UNIVERSITÉ DE
SHERBROOKE

Université de Sherbrooke - Département d'informatique

IFT799 – Sciences de données

Devoir 2

Réalisé par :

HAMMANI Mohamed

Mohamed Amine Mbarki

Email :

hamm1306@usherbrooke.ca

mbam5045@usherbrooke.ca

Professeur : *Shengrui Wang*

Table des matières

Table des figures

Introduction générale

Durant la période de 2020 à 2021, la crise sanitaire du COVID-19 a profondément perturbé les communautés mondiales. En réaction à cette catastrophe sanitaire inédite, plusieurs administrations ont instauré des mesures restrictives pour restreindre la propagation du virus. Ces décisions ont provoqué d'intenses réactions et ont été largement discutées sur les plateformes de médias sociaux, en particulier sur Twitter, qui est désormais une plateforme privilégiée pour exprimer les points de vue des utilisateurs en ligne.

Dans cette situation, les réseaux sociaux ont été une ressource inestimable pour examiner les perceptions générales et les sentiments des usagers. Effectivement, les remarques diffusées sur ces plateformes témoignent fréquemment de sentiments complexes comme la crainte, la colère, le bonheur ou la désolation, qui subissent une influence directe des faits sociaux et politiques.

L'objectif de ce travail est d'extraire et d'analyser ces émotions des informations Twitter recueillies au cours de cette période. En particulier, le but est d'analyser les attributs affectifs des usagers en ligne et de déceler les tendances qui découlent de leurs sentiments en utilisant des méthodes de segmentation comme le regroupement. L'objectif de cette analyse est d'approfondir la compréhension des comportements des internautes et de leur réaction aux mesures instaurées pendant la pandémie.

Problématique

Dans un contexte où les réseaux sociaux jouent un rôle central pour exprimer des émotions et opinions, notamment durant des crises comme la pandémie de COVID-19,il devient crucial d'appréhender les dynamiques émotionnelles des utilisateurs en ligne. En se basant sur des données illustrant des attributs affectifs (valence, crainte, colère, bonheur, tristesse), comment est-il possible de :

1. Repérer les comportements émotionnels répandus (positifs, neutres ou négatifs) au sein des internautes.
2. Mettre automatiquement les utilisateurs en groupes importants grâce à des techniques de regroupement comme *K-means* et le regroupement hiérarchique.
3. Analyser la pertinence des groupes repérés en comparaison avec les véritables sentiments des internautes.

Le but est donc d'évaluer si les attributs extraits correspondent aux sentiments des utilisateurs en ligne, et de mettre en parallèle les résultats des deux méthodes de regroupement dans ce contexte.

Chapitre 1

Description des données

1.1 Introduction

Dans ce chapitre, nous exposons la base de données employée pour examiner les sentiments manifestés par les utilisateurs en ligne durant la période de pandémie de COVID-19. Cinq éléments clés constituent les données, soit la valence, la peur, la colère, le bonheur et la tristesse. Ces caractéristiques reflètent les sentiments exprimés par les usagers des réseaux sociaux via leurs remarques. Chaque élément de caractéristiques est lié à un sentiment général, jugé positif, neutre ou négatif. Cela permet d'organiser les réactions des utilisateurs en ligne aux actions gouvernementales. Ici, nous examinerons en détail la structure du jeu de données, les différentes catégories de variables impliquées et les détails cruciaux pour saisir le contexte d'analyse de ce travail.

1.2 Contexte des données

Les données analysées proviennent de commentaires publiés sur Twitter, avec des caractéristiques émotionnelles extraites pour chaque internaute :

1. Caractéristiques émotionnelles extraites :

- Valence : intensité émotionnelle globale.
- Peur : intensité de la peur.
- Colère : intensité de la colère.
- Joie : intensité de la joie.
- Tristesse : intensité de la tristesse.

2. Classes associées :

— Chaque vecteur d'intensité (valence, peur, colère, joie, tristesse) est associé à un sentiment global :

- -1 : Sentiment négatif.
- 0 : Sentiment neutre.
- 1 : Sentiment positif.

1.3 Type des données

— Toutes les colonnes des intensités émotionnelles sont de type `float64`, car elles contiennent des valeurs numériques continues. Les valeurs dans ces colonnes sont comprises entre des intervalles spécifiques (entre 0 et 1).

— Les valeurs de la colonne sentiment sont 1,0 ou -1 ce qui est de type `int64`.

1.4 Interprétation des données

```
# Get summary statistics of the data
data_description = data.describe()
print(data_description)
```

✓ 0.0s

```
       valence_intensity  fear_intensity  anger_intensity  \
count       30000.000000    30000.000000     30000.000000
mean            0.477280        0.435327         0.423444
std             0.082379        0.082018         0.075165
min             0.091000        0.113000         0.142000
25%             0.432000        0.381937         0.374800
50%             0.486000        0.433500         0.414000
75%             0.527000        0.484000         0.463000
max             0.860000        0.850000         0.860000

       happiness_intensity  sadness_intensity    sentiment
count         30000.000000       30000.000000  30000.00000
mean              0.311358           0.399372      0.00000
std               0.081321           0.068110      0.81651
min               0.008000           0.145000     -1.00000
25%               0.265000           0.356383     -1.00000
50%               0.314000           0.392000      0.00000
75%               0.359000           0.433571      1.00000
max               0.771000           0.833000      1.00000
```

FIGURE 1.1 **:** Stat des données

Selon la figure , on peut constaté que :

— **valence_intensity** : L'intensité moyenne de la valence (représentant le sentiment global, qu'il soit positif ou négatif) est de 0,477, avec une différence typique de 0,082. Cela indique que la plupart des valeurs se situent autour de la moyenne, mais présentent une certaine diversité. La gamme s'étend de 0.091 à 0.860, suggérant un spectre plutôt vaste d'intensité affective, mais principalement favorable.

— **fear_intensity** : En moyenne, l'intensité de la peur est de 0.435, ce qui correspond assez bien à celle de la valence, mais présente une différence typique légèrement inférieure (0.082). Ceci révèle que les niveaux de peur sont aussi modérés et plutôt homogènes, variant entre 0.113 et 0.850.

— **anger_intensity** : L'ampleur de la colère ressemble à celle de la peur, présentant une moyenne de 0.423 et une différence typique de 0.075. Elle suit le mouvement de la peur, présentant des valeurs variant entre 0.142 et 0.860, mais un peu plus concentrées que la moyenne.

— **happiness_intensity** : Bien que la moyenne du bonheur soit de 0.311, on observe une certaine variabilité (une différence typique de 0,068). Les valeurs s'étendent de 0 à 0.771, indiquant que le bonheur se manifeste moins intensément que la peur ou la colère.

— **sadness_intensity** : La moyenne de la tristesse est de 0.399 et l'écart type est de 0,073, ce qui suggère des niveaux modérés de tristesse dans les informations. La gamme d'intensité se situe entre 0.145 et 0.833.

— **sentiment** : Le sentiment global (probablement mesuré sur une échelle de -1 à 1) est très centré autour de 0, avec une moyenne de 0 et un écart type élevé de 0.816. Cela suggère que les sentiments exprimés dans les données varient largement, allant de très négatifs à très positifs.

1.5 Conclusion

Ce chapitre a fourni des détails sur la structure et les attributs des informations employées pour examiner les sentiments des usagers en ligne durant la crise du COVID-19. Des mesures d'intensités émotionnelles comme la valence, la peur, la colère, le bonheur, la tristesse et le sentiment général révèlent une grande variété dans les sentiments des usagers. La peur, la colère et la tristesse présentent des valeurs modérées, tandis que le bonheur se révèle moins intense. Le sentiment général, qui varie entre des valeurs négatives et positives, reflète la diversité des

réactions aux incidents mondiaux. Ces conclusions mettent en lumière l'effet significatif de la pandémie sur les sentiments des individus, et cette étude offre des opportunités pour examiner l'impact des crises sur les émotions collectifs.

Chapitre 2

Analyse exploratoire des données

2.1 Introduction

L'étape de l'analyse exploratoire des données est cruciale pour appréhender la structure et les liens entre diverses variables avant de mettre en œuvre des techniques plus complexes.

Dans ce chapitre, notre attention est portée sur l'étude des répartitions des cinq attributs affectifs (valence, peur, colère, joie et tristesse) en examinant leurs rapports interdépendants.

Nous tenterons de détecter des tendances, des corrélations ou des antagonismes potentiels entre ces attributs en utilisant des visualisations en pair. Cette phase nous aidera à comprendre davantage la dynamique affective des utilisateurs en ligne et à préparer les informations pour les prochaines étapes de clustering.

2.2 Interprétation des figures.

2.2.1 Valence vs fear

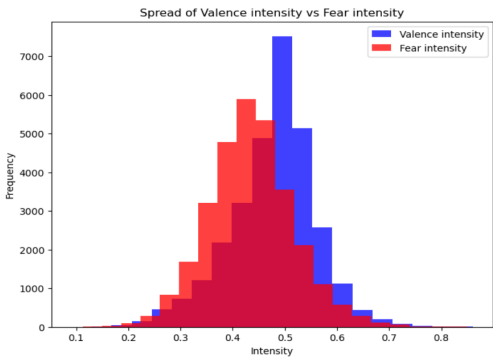

FIGURE 2.1 **:** Valence vs Fear

— **Distribution des intensités :** La valence atteint un pic autour de 0,5, tandis que la peur culmine près de 0,4, indiquant une concentration notable dans ces plages.

— **Fréquence :** L'intensité de la valence présente une fréquence légèrement plus élevée autour de 0,5 par rapport à celle de la peur.

— **Étendue des valeurs :** Les deux distributions varient principalement entre 0,2 et 0,8, avec une concentration marquée entre 0,4 et 0,6.

— **Chevauchement :** Un chevauchement important est observé, ce qui reflète des zones d'intensité communes entre la valence et la peur.

— **Symétrie :** Les deux courbes semblent relativement symétriques, suggérant une distribution proche de la normalité.

— **Conclusion générale** :Les distributions des intensités de valence et de peur semblent proches de distributions normales, avec des pics clairs et une forme symétrique. Le chevauchement observé offre des perspectives intéressantes sur l'interaction complexe des émotions humaines, qui pourrait être explorée davantage dans des analyses futures, notamment dans des contextes d'analyse des émotions sur les réseaux sociaux.

2.2.2 Valence vs Colère

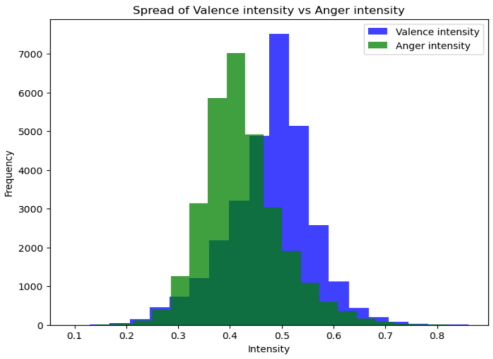

FIGURE 2.2 **:** Valence vs colère

— **Distribution des intensités :** La valence culmine autour de 0,5, tandis que la colère atteint son pic à 0,4, montrant des zones de concentration distinctes.

— **Fréquence :** La valence est plus fréquente que la colère, particulièrement pour des intensités proches de 0,5.

— **Étendue et chevauchement :** Les deux émotions couvrent une plage similaire (0,2 à 0,8) avec un chevauchement notable entre 0,4 et 0,5, suggérant des interactions possibles.

— **Conclusion :** Les distributions révèlent des différences et des similitudes dans les intensités émotionnelles, offrant des pistes pour analyser leurs interactions.

2.2.3 Valence vs Happiness

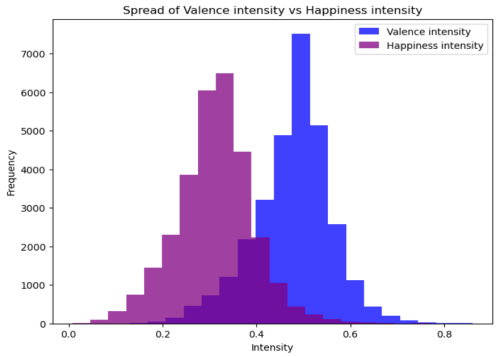

FIGURE 2.3 **:** Valence vs Happiness

— **Distribution des intensités :** La valence culmine autour de 0,5, tandis que le happiness atteint son pic à 0,4, montrant des zones de concentration différentes.

— **Fréquence :** La valence est plus fréquente que le happiness autour de 0,5, indiquant une prédominance des sentiments neutres ou positifs.

— **Étendue et chevauchement :** Les deux émotions s'étendent de 0,2 à 0,8, avec un chevauchement notable entre 0,4 et 0,6, suggérant une interaction ou une co-occurrence possible.

— **Symétrie :** Les distributions sont relativement symétriques autour de leurs pics, évoquant une répartition équilibrée des intensités.

— **Conclusion :** Les deux distributions mettent en évidence une interaction intéressante entre valence et happiness. Le chevauchement dans certaines plages d'intensité pourrait indiquer une relation complémentaire ou une co-occurrence de ces émotions dans des contextes spécifiques

2.2.4 Valence vs Sadness

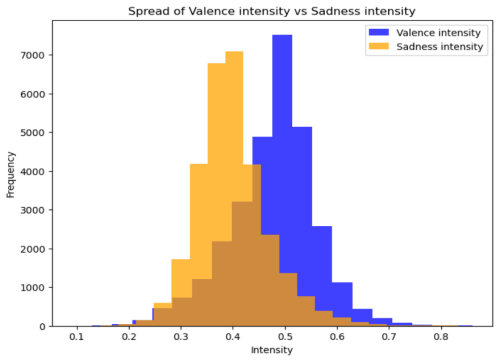

FIGURE 2.4 **:** Valence vs Sadness

— **Distribution des intensités :** La valence (en bleu) culmine autour de 0,5, tandis que la sadness (en orange) atteint son pic à 0,4, reflétant des états émotionnels distincts.

— **Fréquence :** La valence a une fréquence maximale plus élevée, indiquant une prédominance des émotions neutres ou positives.

— **Étendue des valeurs :** Les deux distributions s'étendent principalement entre 0,2 et 0,8, avec une concentration notable entre 0,4 et 0,6.

— **Chevauchement :** Un chevauchement est visible entre 0,4 et 0,5, ce qui pourrait signaler des émotions partagées entre valence et tristesse.

— **Symétrie :** Les deux distributions semblent symétriques, bien que la tristesse montre une légère asymétrie vers des valeurs plus faibles.

— **Conclusion générale :** Les distributions montrent des schémas distincts et des chevauchements, reflétant des nuances émotionnelles et des interactions possibles entre valence et sadness.

2.2.5 Fear vs Anger

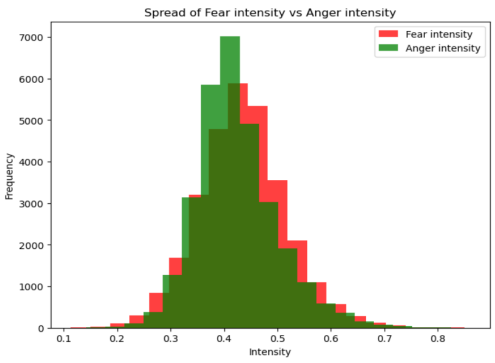

FIGURE 2.5 **:** Fear vs Anger

— **Distribution des intensités :** La peur (rouge) est concentrée entre 0,3 et 0,5, tandis que la colère (verte) couvre une plage plus large, mais reste centrée autour de 0,4.

— **Fréquence :** La peur a une fréquence plus élevée à des intensités faibles, alors que la colère domine aux intensités modérées à élevées.

— **Étendue des valeurs :** Les deux émotions partagent une plage d'intensité similaire, avec une concentration marquée de la peur entre 0,3 et 0,5.

— **Chevauchement :** Un chevauchement significatif apparaît entre 0,3 et 0,5, indiquant une possible coexistence des émotions à ces intensités.

— **Symétrie :** Les distributions sont globalement symétriques, bien que la peur soit plus concentrée, et que la colère montre une augmentation progressive pour les faibles intensités.

— **Conclusion générale :** Les histogrammes révèlent des différences marquées entre la peur et la colère, tout en montrant des chevauchements dans les intensités modérées. La peur est plus fréquente à faible intensité, tandis que la colère se manifeste davantage à des intensités élevées.

2.2.6 Fear vs Happiness

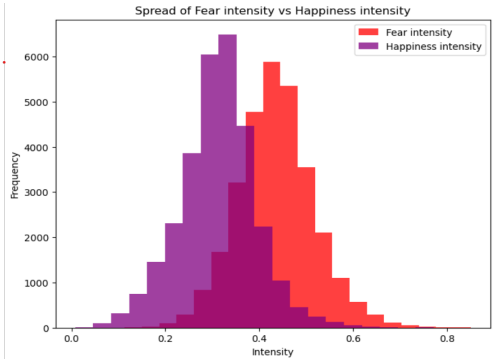

FIGURE 2.6 **:** Fear vs Happiness

— **Distribution des intensités :** La peur (rouge) est concentrée entre 0,3 et 0,5, tandis que le bonheur (violet) est plus marqué autour de 0,4.

— **Fréquence :** Le bonheur atteint une fréquence maximale plus élevée que la peur, indiquant une prévalence plus importante des émotions positives dans cet intervalle.

— **Étendue des valeurs :** Les deux émotions varient principalement entre 0,2 et 0,6, montrant une concentration notable dans cette plage d'intensité.

— **Chevauchement :** Un chevauchement significatif est visible entre 0,3 et 0,5, illustrant une coexistence possible des émotions dans cette plage.

— **Symétrie :** Les deux distributions semblent relativement symétriques, avec une légère dissymétrie pour le bonheur vers les valeurs plus élevées.

— **Conclusion générale :** Les histogrammes montrent des dynamiques opposées, avec une peur plus concentrée à des intensités faibles et un bonheur dominant à des intensités modérées. Le chevauchement des deux émotions mérite une analyse plus approfondie pour explorer leur interdépendance.

2.2.7 Fear vs Sadness

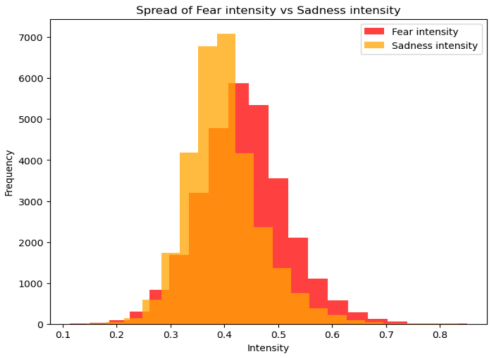

FIGURE 2.7 **:** Fear vs Sadness

— **Distribution des intensités :** Fear (rouge) est concentrée autour de 0,4, tandis que Sadness (orange) s'étend légèrement plus bas, avec un pic autour de 0,35.

— **Fréquence :** Sadness atteint une fréquence maximale plus élevée que Fear, indiquant que cette émotion est plus dominante dans les données.

— **Étendue des valeurs :** Les deux émotions se répartissent principalement entre 0,2 et 0,6, avec une plus grande dispersion pour Fear.

— **Chevauchement :** Les émotions montrent un chevauchement significatif dans la plage de 0,3 à 0,5, suggérant qu'elles peuvent coexister à des intensités modérées.

— **Symétrie :** Les distributions sont globalement symétriques, mais Sadness montre une légère asymétrie vers les valeurs plus faibles.

— **Conclusion générale :** Ces histogrammes révèlent une proximité dans les intensités de Fear et de Sadness, avec des différences notables dans la fréquence maximale et la concentration des valeurs. Cela indique des états émotionnels proches, mais avec des caractéristiques distinctes.

2.2.8 Anger vs Happiness

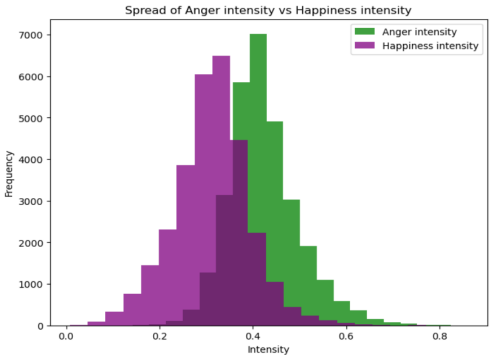

FIGURE 2.8 **:** Anger vs Happiness

— **Distribution des intensités :** Anger (vert) est centrée autour de 0,4, tandis que Happiness (violet) est concentrée autour de 0,3.

— **Fréquence :** Happiness atteint une fréquence maximale plus élevée, suggérant qu'elle est plus présente dans les données.

— **Étendue des valeurs :** Anger montre une répartition légèrement plus large, allant de 0,2 à 0,8, alors que Happiness est plus restreinte.

— **Chevauchement :** Les deux émotions partagent une région significative entre 0,3 et 0,5.

— **Symétrie :** Les deux distributions semblent relativement symétriques, avec une concentration visible vers les intensités modérées.

— **Conclusion générale :** Les émotions Anger et Happiness se distinguent par leur pic d'intensité et leur fréquence, tout en partageant des zones communes qui indiquent une relation potentielle dans les états émotionnels.

2.2.9 Anger vs Sadness

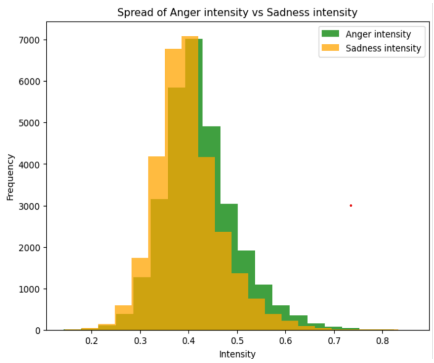

FIGURE 2.9 **:** Anger vs Sadness

— **Distribution des intensités :** Anger (vert) est centrée autour de 0,4, tandis que Sadness (jaune) est également concentrée autour de 0,4, avec des profils très similaires.

— **Fréquence :** Les deux émotions atteignent une fréquence maximale comparable, mais Sadness semble légèrement plus fréquente dans certaines plages d'intensité.

— **Étendue des valeurs :** Anger et Sadness ont une répartition étendue de 0,2 à 0,8, avec une forte concentration autour de leur moyenne.

— **Chevauchement :** Une large région de chevauchement est observée, particulièrement entre 0,3 et 0,5, indiquant des similitudes dans les données des deux émotions.

— **Symétrie :** Les deux distributions apparaissent symétriques, avec une concentration visible autour des valeurs centrales.

— **Conclusion générale :** Les émotions Anger et Sadness présentent des distributions presque superposées, suggérant une coexistence fréquente ou des intensités émotionnelles partagées dans les données.

2.2.10 Happiness vs Sadness

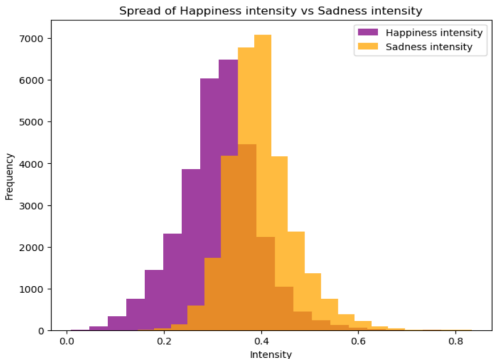

FIGURE 2.10 **:** Happiness vs Sadness

— **Distribution des intensités :** Happiness (violet) et Sadness (orange) sont toutes deux centrées autour de 0,4, montrant des tendances centrales similaires dans leurs intensités.

— **Fréquence :** Sadness atteint une fréquence maximale légèrement supérieure à celle de Happiness, en particulier dans les régions de chevauchement.

— **Étendue des valeurs :** Les intensités de Happiness et Sadness s'étendent approximativement de 0,2 à 0,8, avec un pic bien défini dans les plages intermédiaires.

— **Chevauchement :** Un chevauchement significatif est observé entre 0,3 et 0,5, indiquant des valeurs d'intensité partagées entre les deux émotions.

— **Symétrie :** Les distributions apparaissent symétriques, avec une forte concentration autour de leurs valeurs centrales.

— **Conclusion générale :** Les distributions de Happiness et Sadness sont fortement alignées, suggérant des similitudes dans les intensités émotionnelles et une co-occurrence possible dans les données.

2.3 Conclusion

Ce chapitre a permis d'explorer les interactions entre différents attributs émotionnels (valence, peur, colère, joie et tristesse) au travers d'analyses visuelles et descriptives. Les distributions des émotions étudiées montrent des patterns spécifiques tout en mettant en évidence des chevauchements notables entre certaines d'entre elles.

Les analyses ont révélé que :

— La valence, souvent associée à des émotions neutres ou positives, présente une distribution concentrée et des interactions modérées avec d'autres émotions, comme la peur ou la joie.

— Les émotions négatives (peur, colère, tristesse) présentent des chevauchements significatifs, indiquant des zones d'intensité où ces émotions coexistent, soulignant la complexité des dynamiques émotionnelles humaines.

— Les émotions opposées, comme la joie et la tristesse ou la peur et le bonheur, montrent des patterns distincts, mais des plages d'intensité partagée, ouvrant la voie à une analyse plus approfondie de leur interdépendance.

Ces observations fournissent une base solide pour les prochaines étapes d'analyse, notamment dans le contexte du clustering, où ces dynamiques émotionnelles complexes pourront être exploitées pour segmenter les données en groupes cohérents. Cette compréhension approfondie des interactions émotionnelles enrichit également la préparation des données pour des modèles plus avancés, contribuant ainsi à la précision des futures analyses.

Chapitre 3

K-Means

3.1 Introduction

K-means est un algorithme de classification non supervisée utilisé pour regrouper des données en un nombre prédéfini de clusters. Il fonctionne en trois étapes principales :

— **Initialisation des centroids** : Choix aléatoire de k centroids, représentant les centres des clusters.

— **Attribution des points aux clusters** : Chaque point de données est affecté au centroid le plus proche.

— **Mise à jour des centroids** : Les centroids sont recalculés en prenant la moyenne des points assignés à chaque cluster.

Le processus est répété jusqu'à ce que les centroids ne changent plus de manière significative. L'objectif de l'algorithme est de minimiser la variance intra-cluster et de maximiser la variance inter-cluster.

3.2 Appliquer K-Means

```python
#seperate the features (FTS) and the target (T)

FTS = data.drop('sentiment', axis = 1)
T = data['sentiment']



# we create a dictionary to store all the dataframes related to each k
from sklearn.cluster import KMeans

kmeans_dfs = {}
for k in range(2,11):
    k_means_model = KMeans(n_clusters = k, random_state = 42)
    labels = k_means_model.fit_predict(FTS)
    df = FTS.copy()
    df['Cluster'] = labels
    kmeans_dfs[k] = df
```

FIGURE 3.1 **:** Algorithme K-Means

Cette analyse sépare d'abord les caractéristiques (features) de la cible (sentiment). Ensuite, l'algorithme de K-Means est appliqué pour diviser les données en clusters. Pour chaque nombre de clusters k allant de 2 à 10, un modèle K-Means est utilisé pour assigner chaque observation à un cluster. Les résultats sont stockés dans des DataFrames contenant les caractéristiques originales et une colonne supplémentaire indiquant le numéro du cluster. Cela permet d'explorer comment les données se regroupent pour différentes valeurs de k et de mieux comprendre les structures sous-jacentes dans les données sentimentales.

Les deux figures ci-dessous illustrent quelques résultats obtenus à partir de l'application de l'algorithme de K-Means sur les données.

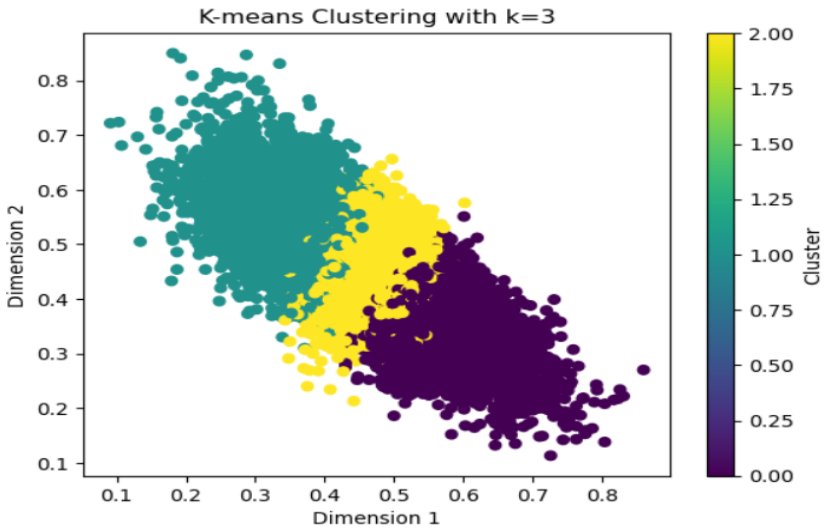

FIGURE 3.2 **:** Algorithme K-Means pour k=3

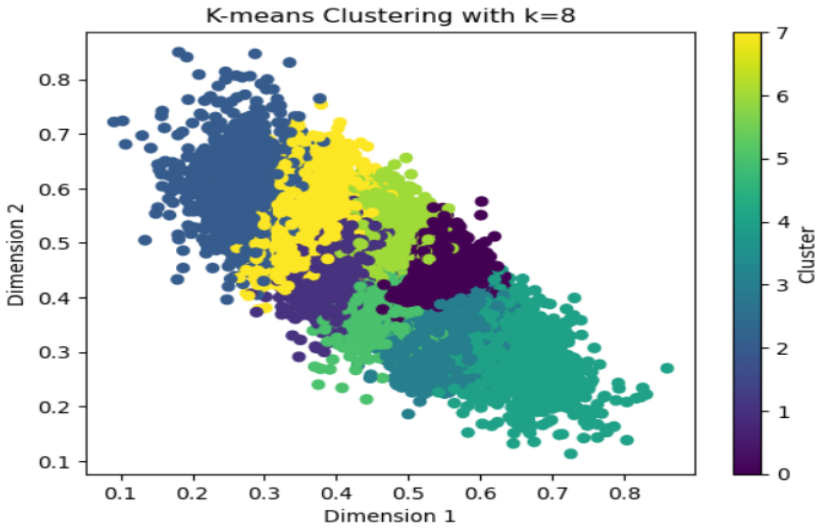

FIGURE 3.3 **:** Algorithme K-Means pour k=8

La première visualisation est normale, mais si nous appliquions UMAP pour un meilleur regroupement, nous obtiendrions une meilleure séparation.

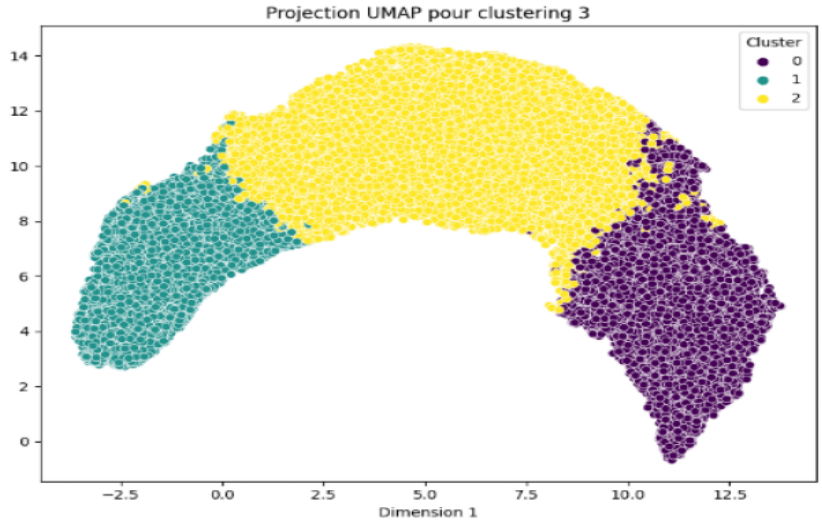

FIGURE 3.4 **:** Algorithme K-Means pour k=3 avec umap

3.3 Silhouette score

Le Silhouette Score est une mesure utilisée pour évaluer la qualité d'un clustering en analysant la cohésion à l'intérieur des clusters et la séparation entre les clusters. Pour chaque point, il compare la distance moyenne aux autres points de son propre cluster (cohésion) à la distance moyenne aux points du cluster le plus proche (séparation). Ce score varie entre -1 et 1 : une

valeur proche de 1 indique que le point est bien assigné à son cluster, une valeur proche de 0 suggère qu'il est à la frontière entre deux clusters, et une valeur négative signifie qu'il est probablement mal assigné. En calculant la moyenne des scores silhouette pour tous les points, on obtient une mesure globale de la qualité du clustering.

Nombre de clusters (K)	Silhouette Score
2	0.4309305032755166
3	0.3423241940746704
4	0.29654022320551854
5	0.26112771660104644
6	0.21515321858961106
7	0.20466052696455905
8	0.19981659470810284
9	0.19586951923840848
10	0.19479334040870913

TABLEAU 3.1 **:** Scores silhouette pour différents nombres de clusters (K).

3.3.1 Analyse des scores pour différents K

— $K = 2$:

Le score est de 0.4309, le plus élevé parmi tous les K. Cela suggère que la meilleure séparation des données en clusters se produit avec 2 groupes. Cela peut refléter une structure binaire ou une division nette dans les données.

— $K = 3$:

Le score diminue à 0.3423, montrant une dégradation de la qualité du clustering. Cela peut indiquer que les données ne sont pas parfaitement adaptées à une division en 3 clusters.

— $K = 4$ **et** $K = 5$:

Les scores continuent de baisser (0.2965 et 0.2611, respectivement), ce qui suggère une moins bonne cohérence interne et séparation entre les clusters.

— $K = 6$ **à** $K = 10$:

Les scores deviennent très bas (entre 0.21 et 0.19), ce qui montre que l'augmentation du nombre de clusters conduit à une fragmentation excessive des groupes, réduisant ainsi la qualité du clustering.

Recommandation sur le choix de K :

Le meilleur K semble être $K = 2$, car il maximise le Silhouette Score.
Bien que $K = 3$ ait un score inférieur, il pourrait être envisagé si l'objectif est de diviser les données en davantage de groupes malgré une qualité légèrement moindre.

Donc pour mieux choisir on passe au calcul de précision , recall et f1 score .

3.4 Précision , recall et f1 score

3.4.1 Précision (Precision)

La précision mesure la proportion des prédictions positives correctes parmi toutes les prédictions positives. Formellement, cela peut être exprimé comme suit :

$$\text{Précision} = \frac{\text{Vrai Positifs (VP)}}{\text{Vrai Positifs (VP)} + \text{Faux Positifs (FP)}}$$

Où :

— **Vrai Positifs (VP)** : Les instances correctement classées comme positives.

— **Faux Positifs (FP)** : Les instances incorrectement classées comme positives.

3.4.2 Rappel (Recall)

Le rappel mesure la proportion des véritables positifs qui ont été correctement identifiés par le modèle. Cela peut être défini par :

$$\text{Rappel} = \frac{\text{Vrai Positifs (VP)}}{\text{Vrai Positifs (VP)} + \text{Faux Négatifs (FN)}}$$

Où :

— **Faux Négatifs (FN)** : Les instances qui sont réellement positives mais ont été classées comme négatives.

3.4.3 3. Score F1 (F1 Score)

Le score F1 est la moyenne harmonique de la précision et du rappel. Il permet d'obtenir un équilibre entre ces deux métriques :

$$\text{F1 Score} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

Ce score est particulièrement utile lorsqu'il y a un déséquilibre entre les classes, car il combine les avantages de la précision et du rappel en un seul nombre.

Summary of Clustering Metrics:

	K	Precision	Recall	F1 Score
0	2	0.432991	0.568567	0.474072
1	3	0.755920	0.710167	0.705398
2	4	0.026051	0.034375	0.029415
3	5	0.141041	0.076840	0.087923
4	6	0.190697	0.110633	0.127960
5	7	0.016981	0.009586	0.012254
6	8	0.147288	0.068163	0.093092
7	9	0.075854	0.022567	0.033220
8	10	0.083489	0.028560	0.040599

FIGURE 3.5 **:** Tableau des scores

3.4.4 Interprétation des résultats des métriques de clustering

Les résultats des métriques de Précision, Rappel et F1 Score pour différents nombres de clusters (K) sont analysés ci-dessous.

3.4.4.1 K=2 (2 clusters)

— **Précision** : 0.43

— **Rappel** : 0.57

— **Score F1** : 0.47

Pour $K = 2$, les performances sont modérées. La précision est de 0.43, ce qui indique que le modèle a un nombre raisonnable de faux positifs. Le rappel est plus élevé (0.57), ce qui suggère que le modèle est assez bon pour identifier les vrais positifs. Le score F1 est modéré (0.47), indiquant un compromis raisonnable entre précision et rappel.

3.4.4.2 K=3 (3 clusters)

— **Précision** : 0.76

— **Rappel** : 0.71

— **Score F1** : 0.71

Pour $K = 3$, les performances s'améliorent significativement. La précision est élevée (0.76), montrant que la majorité des éléments classés comme positifs sont effectivement corrects. Le rappel est également élevé (0.71), ce qui signifie que la plupart des vrais positifs ont été identifiés. Le score F1 (0.71) indique un bon équilibre entre précision et rappel.

3.4.4.3 K=4 (4 clusters)

— **Précision** : 0.03

— **Rappel** : 0.03

— **Score F1** : 0.03

Les performances pour $K = 4$ sont très faibles. La précision, le rappel et le score F1 sont tous proches de zéro, ce qui suggère que le modèle n'a pas bien séparé les données en 4 clusters. Il pourrait y avoir un problème de sur-ajustement (overfitting) ou un mauvais choix du nombre de clusters.

3.4.4.4 K=5 (5 clusters)

— **Précision** : 0.14

— **Rappel** : 0.08

— **Score F1** : 0.09

Pour $K = 5$, les performances restent médiocres, bien que légèrement meilleures que pour $K = 4$. La précision est faible, ce qui signifie que de nombreux éléments classés comme positifs sont incorrects. Le rappel est également faible, ce qui suggère que le modèle a du mal à capturer les vrais positifs. Le score F1 reste faible (0.09).

3.4.4.5 K=6 (6 clusters)

— **Précision** : 0.19

— **Rappel** : 0.11

— **Score F1** : 0.13

Pour $K = 6$, les résultats restent relativement faibles. La précision et le rappel sont légèrement améliorés par rapport à $K = 5$, mais restent insuffisants pour un bon clustering. Le score F1 est également faible (0.13).

3.4.4.6 K=7 à K=10 (7 à 10 clusters)

— **Précision** : Très faible pour tous les K de 7 à 10.

— **Rappel** : Très faible pour tous les K de 7 à 10.

— **Score F1** : Très faible pour tous les K de 7 à 10.

Les performances pour $K = 7$ à $K = 10$ sont extrêmement faibles. Les valeurs de précision, rappel et F1 sont proches de zéro, ce qui indique que le modèle ne parvient pas à trouver des clusters significatifs pour ces valeurs de K. Cela peut être dû à un sur-ajustement ou à un mauvais choix de K.

3.5 Conclusion

En analysant les résultats, $K = 3$ semble être le nombre optimal de clusters. Les scores de précision, rappel et F1 sont nettement meilleurs pour $K = 3$, indiquant un bon compromis entre les deux métriques. Les autres valeurs de K, en particulier pour $K > 3$, montrent des performances médiocres, suggérant que l'augmentation du nombre de clusters perturbe la capacité du modèle à identifier des structures significatives dans les données.

Chapitre 4

Regroupement hiérarchique

4.1 Introduction

Le **regroupement hiérarchique** est une méthode de clustering qui crée une hiérarchie de clusters en fonction de la similarité des éléments. Il existe deux approches principales :

— **Agglomératif** : Commence avec chaque point comme un cluster et fusionne progressivement les plus proches.

— **Divisif** : Commence avec un seul cluster et divise progressivement les groupes.

L'algorithme utilise une matrice de distances pour mesurer la similarité entre les points. Un *dendrogramme* est souvent généré pour visualiser la fusion ou la division des clusters.

4.2 Clustering hiérarchique

4.2.1 Première approche

La première approche consiste a générer des coupes de manière dynamique et contrôlée. Voici comment la coupe a été générée :

1. **Calcul des distances de fusion** : On effectue un clustering hiérarchique agglomératif avec la fonction **linkage**. Cette fonction produit un arbre hiérarchique, où chaque ligne du résultat contient les indices des deux clusters fusionnés et la distance à laquelle cette fusion a eu lieu. Ces distances sont stockées dans la variable **merge_distances**.
2. **Génération des seuils (coupe)** : On génère plusieurs valeurs de seuil (**thresholds**) réparties uniformément entre la distance minimale et maximale des fusions. Cela est effectué

avec la fonction **np.linspace()** qui génère une série de valeurs comprises entre la distance minimale (**merge_distances.min()**) et la distance maximale (**merge_distances.max()**), avec un nombre spécifique de seuils (6 dans ce cas).

3. **Application des seuils** : On applique chaque seuil à l'algorithme de clustering hiérarchique avec **fcluster(hierarchical_clustering, t=threshold, criterion='distance')**, ce qui découpe l'arbre à cette hauteur et détermine les clusters à chaque seuil spécifié.

Quelques résultats des dendrogrammes :

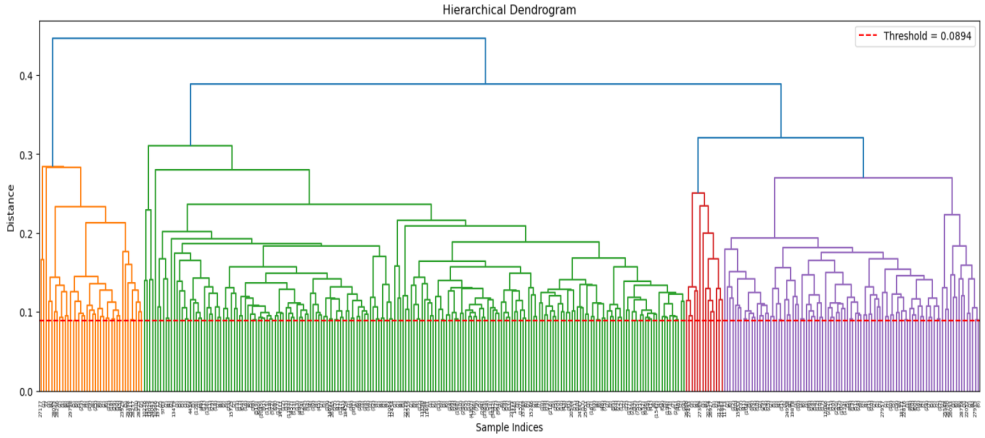

FIGURE 4.1 **:** Dendrogramme pour k=29935

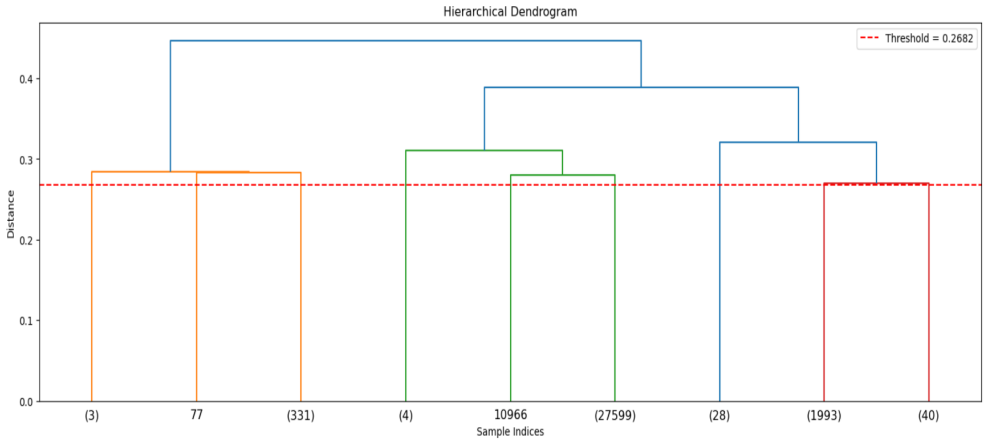

FIGURE 4.2 **:** Dendrogramme pour k=9

Le tableau ci-dessous montre les résultats de la première approche

	Threshold	K	Silhouette Score
0	0.000000	29935	0.003967
1	0.089411	345	0.006867
2	0.178821	28	0.177131
3	0.268232	9	0.207620
4	0.357642	3	0.399158
5	0.447053	1	-1.000000

FIGURE 4.3 **:** Tableau des résultats

4.2.2 Deuxième approche

La deuxième approche consiste à générer des coupes de manière intuitive en spécifiant explicitement une liste de valeurs de seuil, car la première approche a donné un score de silhouette faible.

Quelques résultats des dendrogrammes :

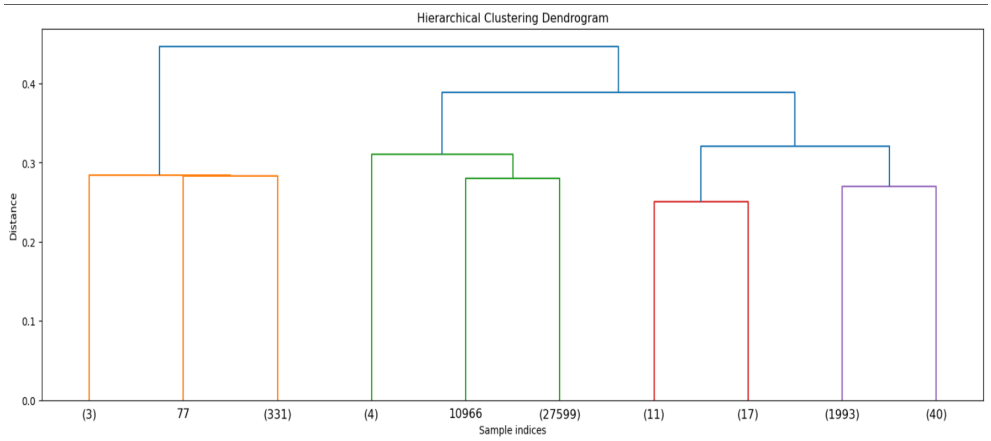

FIGURE 4.4 **:** Dendrogramme pour k=10

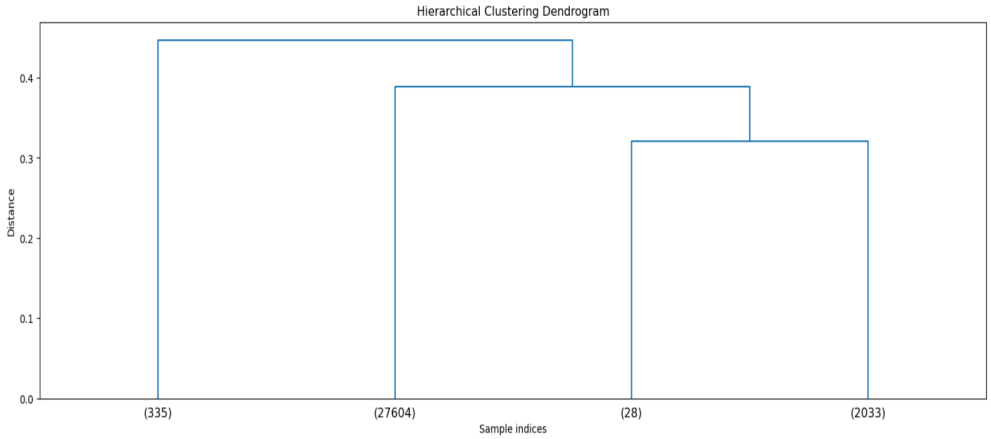

FIGURE 4.5 **:** Dendrogramme pour k=4

Le tableau ci-dessous montre les résultats de la deuxième approche

	Threshold	K	Silhouette Score
0	0.25	10	0.205314
1	0.28	7	0.250427
2	0.30	5	0.322223
3	0.32	4	0.387249
4	0.35	3	0.399158
5	0.40	2	0.502137

FIGURE 4.6 **:** Tableau des résultats

4.3 Conclusion

Dans ce chapitre, nous avons exploré deux approches du **regroupement hiérarchique** pour le clustering de données.

La première approche, qui utilise des seuils générés dynamiquement à partir des distances de fusion, a permis de créer une hiérarchie claire des clusters. Cependant, les résultats obtenus n'ont pas toujours montré une qualité de clustering optimale, notamment en termes de score de silhouette. Cette méthode a permis de visualiser les dendrogrammes à différentes étapes, offrant ainsi une vue d'ensemble de la structure hiérarchique des données.

La deuxième approche, plus intuitive, consistait à définir explicitement des seuils fixes pour la coupe. Cette approche a permis d'obtenir de meilleurs résultats en ajustant les seuils de manière plus ciblée, bien que le score de silhouette reste un critère important à prendre en compte pour évaluer la qualité du clustering.

En conclusion, bien que le regroupement hiérarchique offre une méthode flexible et visuellement intéressante pour le clustering, il est essentiel de choisir les seuils avec soin pour optimiser la qualité des clusters formés. Une évaluation supplémentaire à l'aide de métriques de performance, telles que le score de silhouette, peut aider à affiner les résultats et à choisir la méthode la plus adaptée aux données.

Chapitre 5

K-Means vs clustering hierarchique

5.1 Introduction

Dans ce chapitre, on va faire une comparaison entre K-Means et le clustering hiérarchique.

5.2 Résultats

Méthode	Précision	Rappel	Score F1
Clustering Hiérarchique	0.3358442094766404	0.33573333333333333	0.264638704851612
K-means	0.755919530010392	0.7101666666666667	0.7053979907726137

TABLEAU 5.1 **:** Scores de précision, rappel et F1 pour le clustering hiérarchique et K-means.

Interprétation des résultats :

Les scores de précision, rappel et F1 pour les méthodes de clustering hiérarchique et K-means peuvent être interprétés comme suit :

5.2.1 Clustering Hiérarchique

— **Précision : 0.336**

Cela signifie que, parmi tous les objets classifiés dans un cluster donné, environ 33.6% étaient effectivement des éléments appartenant à ce cluster. La précision relativement faible suggère que le clustering hiérarchique a parfois tendance à inclure des objets erronés dans les clusters.

— **Rappel : 0.336**

Le rappel indique qu'environ 33.6% des objets réels dans chaque cluster ont été correc-

tement identifiés par le modèle. Le faible rappel montre que de nombreux éléments n'ont pas été inclus dans les clusters auxquels ils appartiennent réellement.

— **Score F1 : 0.265**

Le score F1, qui est la moyenne harmonique de la précision et du rappel, est relativement bas (0.265). Cela reflète un compromis faible entre la capacité du modèle à identifier correctement les objets et à éviter les faux positifs. Un score F1 faible indique que le clustering hiérarchique n'a pas été très efficace pour distinguer les clusters de manière précise et complète.

5.2.2 K-means

— **Précision : 0.756**

La précision pour K-means est bien plus élevée (0.756), ce qui signifie que la majorité des objets classifiés dans un cluster donné appartiennent réellement à ce cluster. K-means semble donc être plus efficace pour regrouper des objets similaires.

— **Rappel : 0.710**

Le rappel pour K-means est également plus élevé (0.710), indiquant que la majorité des objets appartenant à un cluster ont été correctement classés dans le bon groupe. Cela suggère que K-means est plus performant pour capturer les éléments de chaque cluster.

— **Score F1 : 0.705**

Le score F1 de K-means (0.705) est bien plus élevé que celui du clustering hiérarchique, ce qui indique une meilleure performance globale. K-means parvient à équilibrer la précision et le rappel, ce qui se traduit par un meilleur modèle global pour le clustering.

5.3 Conclusion

K-means semble offrir de meilleures performances globales que le clustering hiérarchique, avec des scores de précision, rappel et F1 plus élevés. Le clustering hiérarchique, bien que capable de produire des clusters, semble moins précis dans son classement des objets et plus enclin à des erreurs de classification.