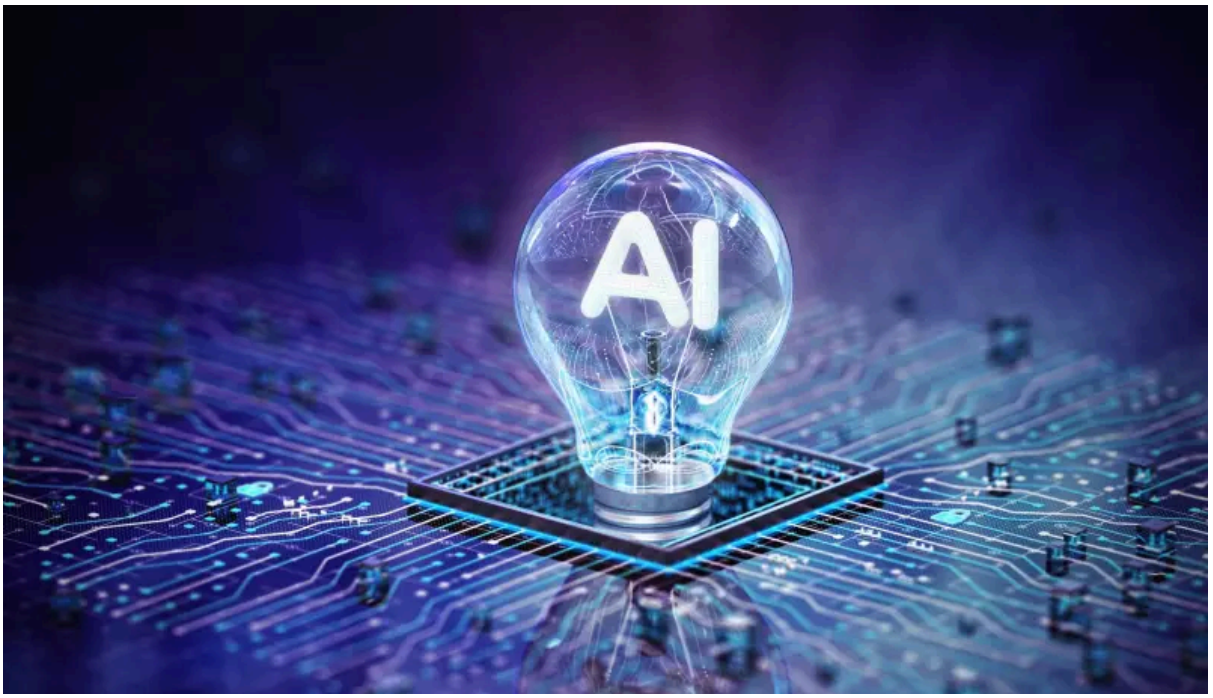
IUT Toulouse 2025 - 2026

# Rapport de SAE : Analyse de Sentiment et Classification des Avis Yelp




Groupe 3
Lucie MASSELIN
Flavien ALCAZAR
Aminata Oumou Rassoul NGOM
Tsinjo Mirantsoa RANDRIANARISON RATSIANDAVANA

# Tables des matières

# Introduction

Dans le cadre de cette Situation d'Apprentissage et d'Évaluation (SAE), il nous a été demandé de mettre en place une chaîne complète de traitement de données textuelles (NLP) afin d'analyser les avis clients issus du dataset Yelp. L'objectif principal est de transformer des données brutes en informations structurées, capables de prédire la note d'un établissement (score) et la polarité d'un commentaire (positif, négatif ou neutre).

Bien que notre phase d'exploration ait porté sur l'ensemble du jeu de données (**Business**, **User** et **Review**) pour bien comprendre les corrélations entre les variables, le cœur de notre travail de modélisation s'est concentré sur le fichier **Review.json**, contenant la matière textuelle nécessaire à l'entraînement de nos modèles.

Pour mener à bien ce projet, notre groupe s'est organisé de manière à couvrir tout le spectre des technologies actuelles en IA. Chaque membre a pris la responsabilité de deux à trois approches spécifiques. Cette stratégie nous a permis de ne pas nous disperser et d'approfondir chaque méthodologie (du Machine Learning classique à l'IA Générative) afin de comparer efficacement les performances et la pertinence de chaque modèle face aux attentes pédagogiques.

# 1. Modèles de Machine Learning Classique : Tsinjo

Tsinjo s'est concentré sur les approches statistiques classiques, qui servent de "baseline" (référence) pour évaluer le gain de performance des modèles plus complexes.

## A. LinearSVC (Support Vector Classification)

Le LinearSVC est un modèle de machine learning qui cherche à trouver l'hyperplan séparateur optimal entre les différentes classes (positif/négatif). Il est particulièrement efficace sur des vecteurs de textes de grande dimension (TF-IDF).

## B. Logistic Regression (Régression Logistique)

Malgré son nom, c'est un modèle de classification qui estime la probabilité qu'un avis appartienne à une catégorie. C'est un modèle robuste, rapide à entraîner et très interprétable.

```
Résultats modèles de machine learning

1. Prédiction de la polarité des avis (positif, négatif, neutre)

================================================
PERFORMANCES BOW
================================================
     accuracy  precision  recall  f1-score
lr     0.8255      0.806  0.8255    0.8143
svm    0.8255      0.806  0.8255    0.8143

================================================
PERFORMANCES TFIDF
================================================
     accuracy  precision  recall  f1-score
lr     0.8255     0.8060  0.8255    0.8143
svm    0.8082     0.7926  0.8082    0.7996

================================================
PERFORMANCES NGRAM
================================================
     accuracy  precision  recall  f1-score
lr     0.8518     0.8237  0.8518    0.8241
svm    0.8425     0.8144  0.8425    0.8227

================================================
PERFORMANCES WORD2VEC
================================================
     accuracy  precision  recall  f1-score
lr     0.8200     0.7857  0.8200    0.7922
svm    0.8182     0.7740  0.8182    0.7789
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

```
2. Prédiction des notes des avis

================================================
PERFORMANCES BOW
================================================
     accuracy  precision  recall  f1-score
lr     0.6128     0.5955  0.6128    0.6025
svm    0.5728     0.5654  0.5728    0.5683

================================================
PERFORMANCES TFIDF
================================================
     accuracy  precision  recall  f1-score
lr     0.6445     0.6040  0.6445    0.6117
svm    0.6218     0.5901  0.6218    0.6009

================================================
PERFORMANCES NGRAM
================================================
     accuracy  precision  recall  f1-score
lr     0.6485     0.6093  0.6485    0.6156
svm    0.6185     0.5895  0.6185    0.5995

================================================
PERFORMANCES WORD2VEC
================================================
     accuracy  precision  recall  f1-score
lr     0.6225     0.5832  0.6225    0.5905
svm    0.6215     0.5767  0.6215    0.5635
```

**Résultat des performances du modèle par rapport à la prédiction des notes**

# 2. Deep Learning & Approches Spatiales : Lucie

Lucie a exploré les réseaux de neurones profonds en utilisant des représentations vectorielles pour capturer des relations plus complexes que le simple comptage de mots.

## A. MLP (Multi-Layer Perceptron)

Le MLP est un réseau de neurones artificiels "dense". Ici, il est utilisé avec un encodage statistique de type BM25 ou TF-IDF. Il permet de classer des données fixes en apprenant des poids sur chaque terme de manière non-linéaire.

```
Deep Learning performance Lucie avec 20 000 dataset
--------------------
--- Modèle MLP ---
--------------------
MLP score (étoiles) :
Précision (Accuracy) : 0.6559
F1-Score (ajusté selon la quantité d'avis par étoile)   : 0.6502
Perte (Log Loss)     : 0.7769

Rapport détaillé :
              precision    recall  f1-score   support

    1 étoile       0.75      0.77      0.76      1506
    2 étoile       0.39      0.49      0.44       781
    3 étoile       0.38      0.44      0.41       933
    4 étoile       0.49      0.48      0.49      2109
    5 étoile       0.83      0.77      0.80      4660

    accuracy                           0.66      9989
   macro avg       0.57      0.59      0.58      9989
weighted avg       0.67      0.66      0.66      9989
```

**Résultat des performances du modèle par rapport à la prédiction des notes**

```
------------------
MLP polarité (label) :
Précision (Accuracy) : 0.7890
F1-Score (Ajusté)    : 0.8187
Perte (Log Loss)     : 0.5013

Rapport détaillé :
              precision    recall  f1-score   support

      neutre       0.30      0.70      0.42      1000
     négatif       0.86      0.78      0.81      2315
     positif       0.97      0.81      0.88      6674

    accuracy                           0.79      9989
   macro avg       0.71      0.76      0.70      9989
weighted avg       0.87      0.79      0.82      9989
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

## B. CNN (Convolutional Neural Network)

Habituellement utilisé pour l'image, le CNN est ici appliqué au texte pour son aspect spatial. Il utilise des filtres pour repérer des "n-grammes" (groupes de mots) clés, un peu comme il repérerait des formes ou des objets dans une image.

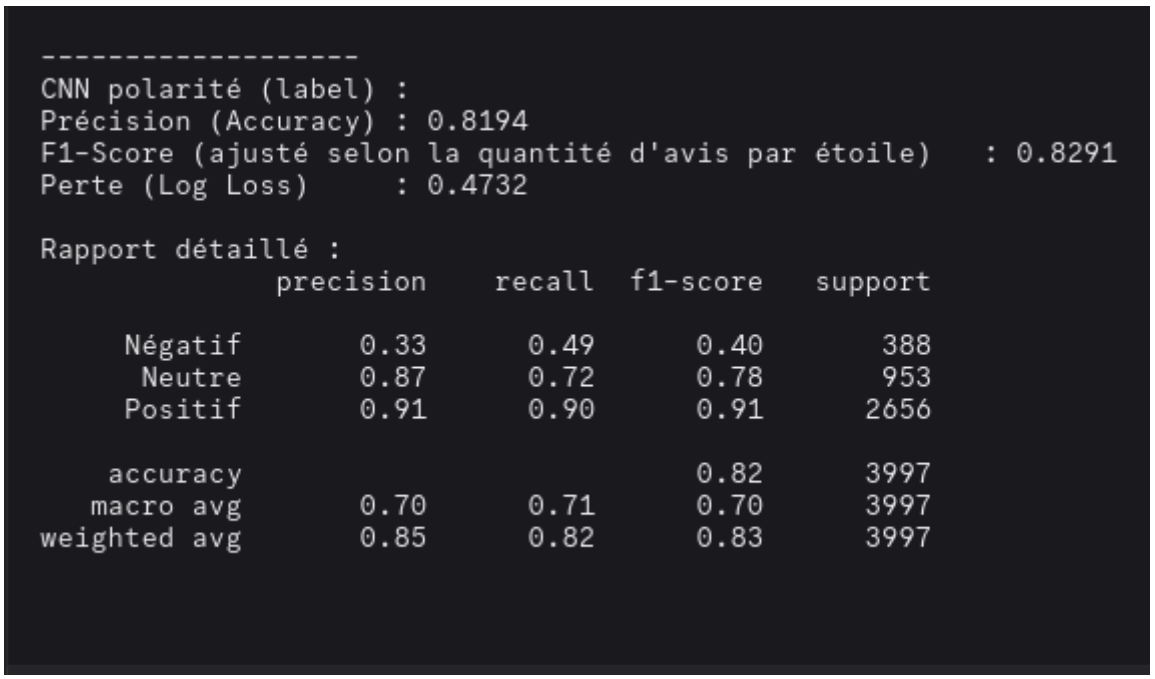
```
------------------
--- Modèle CNN ---
------------------
CNN score (étoiles) :
Précision (Accuracy) : 0.6159
F1-Score (Ajusté selon la quantité d'avis par étoile) : 0.6239
Perte (Log Loss)     : 0.9109

Rapport détaillé :
              precision    recall  f1-score   support

    1 étoile       0.78      0.69      0.73      1527
    2 étoile       0.34      0.48      0.40       764
    3 étoile       0.31      0.53      0.39       917
    4 étoile       0.42      0.32      0.36      2065
    5 étoile       0.81      0.76      0.78      4718

    accuracy                           0.62      9991
   macro avg       0.53      0.56      0.53      9991
   weighted avg       0.64      0.62      0.62      9991
```

**Résultat des performances du modèle par rapport à la prédiction des notes**



```
------------------
CNN polarité (label) :
Précision (Accuracy) : 0.8194
F1-Score (ajusté selon la quantité d'avis par étoile)   : 0.8291
Perte (Log Loss)     : 0.4732

Rapport détaillé :
              precision    recall  f1-score   support

     Négatif       0.33      0.49      0.40       388
      Neutre       0.87      0.72      0.78       953
     Positif       0.91      0.90      0.91      2656

    accuracy                           0.82      3997
   macro avg       0.70      0.71      0.70      3997
weighted avg       0.85      0.82      0.83      3997
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

**Comparaison des performances :** MLP vs CNN Prédiction de la polarité (Positif/Négatif/Neutre) : Avantage au CNN. Le CNN se montre plus précis (Accuracy de 81.9% contre 78.9% pour le MLP). Grâce à ses filtres de convolution, il repère l'ordre des mots (n-grammes), ce qui lui permet de saisir le contexte local (ex: la différence entre "good" et "not good"). Le MLP, en traitant les mots indépendamment, perd cette nuance. Prédiction des notes (1 à 5 étoiles) : Robustesse du MLP. Sur la classification à 5 classes, le MLP l'emporte (65.6% contre 61.6%). L'approche TF-IDF lui permet de capter parfaitement l'intensité globale et la fréquence des mots forts ("amazing", "horrible"), là où le CNN a tendance à se perdre dans les détails locaux de la phrase.

**Analyse des comportements :** Polarisation et "zone grise" L'étude des rapports de classification révèle que les modèles peinent à traiter la nuance humaine : Excellente détection des extrêmes : Les deux modèles identifient très facilement les avis tranchés (1 et 5 étoiles), avec par exemple un F1-score de 0.77 pour le MLP. Les marqueurs de haine ou d'adoration sont mathématiquement très clairs.

**L'effondrement sur les notes intermédiaires :** Pour les notes de 2, 3 ou 4 étoiles, les performances chutent drastiquement (F1-scores autour de 0.40 - 0.49). Une phrase modérée comme "The food was okay but a bit expensive" envoie des signaux contradictoires au réseau. Face à cette ambiguïté sémantique et à la frontière très subjective entre un 3 et un 4 étoiles, les modèles n'arrivent pas à trancher et forcent souvent leurs prédictions vers les extrêmes.

---

# 3. Deep Learning Séquentiel & Attentionnel : Flavien

Flavien a travaillé sur les architectures de pointe basées sur le contexte et l'ordre des mots, utilisant notamment les technologies de type "Transformer".

## A. LSTM (Long Short-Term Memory)

Le LSTM est une architecture de réseau de neurones récurrents (RNN) conçue spécifiquement pour résoudre le problème de la "perte de mémoire" sur les textes longs.

- **Fonctionnement :** Grâce à un système de "portes" (gates), le modèle décide quelles informations de l'avis sont importantes à conserver et lesquelles peuvent être oubliées.
- **Utilité :** C'est un encodeur très efficace pour comprendre la structure d'une phrase complexe où le sens d'un mot peut dépendre d'un autre mot situé beaucoup plus tôt dans le texte.

```
--- RAPPORT DE CLASSIFICATION POLARITE LSTM ---
              precision    recall  f1-score   support

     Négatif       0.72      0.74      0.73      1333
     Positif       0.80      0.69      0.74      1334
      Neutre       0.57      0.63      0.60      1333

    accuracy                           0.69      4000
   macro avg       0.70      0.69      0.69      4000
weighted avg       0.70      0.69      0.69      4000
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

```
--- Rapport de classification LSTM (5 classes) ---
              precision    recall  f1-score   support

    1 étoile       0.64      0.57      0.60       800
   2 étoiles       0.39      0.50      0.44       800
   3 étoiles       0.38      0.41      0.40       800
   4 étoiles       0.44      0.37      0.40       800
   5 étoiles       0.62      0.55      0.58       800

    accuracy                           0.48      4000
   macro avg       0.49      0.48      0.48      4000
weighted avg       0.49      0.48      0.48      4000
```

**Résultat des performances du modèle par rapport à la prédiction des notes**

## B. BERT (Bidirectional Encoder Representations from Transformers)

BERT est un modèle attentionnel. Contrairement aux modèles précédents, il lit le texte dans les deux sens simultanément. Il encode le sens profond et contextuel de chaque mot (par exemple, il distingue le sens du mot "avocat" selon le contexte).

```
--- RAPPORT DE CLASSIFICATION POLARITE BERT ---
              precision    recall  f1-score   support

     Négatif       0.87      0.82      0.84       659
     Positif       0.89      0.82      0.86       704
      Neutre       0.68      0.78      0.73       637

    accuracy                           0.81      2000
   macro avg       0.81      0.81      0.81      2000
weighted avg       0.82      0.81      0.81      2000
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

```
--- Rapport de classification BERT (5 classes) ---
              precision    recall  f1-score   support

    1 étoile       0.78      0.70      0.74       402
   2 étoiles       0.53      0.62      0.57       392
   3 étoiles       0.53      0.49      0.51       397
   4 étoiles       0.58      0.59      0.58       406
   5 étoiles       0.76      0.76      0.76       403

    accuracy                           0.63      2000
   macro avg       0.64      0.63      0.63      2000
weighted avg       0.64      0.63      0.63      2000
```

**Résultat des performances du modèle par rapport à la prédiction des notes**

Le LSTM est un réseau de neurones récurrents conçu pour traiter des données séquentielles. Contrairement aux réseaux classiques, il utilise un système de "portes" pour décider quelles informations textuelles conserver ou oublier sur de longues distances.

• Résultats Polarité (3 classes) : Le LSTM obtient une accuracy de 0,69. Il se montre performant sur les classes "Positif" et "Négatif", mais peine sur la classe "Neutre" avec un F1-score de 0,60. Images résultats LSTM polarité

• Résultats Notes (5 classes) : La précision chute à 0,48. On observe que le modèle a du mal à distinguer les nuances entre 2, 3 et 4 étoiles, où les scores de précision sont particulièrement bas (autour de 0,38 - 0,44). Images résultats LSTM Notes B. BERT (Bidirectional Encoder Representations from Transformers) BERT représente une révolution dans le traitement du langage naturel (NLP) grâce à son architecture basée sur l'attention. Il lit le texte dans les deux sens simultanément, capturant ainsi le sens profond et contextuel de chaque mot.

• Résultats Polarité (3 classes) : BERT surpasse nettement le LSTM avec une accuracy de 0,81. La compréhension de la classe "Neutre" progresse fortement avec un F1-score de 0,73.

Images résultats BERT polarité

• Résultats Notes (5 classes) : Même sur cette tâche complexe, BERT maintient une accuracy de 0,63. Il est particulièrement efficace pour identifier les avis extrêmes (1 étoile et 5 étoiles) avec des précisions respectives de 0,78 et 0,76. Images résultats BERT Notes

**Conclusion** : L'analyse de nos modèles révèle un compromis frappant entre la précision et les ressources nécessaires. Si le modèle LSTM s'avère extrêmement rapide avec un entraînement de seulement 4 minutes, il plafonne à une accuracy de 0,69 sur la polarité et peine sur les nuances des 5 classes (0,48). À l'opposé, BERT exige un investissement computationnel bien plus lourd, nécessitant 1 heure d'entraînement sur le dataset complet. Cependant, ce coût est justifié par une supériorité technique indiscutable : BERT atteint 0,81 d'accuracy sur la polarité et 0,63 sur les notes, offrant une finesse de compréhension contextuelle que la structure séquentielle du LSTM ne peut égaler. En conclusion, si le LSTM reste utile pour des tests rapides, BERT s'impose pour une analyse de sentiments de haute précision.

---

# 4. IA Générative et Modèles de Langage : Aminata

Aminata a exploré l'utilisation des modèles de langage de pointe (LLM) pour évaluer leur capacité à traiter les avis Yelp sans nécessiter un ré-entraînement lourd, en s'appuyant sur le transfert d'apprentissage.

## A. RoBERTa (Robustly Optimized BERT approach)

RoBERTa est une version améliorée et plus robuste du modèle BERT. Contrairement au BERT classique, il a été entraîné sur des volumes de données beaucoup plus vastes et avec des séquences plus longues. Aminata l'a utilisé pour sa grande précision dans la compréhension du contexte global d'un avis, ce qui permet de capter des nuances de langage (ironie, doubles négations) que des modèles plus simples pourraient manquer lors de la prédiction de la polarité.

```
Accuracy polarité : 0.831

Classification report (polarité) :

              precision    recall  f1-score   support

    negative       0.81      0.80      0.81      1165
     neutral       0.23      0.14      0.17       503
    positive       0.89      0.95      0.92      3332

    accuracy                           0.83      5000
   macro avg       0.64      0.63      0.63      5000
weighted avg       0.80      0.83      0.82      5000
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

```
Accuracy globale des notes : 0.601

Classification report détaillé par nombre d'étoiles :
--------------------------------------------------------
              precision    recall  f1-score   support

           1      0.581     0.856     0.692       780
           2      0.000     0.000     0.000       385
           3      0.233     0.139     0.174       503
           4      0.000     0.000     0.000       975
           5      0.638     0.961     0.767      2357

    accuracy                          0.601      5000
   macro avg      0.291     0.391     0.327      5000
weighted avg      0.415     0.601     0.487      5000
```

**Résultat des performances du modèle par rapport à la prédiction des notes**

## La subjectivité du neutre

L'analyse des premiers résultats a révélé une difficulté majeure : la classe "neutre" n'est absolument pas traitée de manière efficace par les modèles, avec une précision tombant à 0.00. En approfondissant les recherches sur le dataset, il apparaît que la neutralité est une notion extrêmement subjective.

Ce phénomène s'explique par plusieurs facteurs observés lors de l'entraînement :

- **Ambiguïté sémantique :** Des expressions comme *"the food can be better"* sont interprétées différemment selon le contexte ; pour certains, cela traduit une attente non comblée (négatif), pour d'autres, une simple observation constructive (neutre).
- **Instabilité des notes intermédiaires :** Le modèle peine à se stabiliser sur les notes pivots. Il a une forte tendance à basculer vers les extrêmes : il prédit plus facilement un 1 qu'un 2, et un 5 qu'un 4. La note 3, censée représenter le neutre, devient un "no man's land" statistique où le modèle ne parvient pas à trancher.

Face à ce constat, nous avons pris la décision stratégique de supprimer la catégorie neutre. Cette étape a permis de supprimer le bruit généré par ces avis ambigus afin de maximiser la performance du modèle sur les pôles positif et négatif, garantissant ainsi une classification beaucoup plus fiable et une extraction d'aspects plus nette.

## B. Google T5 (Text-to-Text Transfer Transformer) : flan-t5-base

Pour l'extraction d'aspects et la classification, Aminata a utilisé le modèle **flan-t5-base**. Ce modèle se distingue par son approche "Text-to-Text" : chaque tâche (qu'il s'agisse de donner une note ou d'extraire un sentiment) est formulée comme une instruction textuelle générant une réponse textuelle.

### 1. L'approche Zero-shot

Dans cette configuration, nous soumettons l'avis au modèle accompagné d'une commande directe (ex: *"Is the sentiment of this review positive or negative?"*). Le modèle utilise ses connaissances pré-entraînées pour répondre sans avoir jamais vu d'exemples spécifiques à notre dataset Yelp.

### 2. L'approche Few-shot

Ici, nous intégrons dans le "prompt" (la requête) deux ou trois exemples d'avis déjà annotés. Cette méthode est cruciale pour **flan-t5-base** car elle lui "apprend" en temps réel le format de sortie souhaité. Cependant, même après, nous observons quasiment le même comportement, montrant réellement que ce modèle n'est pas adapté aux tâches qu'on lui propose.

```
Accuracy polarité : 0.699
Classification report (polarité) :

              precision    recall  f1-score   support

    negative       0.00      0.00      0.00        28
    positive       0.70      1.00      0.82        65

    accuracy                           0.70        93
   macro avg       0.35      0.50      0.41        93
weighted avg       0.49      0.70      0.58        93
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

```
Accuracy globale des notes : 0.20
Classification report détaillé par nombre d'étoiles :

              precision    recall  f1-score   support

           1       0.20      1.00      0.33        20
           2       0.00      0.00      0.00         8
           3       0.00      0.00      0.00         7
           4       0.00      0.00      0.00        22
           5       0.00      0.00      0.00        43

    accuracy                           0.20       100
   macro avg       0.04      0.20      0.07       100
weighted avg       0.04      0.20      0.07       100
```

**Résultat des performances du modèle par rapport à la prédiction des notes**

## C. DistilBERT (Version optimisée de BERT)

En complément, Aminata a testé DistilBERT, une version plus légère, plus rapide et moins gourmande en mémoire que le modèle BERT original, tout en conservant environ 95% de ses performances.

- **Rôle** : Ce modèle a été utilisé principalement pour la prédiction de la polarité et des notes. Étant un modèle d'encodeur, il excelle à comprendre le sens global d'une phrase pour la classer dans une catégorie précise.
- **Observation** : Bien que plus efficace que T5 pour la classification pure, DistilBERT a lui aussi buté sur la fameuse "zone grise" des avis. Comme les autres modèles, il s'est montré très performant sur les avis extrêmement positifs ou négatifs, mais a montré ses limites sur les notes intermédiaires (2, 3, 4).

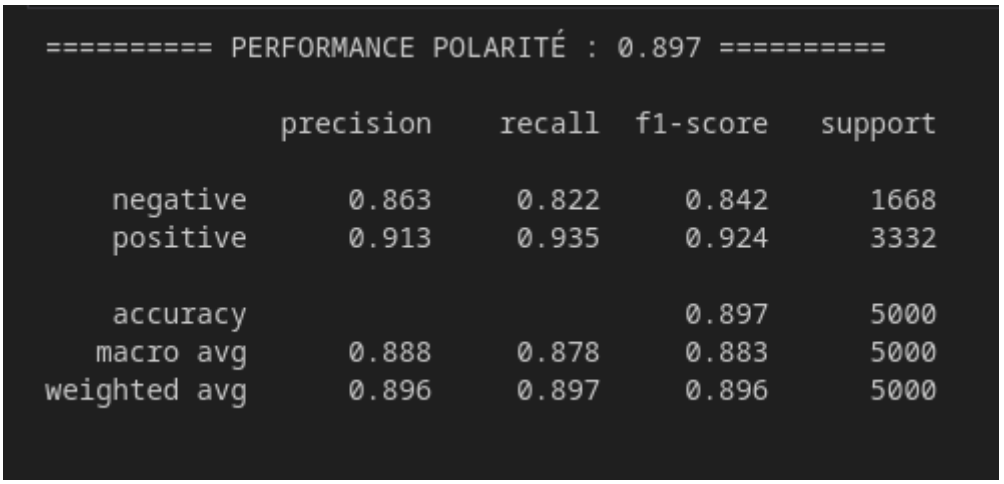
```
========== PERFORMANCE RATING (1-5) : 0.603 ==========

              precision    recall  f1-score   support

           1      0.480     0.978     0.644       780
           2      0.000     0.000     0.000       385
           3      0.000     0.000     0.000       503
           4      0.000     0.000     0.000       975
           5      0.661     0.956     0.782      2357

    accuracy                          0.603      5000
   macro avg      0.228     0.387     0.285      5000
weighted avg      0.386     0.603     0.469      5000
```

**Résultat des performances du modèle par rapport à la prédiction des notes**



```
========== PERFORMANCE POLARITÉ : 0.897 ==========

              precision    recall  f1-score   support

    negative      0.863     0.822     0.842      1668
    positive      0.913     0.935     0.924      3332

    accuracy                          0.897      5000
   macro avg      0.888     0.878     0.883      5000
weighted avg      0.896     0.897     0.896      5000
```

**Résultat des performances du modèle par rapport à la prédiction de la polarité**

---

# Conclusion

Ce travail collaboratif nous a permis de constater l'évolution des performances en fonction de la complexité des modèles. Si le Machine Learning classique reste très performant pour de la classification binaire simple, les modèles de Deep Learning (BERT, GRU) et l'IA Générative apportent une finesse de compréhension indispensable pour l'extraction d'aspects et l'analyse de nuances complexes dans les commentaires Yelp. La complémentarité de nos recherches nous offre aujourd'hui une vision globale des solutions NLP applicables au monde du business.