Projet P2
Clustering

Université du Québec en Outaouais
Titre du cours : INF1473-Entreposage et prospection des données

Nom et Prénom : Sankara Kabem Abdoul Charif
Kouyate Yasmine Jawad
C

Introction

Ce projet vise à approfondir la compréhension et la maîtrise des concepts de clustering. À partir d'un jeu de données tiré d'analyses menées sur des commentaires d'internautes canadiens pendant la pandémie de COVID-19, le travail consiste à explorer les relations entre différentes caractéristiques émotionnelles et à segmenter ces données en groupes significatifs. L'objectif est de mettre en œuvre des approches variées de clustering, telles que K-means et le clustering hiérarchique, tout en évaluant leurs performances avec des critères quantitatifs et des métriques comme le score silhouette et le F1-score.

I Analyse des caractéristiques émotionnelles

Dans un premier temps, analysons conjointement les distributions des cinq caractéristiques émotionnelles (*valence*, *peur*, *colère*, *joie*, *tristesse*) afin de comprendre les relations entre elles. Pour ce faire :

- Nous avons créé une représentation graphique pour chaque paire de caractéristiques (un total de 10 figures est attendu).
- Pour chaque figure, nous avons interprété les relations observées entre les deux caractéristiques et leur possible lien avec les sentiments exprimés (*négatif*, *neutre* ou *positif*).

I.1 les distributions de l'intensité de la colère (anger_intensity) et de l'intensité de la joie (happiness_intensity)

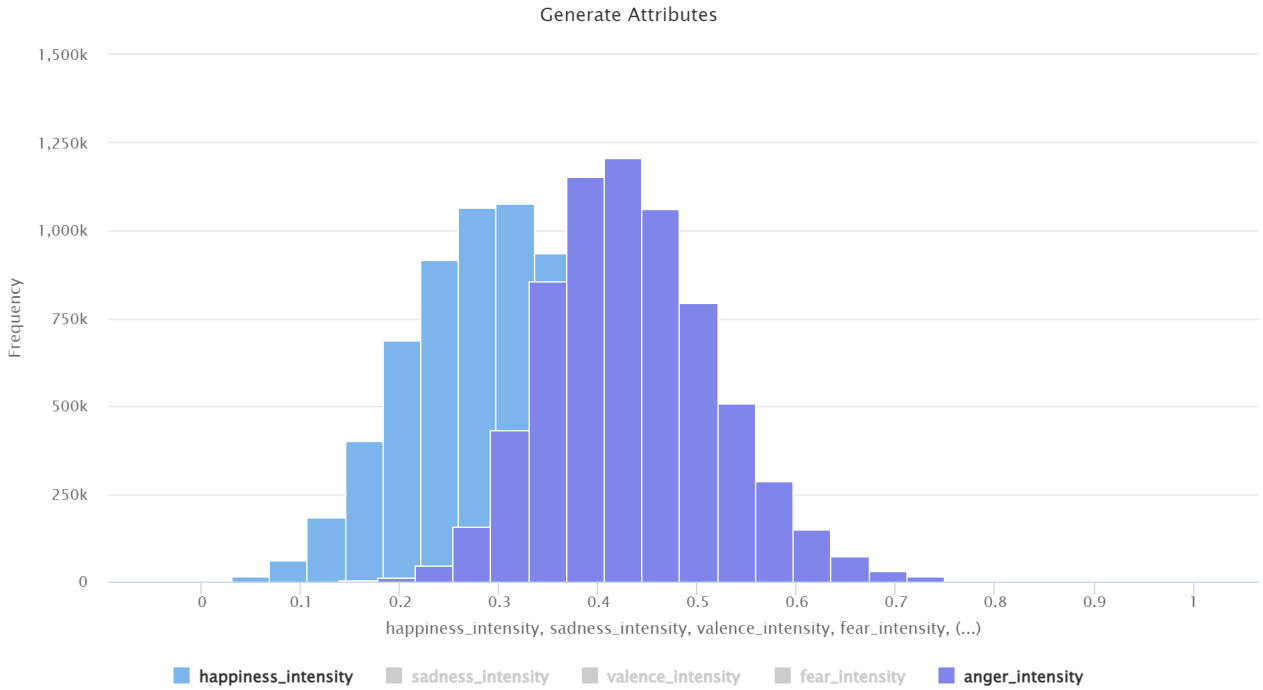


FIGURE 1 – les distributions de l'intensité de la colère (anger_intensity) et de l'intensité de la joie (happiness_intensity)

Interprétation

— **Superposition des distributions :** On observe que les distributions de l'intensité de la joie (*barres bleu clair*) et de la colère (*barres bleu foncé*) sont assez distinctes, avec peu de superposition. L'intensité de la joie est plus élevée dans les zones de faibles valeurs (entre 0.1 et 0.3), tandis que la colère présente une distribution plus concentrée autour de valeurs modérées (0.3 à 0.5).

— **Caractère antagoniste :** La séparation entre ces deux distributions suggère un caractère antagoniste entre la joie et la colère. Cela signifie que les internautes ressentent rarement une intensité élevée de ces deux émotions en même temps. Ce résultat est cohérent vu qu'il est peu probable qu'une personne exprime simultanément des sentiments de bonheur et de colère.

— **Détection de sentiments :** Cette distinction claire entre les distributions nous permet d'identifier une intensité élevée de joie comme indicateur de sentiment positif {1}, tandis qu'une intensité élevée de colère comme un sentiment négatif {-1}.

I.2 Les distributions de l'intensité de la colère (anger_intensity) et de l'intensité de la tristesse (sadness_intensity)

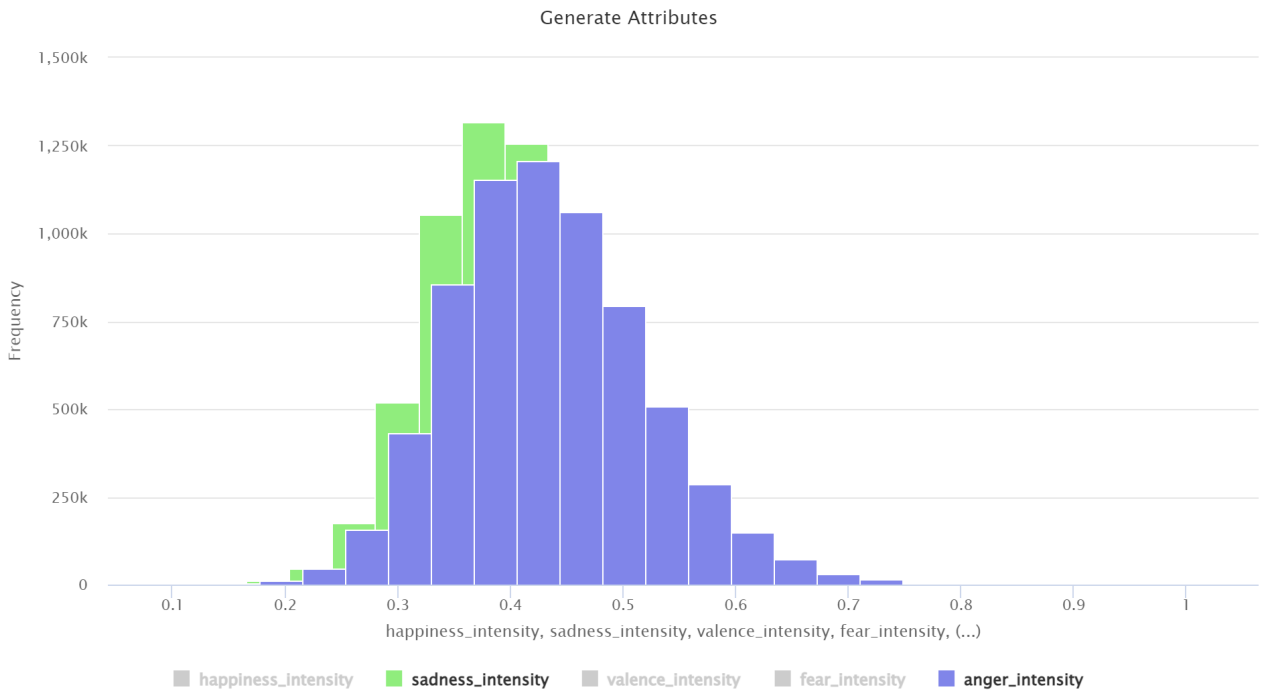


FIGURE 2 – Les distributions de l'intensité de la colère (anger_intensity) et de l'intensité de la tristesse (sadness_intensity)

Interprétation

— **Superposition des distributions :** La distribution de l'intensité de la tristesse (*barres vertes*) est principalement concentrée autour de valeurs entre 0.2 et 0.4. Elle semble avoir une forme plus étroite que celle de l'intensité de la colère (*barres bleu foncé*), qui s'étend d'avantage jusqu'à 0.7. Il y a **une superposition partielle entre les deux distributions** dans la région autour de **0.3 à 0.4**. Cela montre que, dans

certains cas, les individus peuvent ressentir de la tristesse et de la colère, bien que la tristesse ait tendance à être moins intense dans cette distribution.

— **Détection de sentiments :** Une forte intensité de colère, sans tristesse, pourrait signaler une colère isolée donc un sentiment négatif {-1}, tandis qu'une intensité modérée dans les deux émotions pourrait indiquer un sentiment mélangé donc avoir la valeur {0}.

I.3 Les distributions de l'intensité de la peur (fear_intensity) et de la joie (happiness_intensity)

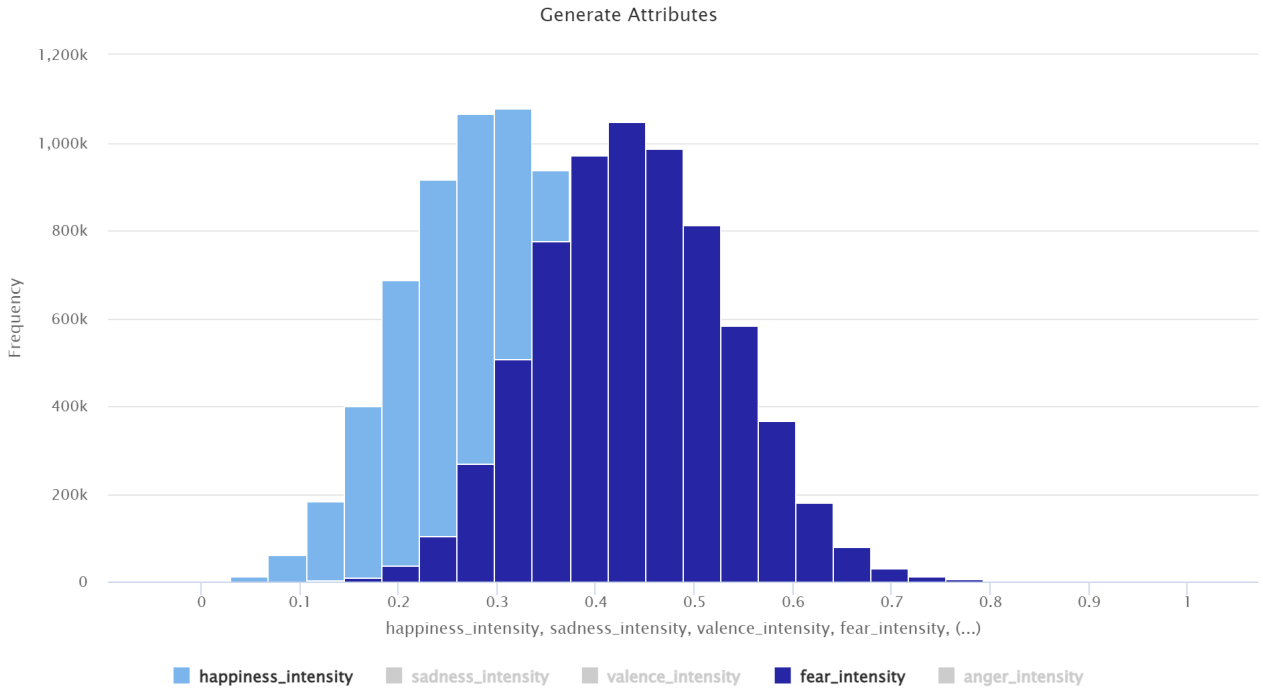


FIGURE 3 – Les distributions de l'intensité de la peur (fear_intensity) et de la joie (happiness_intensity)

Interprétation

— **Superposition des distributions :** Les distributions de la joie et de la peur semblent très distinctes. La joie est concentrée autour de valeurs basses (entre 0.1 et 0.3), tandis que la peur est plus centrée sur des valeurs légèrement plus élevées (autour de 0.4 à 0.5). Très peu de chevauchement observable, suggérant que ces deux émotions sont rarement présentes avec des intensités similaires. les distributions de la joie et de la peur montre que ces deux émotions sont très opposées. Cela reflète bien la nature antagoniste des émotions positives et négatives : la joie étant une émotion positive, et la peur, une émotion négative.

— **Détection de sentiments :** Joie et peur sont clairement distinctes, renforçant l'idée que ces émotions sont antagonistes et rarement présentes ensemble. La peur tend à apparaître avec des intensités modérées à élevées, tandis que la joie se manifeste plutôt avec des intensités faibles à modérées. Cette relation confirme que ces émotions peuvent être exploitées pour différencier des sentiments positifs **(joie)** {1}, des sentiments négatifs **(peur)** {-1}.

I.4 Les distributions de l'intensité de la peur (fear_intensity) et de la tristesse (sadness_intensity)

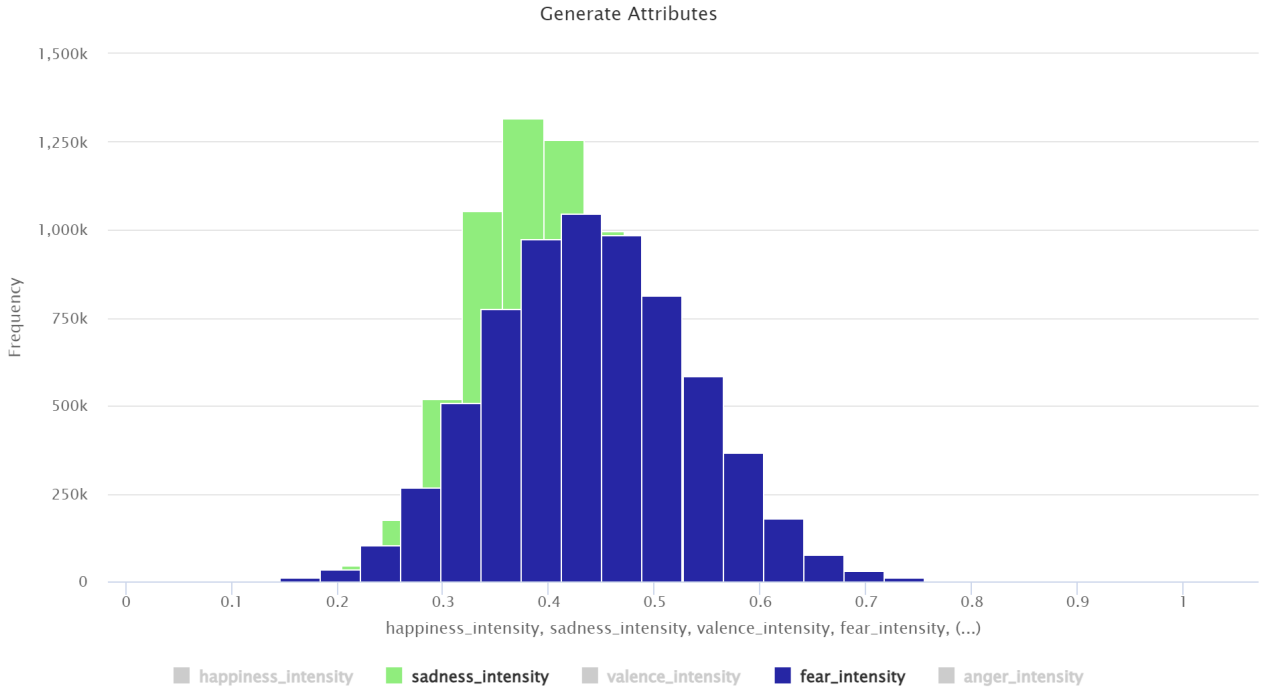


FIGURE 4 – Les distributions de l'intensité de la peur (fear_intensity) et de la tristesse (sadness_intensity)

Interprétation

- **Superposition des distributions :** La distribution de la peur est centrée autour de 0,4-0,5, avec une répartition modérément large. La distribution de la tristesse est davantage concentrée autour de 0,3, avec une variance plus faible par rapport à celle de la peur. Il existe un chevauchement dans la plage de 0,3 à 0,4, bien qu'elles restent globalement distinctes. Ce chevauchement partiel montre que peur et tristesse peuvent être liées dans certaines situations. Cela pourrait refléter des expériences où une intensité modérée de peur coexiste avec de la tristesse, comme dans des contextes de détresse émotionnelle.
- **Détection de sentiments :** Le chevauchement limité et les différences dans les pics des distributions montrent que ces deux caractéristiques peuvent être utilisées pour distinguer des nuances dans les sentiments négatifs. Par exemple :
  - Une intensité élevée de peur accompagnée d'une intensité modérée de tristesse pourrait indiquer une situation de stress aigu.
  - Une intensité faible de peur associée à une intensité élevée de tristesse pourrait correspondre à un sentiment de désespoir.

  Bien que la peur et la tristesse présentent un chevauchement modéré, leur distinction dans les distributions reflète leur complémentarité dans l'expression des sentiments négatifs.

I.5 Les distributions de l'intensité du bonheur (happiness_intensity) et de la tristesse (sadness_intensity)

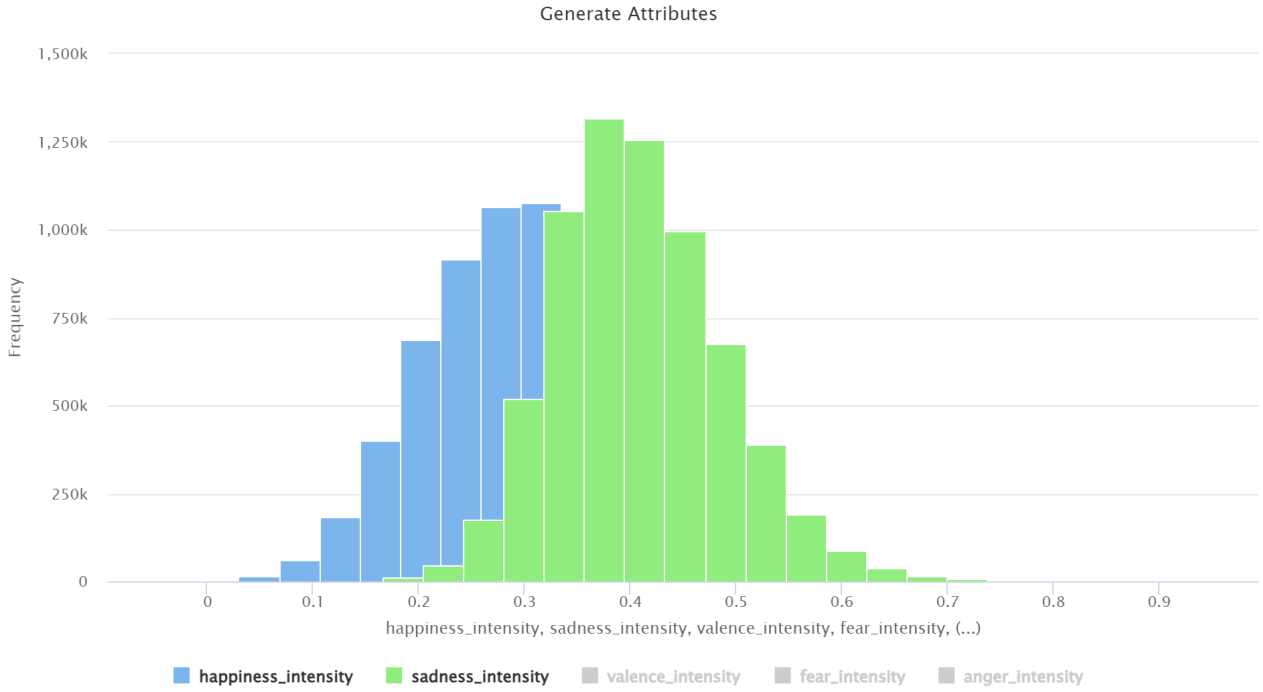


FIGURE 5 – Les distributions de l'intensité du bonheur (happiness_intensity) et de la tristesse (sadness_intensity)

Interprétation

— **Superposition des distributions :**

1. **Répartition des intensités :**
   — Les intensités de *happiness* et de *sadness* montrent des distributions distinctes mais partiellement chevauchantes.
   — **Bonheur** : La distribution est centrée autour de 0.2-0.3, avec une décroissance rapide après 0.3. Elle est plus concentrée dans les faibles valeurs.
   — **Tristesse** : La distribution est centrée autour de 0.3-0.4, avec une amplitude légèrement plus large et une décroissance plus progressive jusqu'à 0.6.
2. **Chevauchement :**
   — La zone de chevauchement la plus importante se situe autour de 0.3.
   — En revanche, pour des intensités faibles (<0.2), le *bonheur* est beaucoup plus représenté que la *tristesse*.
   — À partir de 0.4, la *tristesse* devient dominante, tandis que le *bonheur* diminue rapidement.

— **Détection de sentiments :**

Ces distributions suggèrent que *happiness* et *sadness* sont généralement opposé, mais il existe une plage intermédiaire où ces deux émotions peuvent coexister à faible ou moyenne intensité (autour de 0.3). Cela peut refléter des expériences émotionnelles mixtes, où une personne pourrait ressentir des traces de *bonheur* même dans des situations empreintes de *tristesse*, ou vice versa.

I.6 Les distributions de l'intensité de la valence (valence_intensity) et de la colère (anger_intensity)

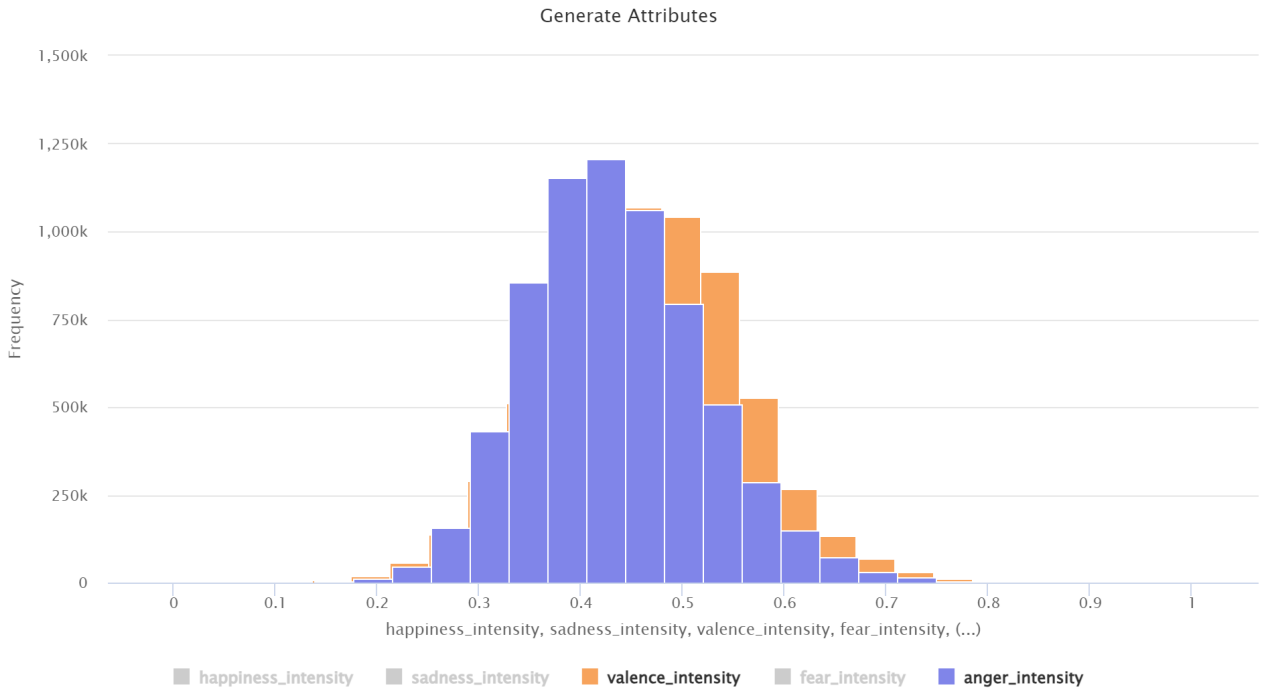


FIGURE 6 – Les distributions de l'intensité de la valence (valence_intensity) et de la colère (anger_intensity)

Interprétation

- **Superposition des distributions :**
  - **Valence Intensity :** La distribution est centrée autour de 0.4-0.5, avec une décroissance symétrique des deux côtés. La valence semble présenter une variation relativement équilibrée et large, couvrant des intensités allant de 0.2 à 0.7.
  - **Anger Intensity :** La distribution est également centrée autour de 0.4, mais elle est plus condensée, avec une chute plus rapide en dehors de cette plage. Les intensités élevées (>0.6) sont rares pour la colère.
  - **Zone de chevauchement** : Les distributions se chevauchent presque complètement, en particulier autour de la zone 0.4-0.5, ce qui indique que ces deux attributs sont souvent présents simultanément à ces niveaux.
- **Détection de sentiments :**
  - **Relation entre valence et colère** : La valence mesure généralement la "valeur émotionnelle" globale (positive ou négative). Sa distribution plus large reflète probablement une plus grande variété de contextes émotionnels représentés. La colère est plus spécifique et semble souvent modérée en intensité, ce qui explique sa concentration autour de 0.4.
  - **Zones de co-occurrence** La forte superposition à des intensités modérées (0.3-0.5) suggère que la colère est souvent liée à une valence légèrement négative ou neutre. Par exemple, une colère modérée pourrait s'exprimer dans des situations où l'émotion générale n'est pas extrême.

  Valence et colère montrent une forte corrélation dans leur plage centrale, mais la valence a une plus grande variabilité.

I.7 Les distributions de l'intensité de la valence (valence_intensity) et de la peur (fear_intensity)

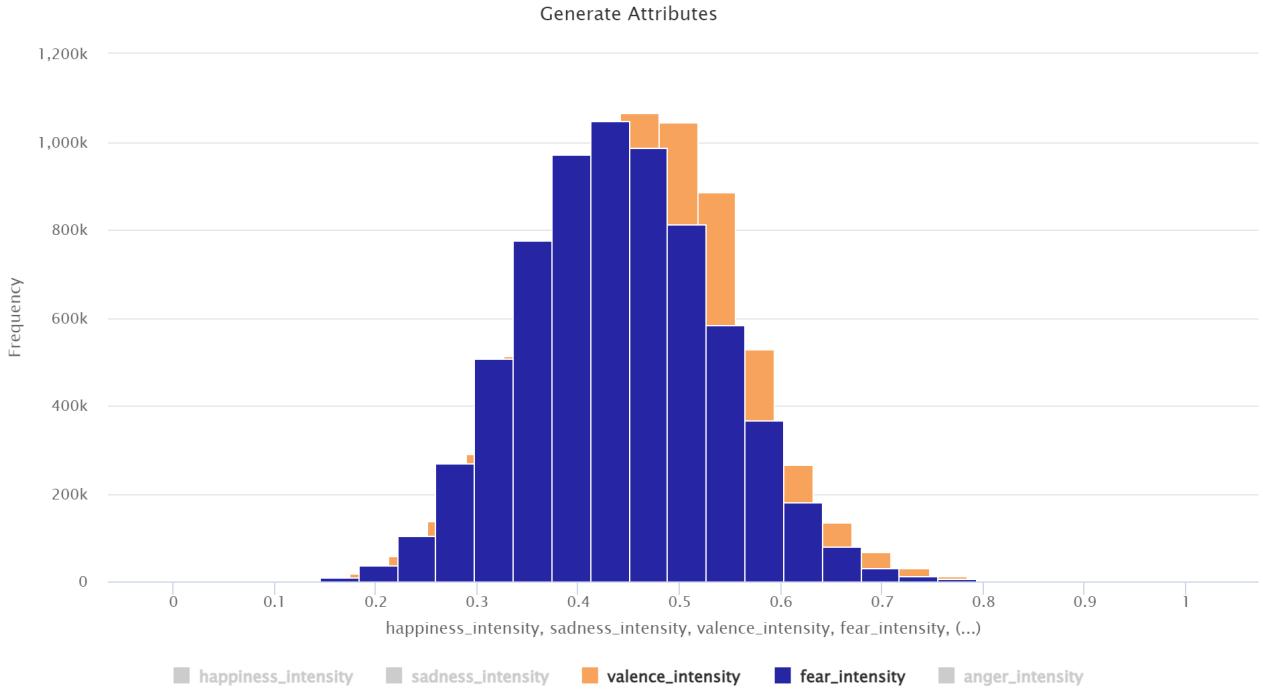


FIGURE 7 – Les distributions de l'intensité de la valence (valence_intensity) et de la peur (fear_intensity)

- **Superposition des distributions :**
  - **Valence Intensity :** Distribution centrée autour de 0.4-0.5, avec une légère dissymétrie. La valence couvre une plage relativement étendue de 0.2 à 0.7.
  - **Fear Intensity :** Distribution également centrée autour de 0.4-0.5, mais elle est plus large, montrant une présence notable de la peur sur une gamme plus variée d'intensités.
  - **Superposition :** Les deux distributions se chevauchent largement autour des intensités modérées (0.4-0.6), indiquant une co-occurrence fréquente de ces deux attributs émotionnels.

Interprétation

Relation entre valence et peur

- La valence capture une évaluation émotionnelle globale (positive/négative), alors que la peur est une émotion spécifique. La co-occurrence autour des valeurs moyennes peut indiquer des situations où une évaluation neutre ou légèrement négative est accompagnée d'une intensité modérée de peur.

Zones de co-occurrence

- La plage 0.4-0.6, où les deux distributions atteignent leur pic, pourrait correspondre à des états émotionnels ambigus ou modérés, tels qu'une légère appréhension ou inquiétude.

— Les faibles intensités de peur (0.2-0.3) associées à des intensités similaires de valence pourraient correspondre à des contextes légèrement négatifs mais pas excessivement menaçants.

I.8 Les distributions de l'intensité de la valence (valence_intensity) et de l'intensité de la joie (happiness_intensity).

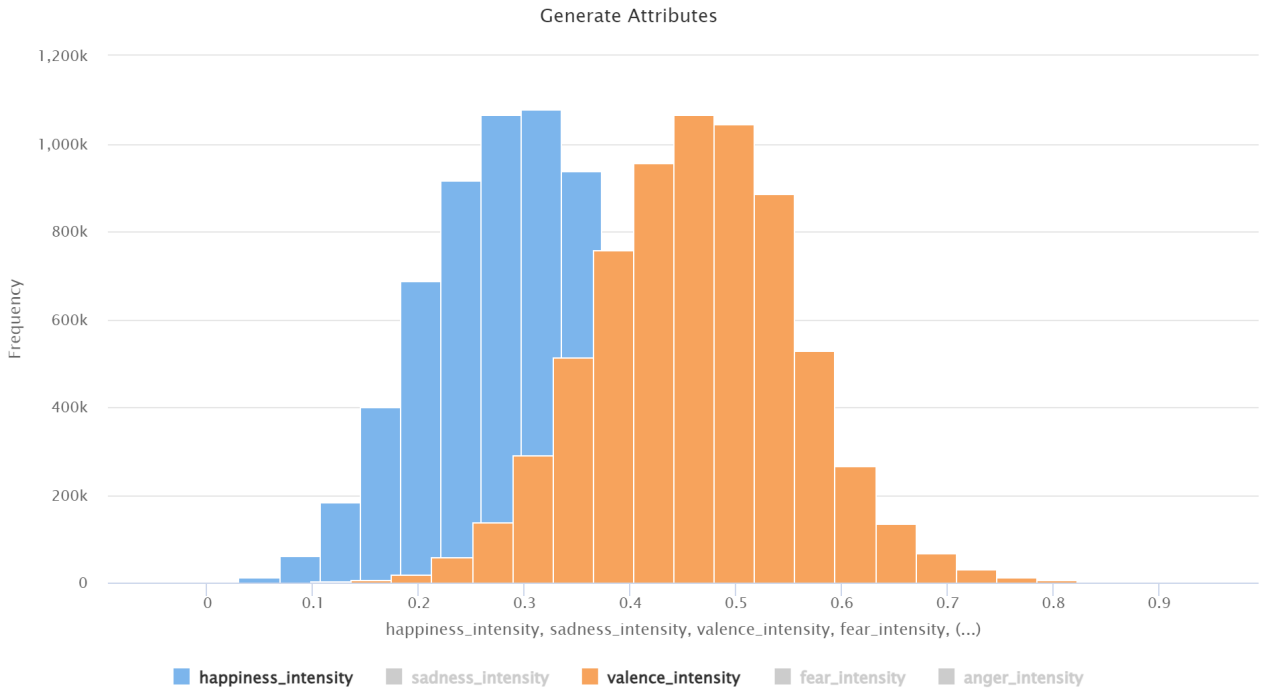


FIGURE 8 – Les distributions de l'intensité de la valence (valence_intensity) et de l'intensité de la joie (happiness_intensity)

Interprétation

Superposition des distributions

— Les deux distributions (valence en orange et bonheur en bleu) se chevauchent de manière significative dans la zone allant d'environ 0.2 à 0.5, bien que la distribution de la valence ait une tendance plus étendue vers des valeurs plus élevées (jusqu'à 0.7).

— La distribution de la joie a son pic autour de 0.3, tandis que celle de la valence atteint un pic un peu plus large, autour de 0.4 à 0.5. Cela indique que la joie et la valence (c'est-à-dire le caractère positif d'une émotion) sont fortement corrélées, bien que la valence ait tendance à être légèrement plus élevée.

Relation entre valence et joie

— La superposition importante et les valeurs moyennes proches de ces deux caractéristiques suggèrent que, dans la plupart des cas, une forte intensité de valence est associée à une intensité de joie similaire. Cependant, la valence n'est pas toujours strictement équivalente à la joie et peut aussi représenter d'autres émotions positives, ce qui explique pourquoi elle couvre une plus grande plage de valeurs.

— Ces distributions peuvent aider à détecter des sentiments positifs. Par exemple, une forte valence avec une joie élevée pourrait indiquer un sentiment très positif, tandis qu'une valence élevée avec une joie plus faible pourrait signaler une émotion positive plus nuancée, comme de la satisfaction ou de la gratitude.

Conclusion

La joie et la valence ont des distributions très similaires, ce qui est cohérent avec le fait que la valence reflète la tonalité positive des émotions. Cette superposition suggère que ces deux caractéristiques peuvent être exploitées conjointement pour la détection de sentiments positifs.

I.9 Les distributions de l'intensité de la valence (valence_intensity) et de l'intensité de la tristesse (sadness_intensity)

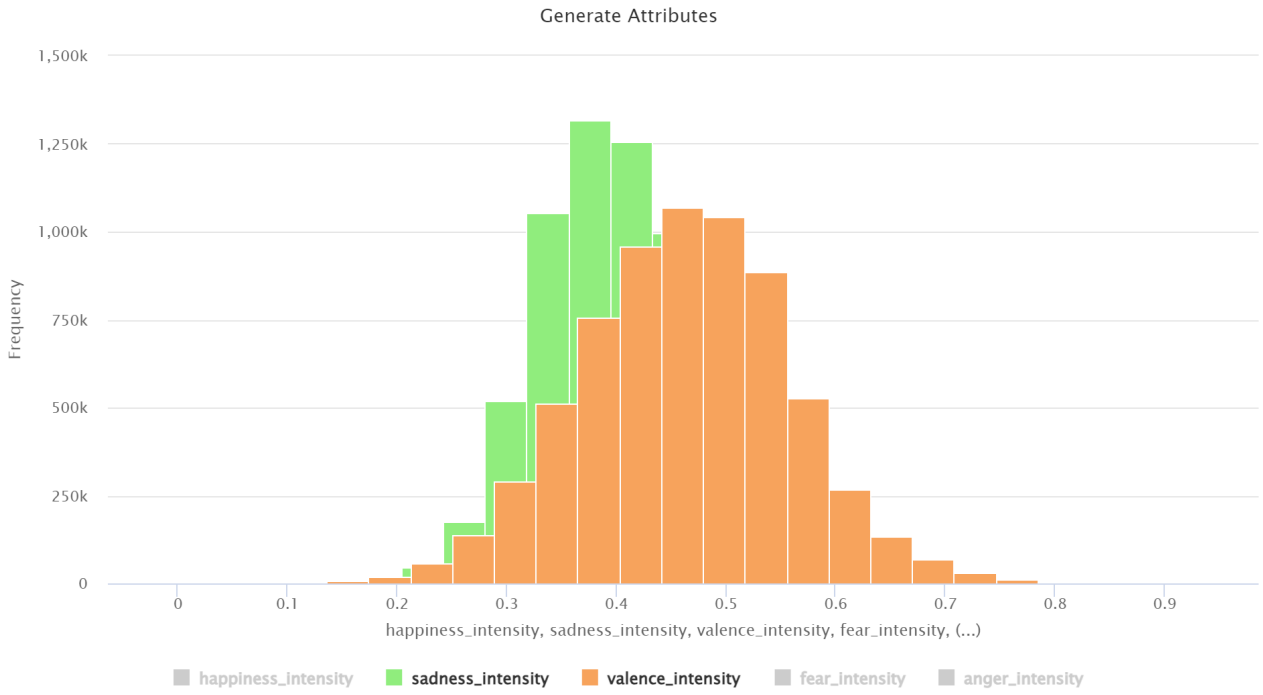


FIGURE 9 – Les distributions de l'intensité de la valence (valence_intensity) et de l'intensité de la tristesse (sadness_intensity)

Interprétation

Comparaison des distributions

— La distribution de la tristesse (en vert) est décalée vers des valeurs plus faibles par rapport à celle de la valence (en orange). La tristesse est concentrée principalement entre 0.2 et 0.4, avec un pic autour de 0.3, tandis que la valence est plus étendue et atteint des valeurs plus élevées, avec un pic autour de 0.4-0.5.

— Cela signifie qu'en général, les sentiments de tristesse, lorsqu'ils sont présents, ont tendance à être de faible à modérée intensité. La valence, quant à elle, semble englober une gamme plus large d'intensités, représentant ainsi une plus grande diversité d'émotions, dont certaines peuvent être moins négatives que la tristesse pure.

Relation entre valence et tristesse

— La faible superposition entre les deux distributions confirme que la valence et la tristesse capturent des aspects émotionnels différents. Une intensité de tristesse élevée est souvent associée à une valence faible, mais les deux ne sont pas totalement corrélées. En effet, certaines émotions négatives légères peuvent avoir une valence non négligeable sans être associées à la tristesse.

— Ces distributions pourraient être utiles pour distinguer les nuances émotionnelles. Par exemple, une intensité de tristesse élevée combinée à une faible valence pourrait indiquer un sentiment très négatif, tandis qu'une faible tristesse associée à une valence plus élevée pourrait refléter une émotion neutre à légèrement négative.

I.10 Distribution entre la peur (fear_intensity) et la colère (anger_intensity)

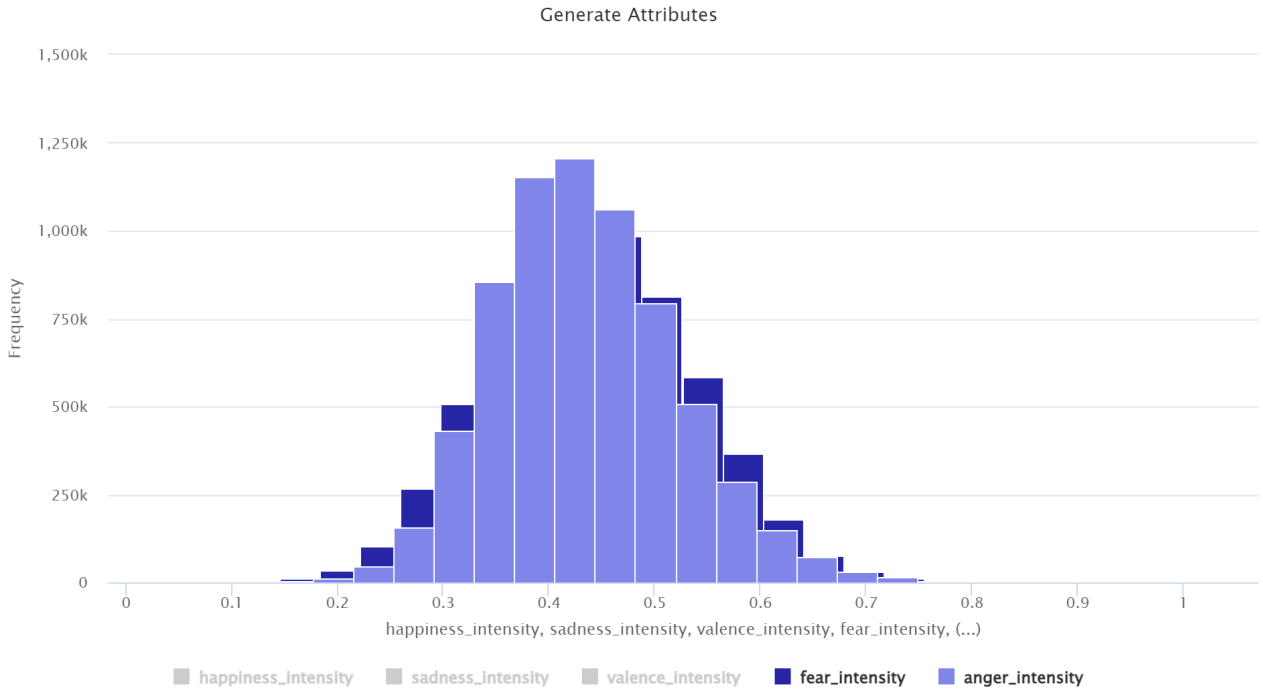


FIGURE 10 – Distribution entre la peur (fear_intensity) et la colère (anger_intensity)

Analyse spécifique par rapport à l'énoncé

Relation entre la peur (`fear_intensity`) et la colère (`anger_intensity`)

— Ces deux émotions semblent avoir des distributions similaires en termes de forme et de plages d'intensité (situées entre 0.2 et 0.6 principalement).

— **Interprétation** : Cela pourrait indiquer que la peur et la colère coexistent fréquemment dans certaines situations émotionnelles, comme lors d'expériences de danger ou de conflit. Ces deux émotions pourraient être exploitées ensemble pour identifier un sentiment négatif.

II Utilisation de l'algorithme K-means avec $K \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$ et visualisation des clusters à l'aide d'une projection UMAP en deux dimensions.

Pour utiliser l'algorithme K-means avec des valeurs de k allant de 1 à 10, nous avons d'abord isolé les attributs pertinents liés au sentiment à l'aide de l'opérateur **Select Attributes**. Ensuite, nous avons utilisé l'opérateur **Multiply**, qui permet d'utiliser le jeu de données (*canada_data*) plusieurs fois sans que les opérations effectuées sur une copie n'affectent les autres ou les résultats globaux. Par la suite, nous avons appliqué l'opérateur **Cluster Distance Performance**, en choisissant comme critère principal l'indice de Davies-Bouldin, afin d'identifier la valeur de k correspondant au meilleur cluster et, par conséquent, à la meilleure classification des données.

Pour effectuer la projection UMAP en deux dimensions, nous avonsexecuter un code sur google colab que nous avons envoyé en pieces jointes. Cette méthode réduit la dimensionnalité du jeu de données tout en préservant autant que possible les informations essentielles.

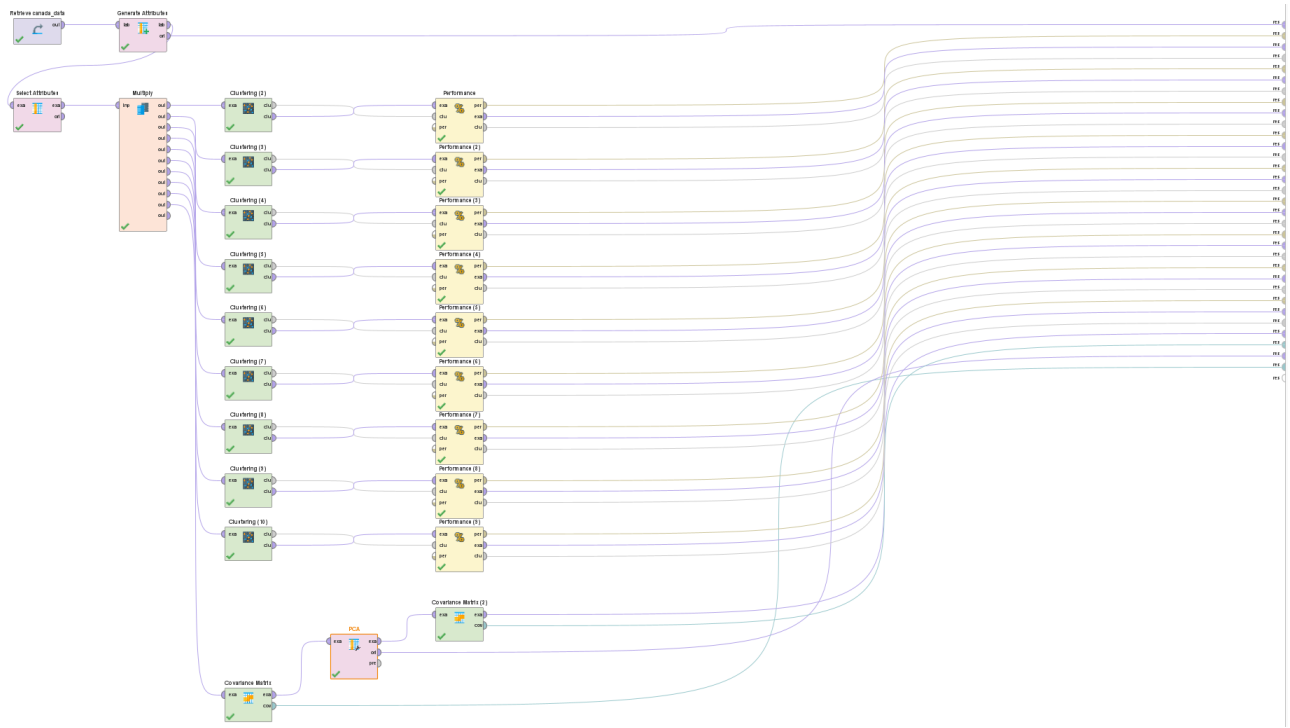


FIGURE 11 – Process2

II.1 K-means avec $K \in \{2, 3, 4, 5, 6, 7, 8, 9, 10\}$

Analyse des résultats

- L'indice de Davies-Bouldin mesure la qualité des clusters, avec des valeurs plus faibles indiquant une meilleure compacité intra-cluster et séparation inter-cluster. - Le meilleur résultat est obtenu pour $k = 2$, avec un indice de 0.863, ce qui est significativement plus faible que les autres valeurs de k. - En augmentant k, l'indice tend à se dégrader légèrement, suggérant que des valeurs élevées de k réduisent la qualité des clusters.

k	**Indice de Davies-Bouldin**
2	0.863
3	0.975
4	1.076
5	1.193
6	1.326
7	1.313
8	1.302
9	1.333
10	1.313

TABLE 1 – Performances des clusters pour k allant de 2 à 10 selon l'indice de Davies-Bouldin.

Conclusion

Le meilleur clustering est obtenu avec $k = 2$, car il présente l'indice de Davies-Bouldin le plus faible.

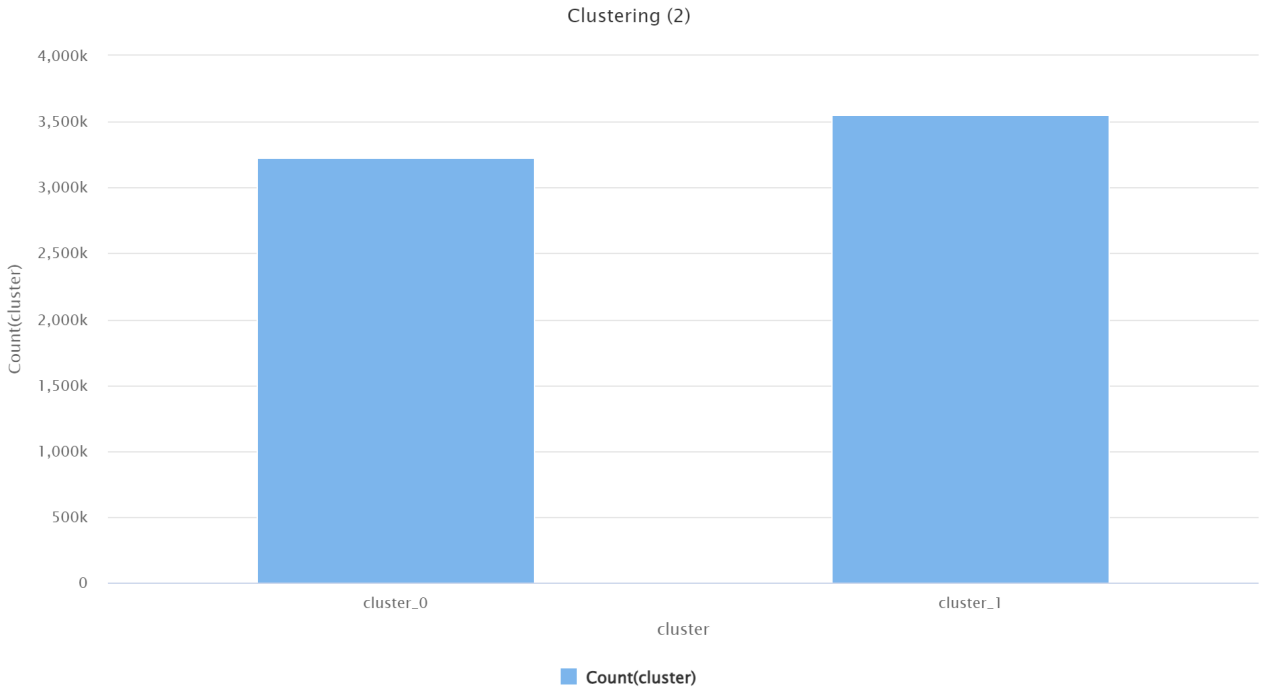


FIGURE 12 – Cluster 2

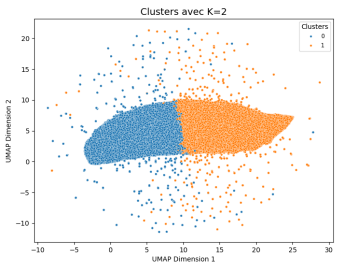
Nous présentons ici une représentation graphique sous forme de diagramme à barres illustrant les deux clusters identifiés. Après analyse, le **cluster 0** est associé aux sentiments négatifs, tandis que le **cluster 1** représente les sentiments positifs. Cependant, notre étude révèle que certaines personnes peuvent exprimer des sentiments mitigés. Pour mieux différencier les sentiments positifs des négatifs, nous envisageons l'introduction d'un **cluster 3**, qui permettrait de mieux capturer ces nuances.

II.2 la projection UMAP en deux dimensions

Pour cette question, un script Python a été développé pour analyser et visualiser des données de sentiments en appliquant des algorithmes de clustering. Les principales étapes comprennent :

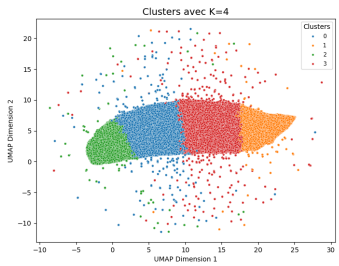
— **Réduction de dimensionnalité** avec l'algorithme **UMAP**, projetant les données en deux dimensions pour faciliter la visualisation.
— **Clustering K-means**, testé pour des nombres de clusters (k) allant de 2 à 10.
— **Visualisation des clusters**, où les résultats ont été tracés sur une grille de graphiques 2D pour chaque k.

Ce code permet d'identifier des sous-groupes dans les données sentimentales tout en explorant la séparation et la structure des clusters. Le code détaillé sera fourni en **annexe** pour référence.

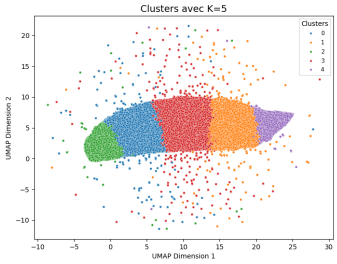


(a) `clusters_k_2`

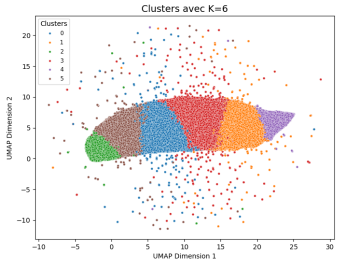
(b) `clusters_k_3`



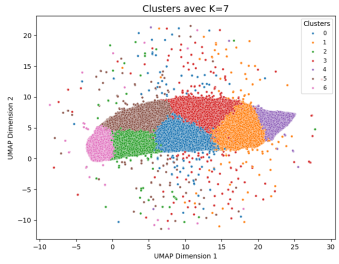
(c) `clusters_k_4`



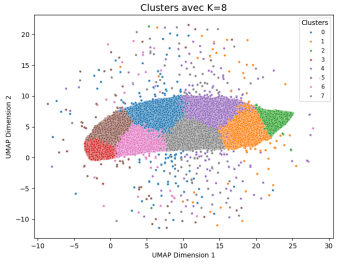
(d) `clusters_k_5`



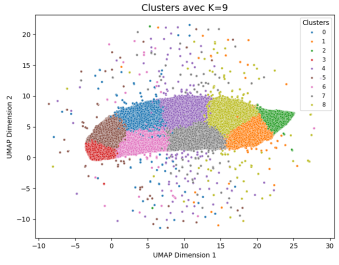
(e) `clusters_k_6`



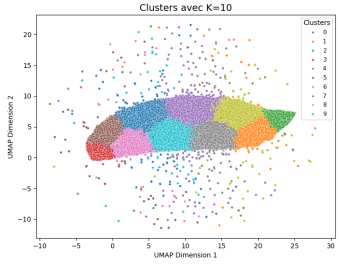
(f) `clusters_k_7`



(g) `clusters_k_8`



(h) `clusters_k_9`



(i) `clusters_k_10`

FIGURE 13 – Grille de clusters pour différentes valeurs de k (de 2 à 10).

Interprétation :

1. **Évolution des clusters :**
   — Pour $k = 2$, les données sont divisées en deux grandes catégories, ce qui peut refléter une séparation globale entre **sentiments positifs** et **sentiments négatifs**.

— À mesure que k augmente, les clusters deviennent plus nombreux, révélant des sous-groupes plus précis au sein des deux catégories principales. Cela pourrait indiquer des nuances telles que des émotions mixtes (par exemple, tristesse mêlée de joie ou colère modérée).

2. **Complexité croissante :**
   — À partir de $k = 6$ ou $k = 7$, certains clusters commencent à se chevaucher ou deviennent moins clairement définis. Cela peut suggérer que l'augmentation du nombre de clusters introduit une segmentation excessive, réduisant la pertinence de l'analyse.
3. **Interprétation des dimensions projetées :**
   — Les deux axes générés par **UMAP** simplifient la structure multidimensionnelle des données. Cette réduction peut entraîner des pertes d'informations, et les clusters doivent être interprétés avec prudence. Chaque cluster peut représenter un ensemble d'émotions dominantes, telles que la joie, la colère ou la tristesse, en fonction de la position des données dans l'espace projeté.

III Evaluation Des differents Clusters

III.1 Tableau des Scores de Silhouette et des Indices de Davies-Bouldin

le tableau combiné des scores de silhouette et des indices de Davies-Bouldin pour chaque nombre de clusters k allant de 2 à 10 :

k	$ScoredeSilhouette$	$IndicedeDavies - Bouldin$
2	0.4010	0.863
3	0.3105	0.975
4	0.2565	1.076
5	0.2252	1.193
6	0.1928	1.326
7	0.2011	1.313
8	0.1992	1.302
9	0.1816	1.333
10	0.1868	1.313

TABLE 2 – Performances des clusters pour k allant de 2 à 10, selon les scores de silhouette et les indices de Davies-Bouldin.

Analyse et Comparaison

Le score de silhouette diminue généralement avec l'augmentation de k, tandis que l'indice de Davies-Bouldin augmente, indiquant que la cohésion et la séparation des clusters se dégradent au fur et à mesure que k augmente. Cela est généralement interprété comme un signe de baisse de la qualité du clustering.

— **Meilleur score de silhouette** : Le meilleur score de silhouette est obtenu pour $k = 2$ avec un score de 0.4010.
— **Indice de Davies-Bouldin minimal** : L'indice de Davies-Bouldin le plus bas est également pour $k = 2$ avec une valeur de 0.863.
— **Tendance** : Les scores de silhouette diminuent et les indices de Davies-Bouldin augmentent, indiquant que la qualité des clusters se détériore avec l'augmentation de k.

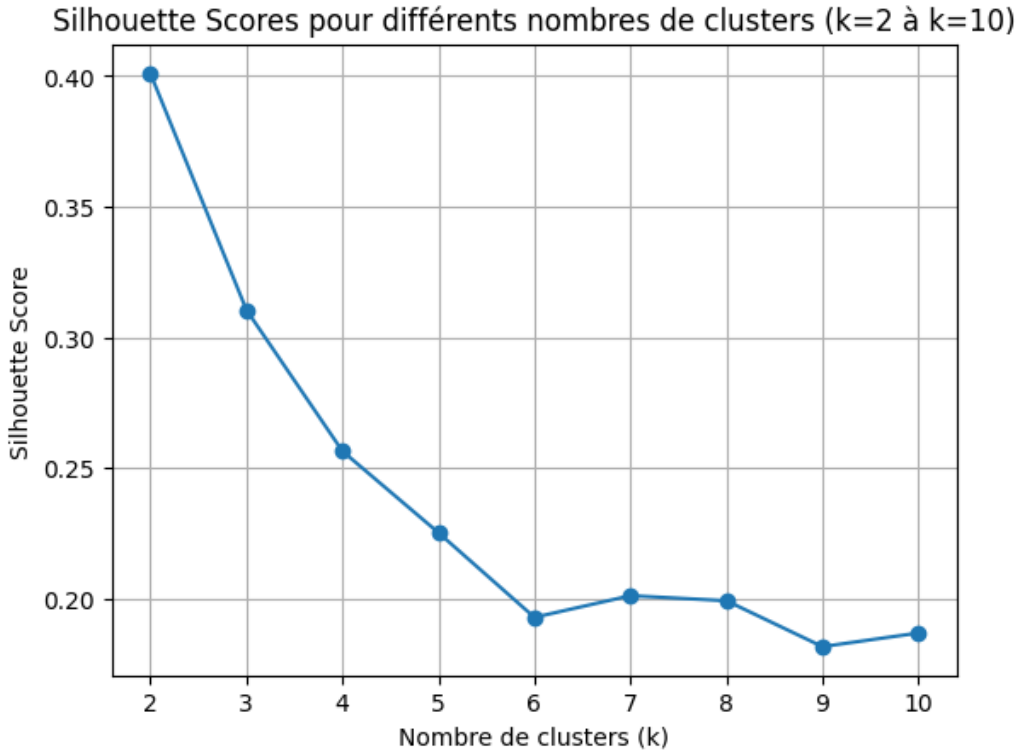


FIGURE 14 – Les scores de silhouette

Conclusion

Le meilleur nombre de clusters pour cette analyse, en se basant à la fois sur le score de silhouette et l'indice de Davies-Bouldin, est $k = 2$, car c'est le meilleur cluster expliquant la cohésion et la séparation des clusters.

III.2 Interprétation et Calcul des Précision, Rappel et F1-score

Après avoir effectué une nouvelle évaluation et utilisé les critères de précision, rappel (recall) et F1-score, nous avons obtenu une matrice de confusion qui nous a permis d'obtenir le tableau ci-dessous, accompagné des conclusions suivantes :

PerformanceVector

PerformanceVector:
accuracy: 99.13%
ConfusionMatrix:

True:	negative	neutral or mixed	positive	very negative	very positive
negative:	1054992	0	0	0	
neutral or mixed:	0	336998	0	0	0
positive:	0	0	519876	0	208
very negative:	0	0	0	86374	0
very positive:	0	0	17494	0	18221

FIGURE 15 – Matrice de confusion

Évaluation des métriques de classification

Calculs par classe

Pour chaque classe, nous calculons la précision, le rappel et le F1-score à partir de la matrice de confusion donnée.

1. Classe "negative"

- **Vrai Positifs (TP)** : 1054992
- **Faux Positifs (FP)** : 0
- **Faux Négatifs (FN)** : 0

Précision :

$$Precision = \frac{TP}{TP + FP} = \frac{1054992}{1054992 + 0} = 1.0$$

Rappel (Recall) :

$$Recall = \frac{TP}{TP + FN} = \frac{1054992}{1054992 + 0} = 1.0$$

F1-score :

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 2 \cdot \frac{1.0 \cdot 1.0}{1.0 + 1.0} = 1.0$$

2. Classe "neutral or mixed"

- **Vrai Positifs (TP)** : 336998
- **Faux Positifs (FP)** : 0
- **Faux Négatifs (FN)** : 0

Précision :

$$Precision = \frac{TP}{TP + FP} = \frac{336998}{336998 + 0} = 1.0$$

Rappel (Recall) :

$$Recall = \frac{TP}{TP + FN} = \frac{336998}{336998 + 0} = 1.0$$

F1-score :

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 2 \cdot \frac{1.0 \cdot 1.0}{1.0 + 1.0} = 1.0$$

3. Classe "positive"

- **Vrai Positifs (TP)** : 519876
- **Faux Positifs (FP)** : 17494
- **Faux Négatifs (FN)** : 208

Précision :

$$Precision = \frac{TP}{TP + FP} = \frac{519876}{519876 + 17494} \approx 0.9675$$

Rappel (Recall) :

$$Recall = \frac{TP}{TP + FN} = \frac{519876}{519876 + 208} \approx 0.9996$$

F1-score :

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = 2 \cdot \frac{0.9675 \cdot 0.9996}{0.9675 + 0.9996} \approx 0.9833$$

Conclusion

Les résultats montrent une excellente performance pour les classes "negative", "neutral or mixed", "positive", et "very negative" avec des F1-scores élevés proches de 1.0. Cependant, la classe "very positive" présente un F1-score inférieur (≈ 0.6760), en raison d'un faible rappel (≈ 0.5103). Cela indique que ce modèle a des difficultés à capturer correctement les données de cette classe, ce qui pourrait nécessiter une optimisation du modèle ou une meilleure séparation des clusters.

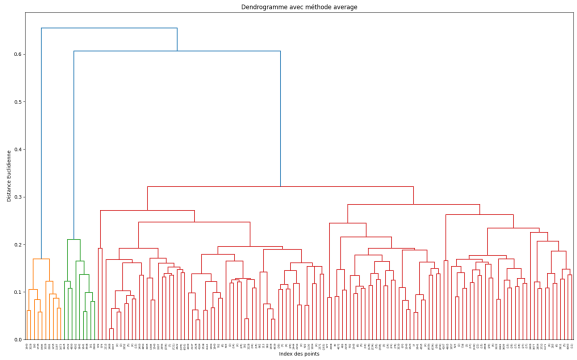
IV Approche hiérarchique et évaluation des clusters obtenus

IV.1 Présentons les différents dendogrammes correspondant

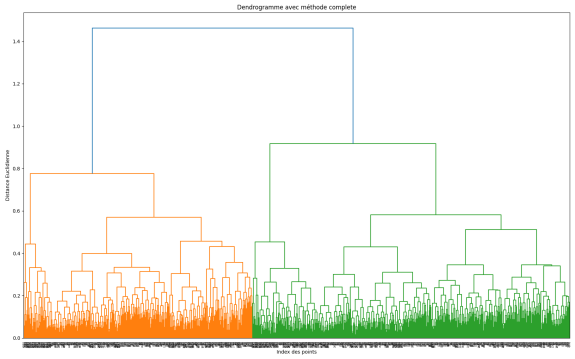
Nous avons exécuté un code en `Python` pour réaliser une analyse de clustering hiérarchique sur un jeu de données en utilisant plusieurs méthodes de liaison, à savoir :

- **Single** : distance minimale entre les points de deux clusters ;
- **Complete** : distance maximale entre les points de deux clusters ;
- **Average** : moyenne des distances entre tous les points de deux clusters ;
- **Ward** : méthode visant à minimiser la variance intra-cluster.

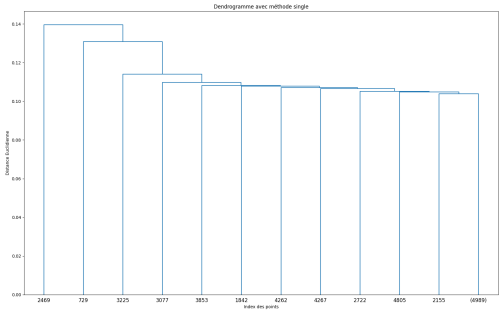
Les dendrogrammes générés permettent de visualiser la structure des clusters et d'identifier un nombre optimal de regroupements. Par ailleurs, l'ajustement de la limite de récursion assure le bon fonctionnement de l'algorithme, même pour des ensembles de données volumineux ou des structures complexes



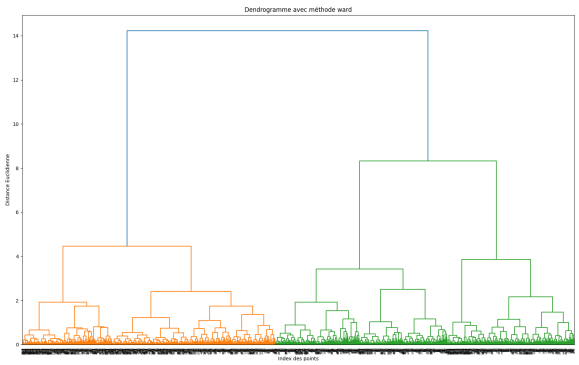
(a) Dendrogramme avec Average



(b) Dendrogramme avec Complete



(c) Dendrogramme avec Single



(d) Dendrogramme avec Ward

FIGURE 16 – Comparaison des dendrogrammes obtenus avec différentes méthodes de liaison

IV.2 Évaluation des clusters obtenus à travers différentes métriques

Après avoir fixé différentes valeurs de seuil, nous avons obtenu les résultats suivants, que nous avons interprétés de la manière suivante :

Méthode	Seuil	Nombre de clusters	Silhouette
Single	5	1	Non calculé
Single	10	1	Non calculé
Single	15	1	Non calculé
Single	20	1	Non calculé
Complete	5	1	Non calculé
Complete	10	1	Non calculé
Complete	15	1	Non calculé
Complete	20	1	Non calculé
Average	5	1	Non calculé
Average	10	1	Non calculé
Average	15	1	Non calculé
Average	20	1	Non calculé
Ward	5	3	0.309
Ward	10	2	0.399
Ward	15	1	Non calculé
Ward	20	1	Non calculé

TABLE 3 – Résultats obtenus selon les méthodes et les seuils.

Interprétation

- Les méthodes **single**, **complete** et **average** ne donnent pas de résultats exploitables à ces seuils, car elles forment un seul cluster. Cela pourrait être dû à des seuils mal ajustés ou à une structure particulière dans les données.
- La méthode **ward** produit des clusters significatifs :
  - À un seuil de **5**, elle identifie **3 clusters**, ce qui semble prometteur pour répondre à la question sur la formation de 3 groupes naturels.
  - À un seuil de **10**, elle identifie **2 clusters** avec une meilleure cohérence (*silhouette* = 0.399).
- Les résultats suggèrent que **la méthode "ward" est la plus adaptée aux données**, et qu'il pourrait être pertinent d'ajuster les seuils autour de **5** pour explorer plus en détail la structure en 3 clusters.
- **Le seuil optimal** pour trouver naturellement trois clusters est : **5** avec **la méthode Ward**.

V Clustering hiérarchique avec $k = 3$ et évaluation des résultats obtenus

Pour répondre à cette question, nous avons d'abord converti le label "sentiment" en valeurs numériques : négatif (-1), neutre (0) et positif (1). Ensuite, nous avons appliqué un clustering hiérarchique avec la méthode "Ward" et la distance euclidienne, afin de créer trois clusters à partir des données. Les clusters prédits ont été affichés et comparés aux sentiments des internautes. Les résultats suivants ont été obtenus :

— **Précision** : 0.048
— **Rappel** : 0.1303
— **Score F1** : 0.0702

Les résultats des métriques montrent que les clusters obtenus ne correspondent pas bien aux catégories de sentiments (négatif, neutre, positif) des internautes. Ces résultats peuvent être expliqués par les éléments suivants :

1. **Précision (0.048)** : La précision mesure la proportion de prédictions positives correctes parmi toutes les prédictions positives. Une précision aussi faible indique que la majorité des éléments identifiés comme appartenant à une catégorie (par exemple, positif) ne sont en réalité pas de cette catégorie. Cela suggère que les clusters formés par le modèle ne sont pas bien alignés avec les sentiments réels des internautes.
2. **Rappel (0.1303)** : Le rappel mesure la proportion d'éléments d'une catégorie qui ont été correctement identifiés comme appartenant à cette catégorie. Un faible rappel indique que beaucoup d'exemples des catégories de sentiment (négatif, neutre, positif) n'ont pas été correctement capturés par les clusters, ce qui suggère que le clustering ne réussit pas à saisir la diversité des sentiments.
3. **Score F1 (0.0702)** : Le score F1 est la moyenne harmonique de la précision et du rappel. Un score aussi faible reflète la mauvaise performance générale du modèle en termes de classification des sentiments. Cela confirme que les clusters obtenus ne correspondent pas bien aux catégories de sentiments des internautes et qu'il existe une grande divergence entre les clusters formés et les véritables sentiments.

Ces faibles valeurs de précision, rappel et score F1 suggèrent que le modèle de clustering hiérarchique avec la méthode "Ward" et la distance euclidienne n'est pas bien adapté pour capturer les nuances des sentiments dans les données.