signifie que le modèle prédit correctement la classe de l'entrée 55% du temps. Pour les modèles de classification binaire, surtout lorsque les classes sont équilibrées, ce taux est peu élevé et seulement légèrement meilleur que le hasard.

## Conclusion

En résumé, ce script effectue une analyse détaillée et multifacette des commentaires de films et séries, combinant traitement du langage naturel et apprentissage automatique pour prédire le sentiment des commentaires. Il va au-delà d'une simple analyse de sentiment en intégrant des aspects linguistiques tels que la structure grammaticale et le temps verbal, offrant ainsi une compréhension plus nuancée des données textuelles.