**Nom : DIALLO**

**Prénoms : Mamadou Arabiou**

# ***Projet 1 : Topic Modeling des Avis des Produits***

## PAETIE II - Clustering non supervisé des documents pour identifier des topics et mots-clés

1. Clustering avec KMeans

```
Distribution des documents par cluster (KMeans) :
cluster
4    356
2    271
1    185
3    182
0      6
Name: count, dtype: int64
```

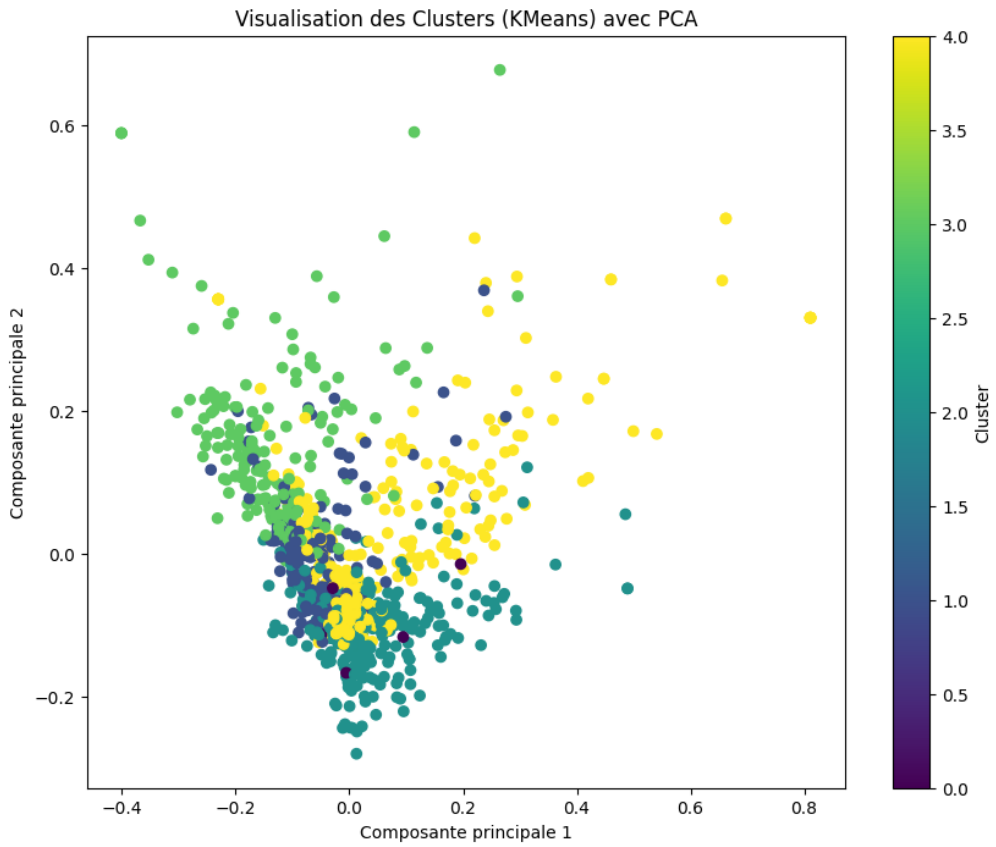
Les résultats montrent la distribution des documents à travers les cinq clusters formés par l'algorithme KMeans.

Pourquoi Cette méthode :

**KMeans** est un algorithme de **clustering non supervisé** efficace pour partitionner les données en **k groupes** basés sur des similarités entre les documents. Dans ce cas, il semble avoir bien séparé les documents en différents clusters, bien que la taille des clusters soit très variable.

Les résultats montrent une distribution des documents inégale entre les clusters. Le **Cluster 4** regroupe la majorité des documents, suggérant un thème principal. Les **Clusters 2** et **1** contiennent également un nombre conséquent de documents, représentant probablement des sous-thèmes. Le **Cluster 3** est plus petit, reflétant une catégorie spécifique, tandis que le **Cluster 0**, avec seulement 6 documents, pourrait indiquer un groupe minoritaire avec des caractéristiques uniques.

- Visualisation des clusters




**Partie III – Analyse des sentiments des avis clients**

1. Choix Effectués

a. Modèle Utilisé :

- Modèle : nlptown/bert-base-multilingual-uncased-sentiment (BERT Multilingue Pré-entraîné)
- Ce modèle a été choisi pour sa capacité à gérer plusieurs langues et sa spécialisation dans la classification des sentiments.

b. Techniques et Outils :

- Librarie principale : *Transformers* de Hugging Face pour le chargement du modèle pré-entraîné.
- Pandas pour la gestion des données (chargement et manipulation des fichiers JSON).
- Seaborn et Matplotlib pour les visualisations.

- Scipy.stats pour le calcul de la corrélation de Pearson.

c. Prétraitement des Données :

- Limitation de la taille des textes à 512 caractères pour s'adapter aux contraintes du modèle BERT.
- Conversion des étiquettes de sortie (sentiment) en scores numériques (de "1 étoile" à "5 étoiles").

d. Évaluation :

- Calcul de la corrélation entre les notes réelles (issues des données) et les prédictions pour évaluer les performances.
- Corrélation observée : **0.81**, indiquant une relation forte entre les prédictions du modèle et les notes réelles.

2. Problématiques Rencontrées

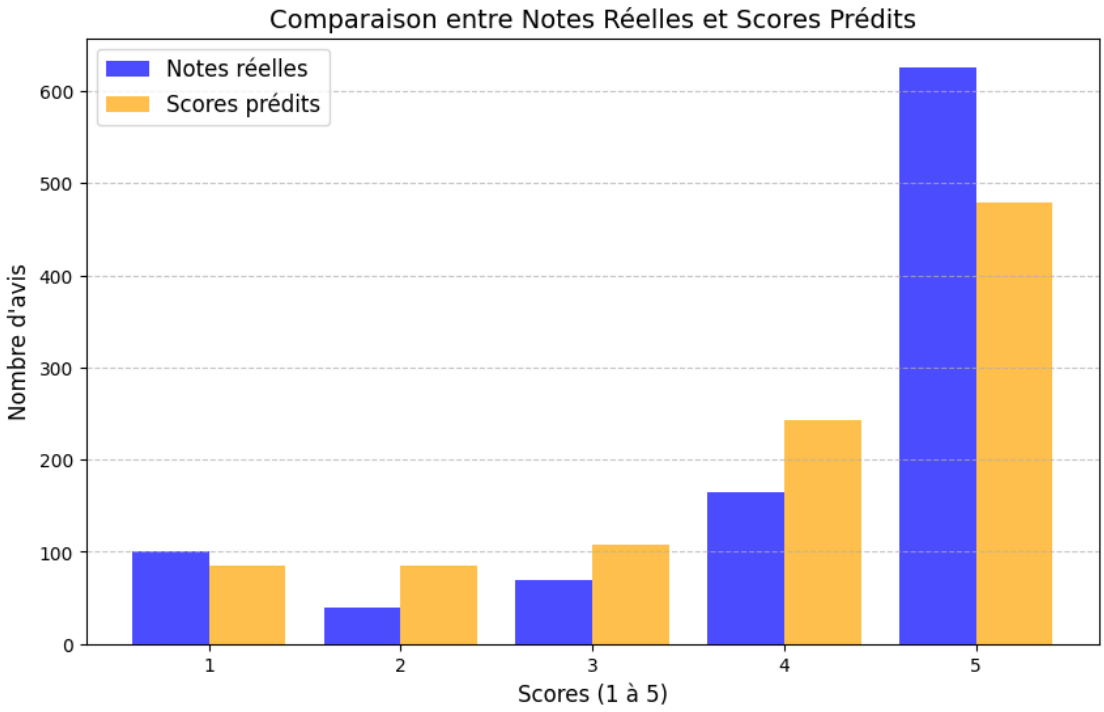
1. Longueur des Textes
    - Certains avis dépassaient la limite de 512 tokens imposée par le modèle BERT. La solution retenue a été de tronquer les textes.
2. Déséquilibre des Classes
    - Les données présentaient un déséquilibre, avec une surreprésentation des avis de 5 étoiles. Cela pourrait biaiser le modèle en le poussant à sur-prédire des scores élevés.
3. Erreurs de Prédiction pour Certaines Classes
    - Les prédictions du modèle montrent des erreurs plus fréquentes pour des notes moyennes (comme 3 étoiles), probablement dues à la difficulté de discriminer des sentiments nuancés.
4. Interprétation des Résultats
    - Bien que la corrélation de 0.81 soit satisfaisante, elle ne garantit pas une performance parfaite. Une matrice de confusion a été envisagée pour mieux comprendre les erreurs.

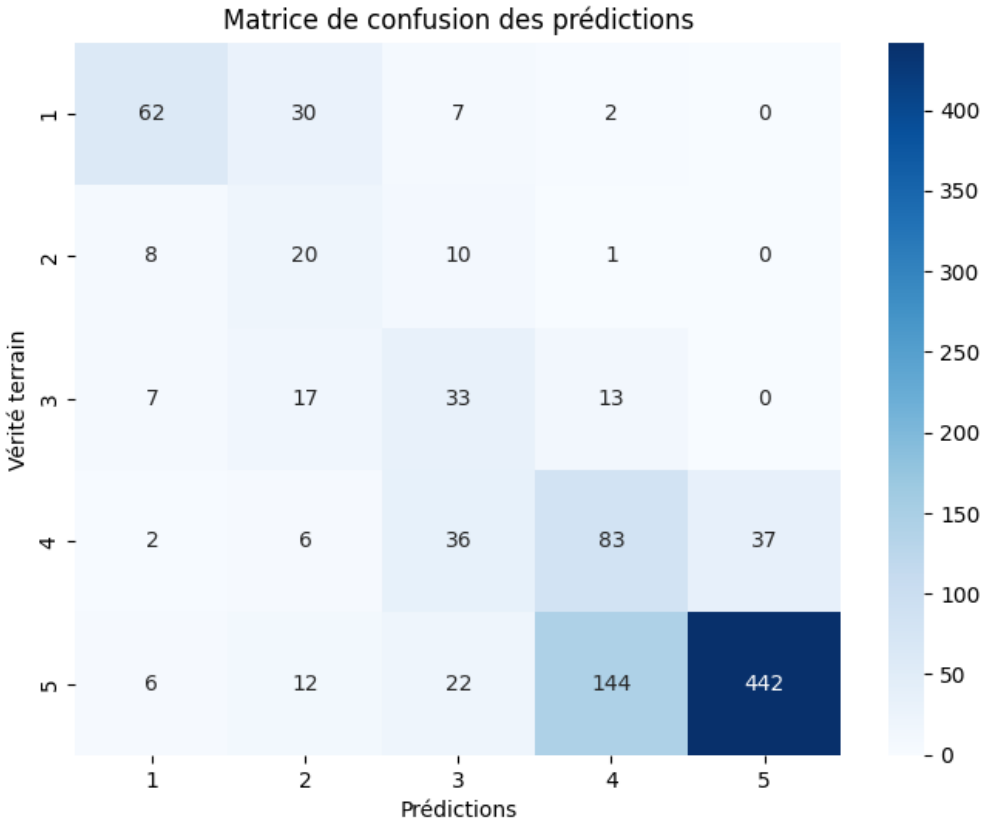
Résultat Représentatif

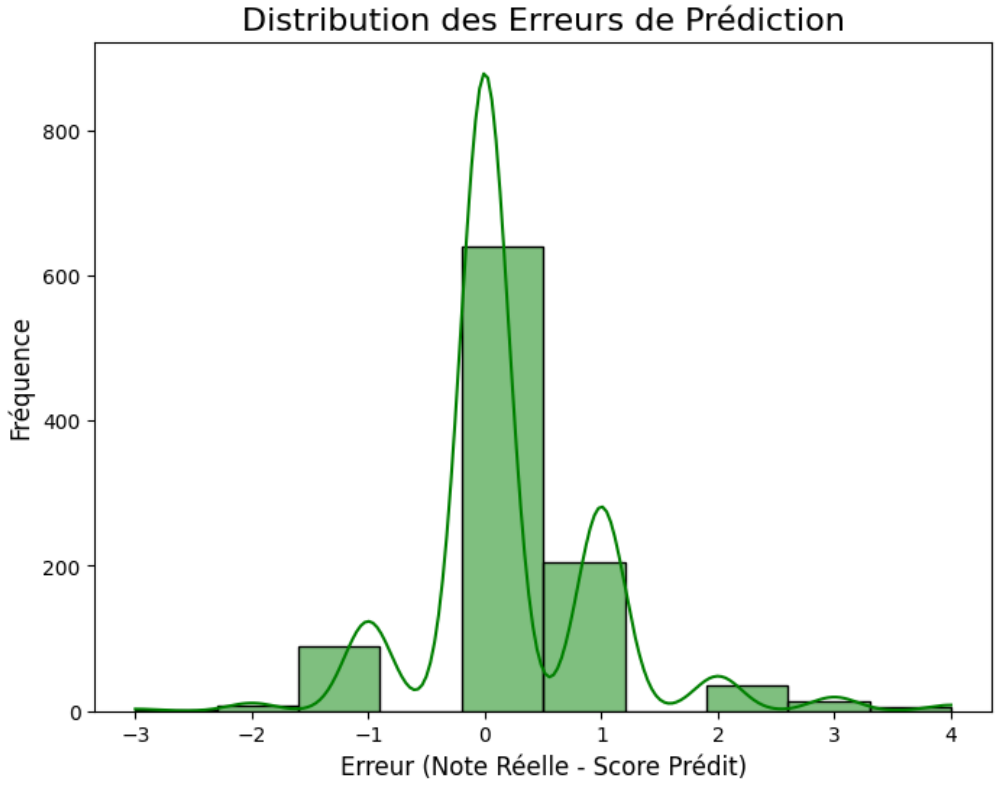
```
Accuracy : 0.64
Precision : 0.74
Recall : 0.64
F1 Score : 0.67
```

Le modèle obtient une **accuracy de 64%**, ce qui signifie qu'il fait des erreurs dans un tiers des prédictions. Sa **précision (74%)** est relativement bonne, mais le **rappel (64%)** montre qu'il manque encore certaines prédictions positives. Le **F1 score de 0.67** indique un compromis raisonnable entre la précision et le rappel, avec une marge d'amélioration possible.

Donc, Le modèle montre des performances correctes, mais nécessite des ajustements pour améliorer la précision et le rappel, notamment pour mieux identifier les prédictions positives.




Matrice de confusion des prédictions
Vérité terrain
Prédictions
1
2
3
4
5
62
30
7
2
0
8
20
10
1
0
7
17
33
13
0
2
6
36
83
37
6
12
22
144
442
400
350
300
250
200
150
100
50
0


Distribution des Erreurs de Prédiction
Fréquence
Erreur (Note Réelle - Score Prédit)
800
600
400
200
0
−3
−2
−1
0
1
2
3
4