Rapport de Projet

Analyse de Sentiment

Dans le cadre du master informatique 1ère année de l'Université de Paris, il nous a été donné pour objectif d'explorer et d'analyser des données réelles portant sur des tweets et de proposer un sujet d'analyse.

Concernant les critères de conception, le langage de programmation utilisé pour réaliser l'ensemble du projet est Python 3.0.

Les différentes bibliothèques utilisées sont :

- **Scikit-Learn :** permet d'utiliser des méthodes de machine learning.
- **Pandas :** permet la manipulation et l'analyse des données.
- **Seaborn :** permet de tracer et visualiser les données sous forme graphique.

Ce rapport retrace les travaux et les choix effectués durant ce projet.

Sommaire

I.Objet de l'analyse

L'analyse des sentiments est le processus automatisé d'analyse des données textuelles et de leur tri en sentiments positifs, négatifs ou neutres.

Nous pouvons travailler sur des tweets choisis à partir d'un nom d'utilisateur ou bien à partir d'un thème actuel.

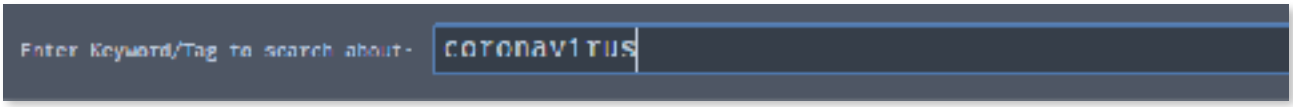

Figure 1. Sélection des tweets ayant pour thème le coronavirus.

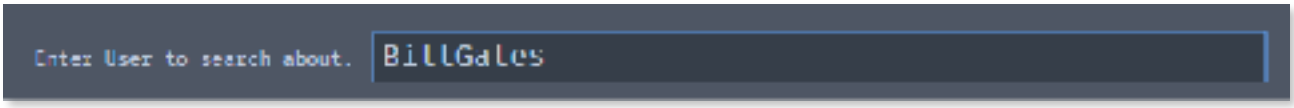

Figure 2. Sélection des tweets ayant pour auteur Bill Gates.

Notre analyse se fera sur les tweet de Bill Gates, ce choix se justifie par le fait qu'ils peuvent toucher plusieurs domaines d'actualité, que ce soit la situation géo-politique de certains pays ou bien l'épidémie du coronavirus.

De plus les tweets ont un certains intérêt, en effet ils peuvent influencer de nombreuses personnes pour ne pas dire des nations en effet la critique d'un choix gouvernemental ou une expression de soutiens sur un événement quelconque peut avoir des conséquences.

Enfin, il faut souligner que Bill Gates est une idole et l'un des influenceurs les plus suivis au monde (51 millions d'abonnements) pour l'ensemble de ses travaux réalisés et l'effort qu'il a consacré pour faciliter la vie humaine.

II.Statistique Descriptive

Le jeu de donnée est composée de 200 tweets, tous en anglais et ayant pour auteur Bill Gates.

	Tweets
0	@narendramodi @gatesfoundation Thank you for t...
1	I'm hopeful that this program will improve our...
2	RT @melincagates: When will America be able to...
3	Class of 2020, these are not easy times. But w...
4	RT @melindagates: To overcome #COVID19 the wor...

Figure 3. Exemple des 5 premiers tweets (avec bruit).

On remarque la présence de bruit dans notre dataset, nous devons donc nettoyer notre jeu de donnée en supprimant toutes les mentions (commence par @), les hashtags (commence par #), tout les retweets(commence par RT) ,les hyper-liens (commence par http) ainsi que tout les stopwords.

	Tweets
0	Thank you for the conversation and partnersh...
1	I m hopeful that this program will improve our...
2	: When will America be able to get back to wor...
3	Class o' 2020, these are not easy times. But w...
4	: To overcome COVID19 the world doesn't just n...

Figure 4. Exemple des 5 premiers tweets (sans bruit).

A ce stade, nous pouvons avoir un aperçu des thèmes dans les tweets.

Figure 5. Nuage de mots

A travers le nuage de mots, on a une visualisation des différents termes les plus fréquents qui constitue le discours de Bill Gates, on s'aperçois que les tendances du moment ressorte, c'est-à-dire l'épidémie du coronavirus et les actions de la Gates Foundation.

Parmi les différentes variables utilisé, nous allons ajouter la subjectivité ainsi que la polarité.

	Tweets	Subjectivity	Polarity
0	Thank you for the conversation and partnersh...	0.316667	0.058333
1	I'm hopeful that this program will improve our...	0.375000	-0.125000
2	: When will America be able to get back to wor...	0.312500	0.250000
3	Class of 2020, these are not easy times. But w...	0.833333	-0.216667
4	: To overcome COVID19 the world doesn't just n...	0.000000	0.000000

Figure 6. Ajout de la Subjectivité et de la Polarité

En effet, l'ajout de ces deux attributs nous permet d'avoir une idée général concernant l'analyse de sentiment dans son ensemble, c'est-à-dire si les tweets positifs/négatifs sont plutôt subjectif ou bien objectif.

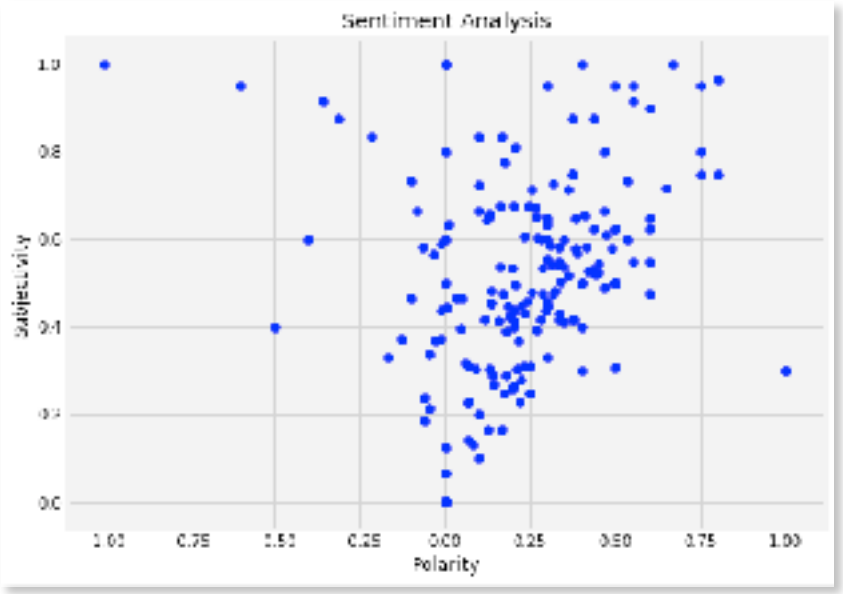

Figure 7. Affichage des tweets en fonction de leurs polarités et de leurs subjectivités .

On voit bien sur la figure précédente que les tweets positifs sont très majoritairement subjectif, on a un partage pour les tweets neutres tandis que les tweets négatifs sont majoritairement subjectifs.

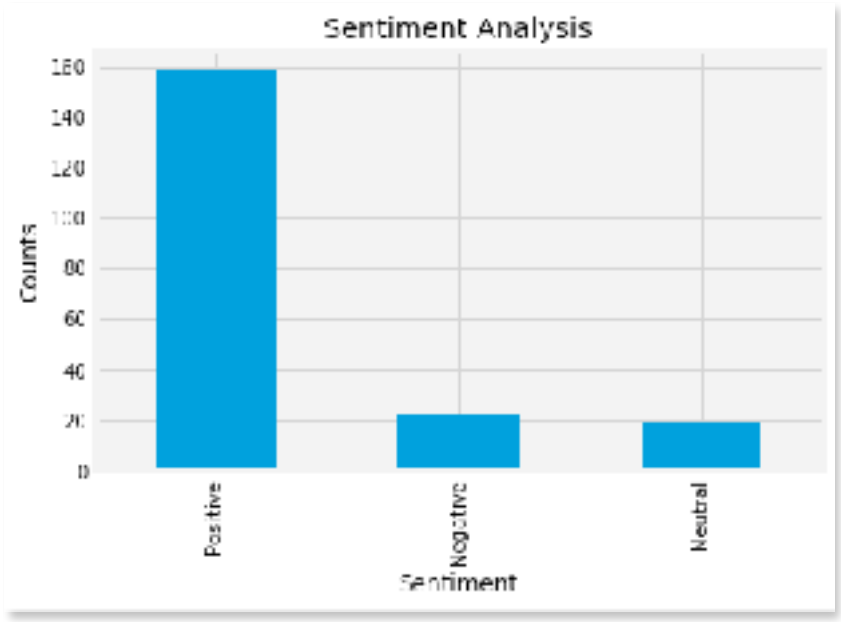

Figure 8. Barplot de la dispersion des tweets en fonction de leurs polarités.

Le pourcentage de tweet positif est de : 79.5%.
Le pourcentage de tweet neutre est de : 9.5%.
Le pourcentage de tweet négatif est de : 11%.

	Tweets	Subjectivity	Polarity	Analysis
0	Thank you for the conversation and partnersh...	0.316667	0.058333	Positive
1	I'm hopeful that this program will improve our...	0.375000	-0.125000	Negative
2	: When will America be able to get back to wor...	0.312500	0.250000	Positive
3	Class of 2020, these are not easy times. But w...	0.833333	-0.216667	Negative
4	: To overcome COVID19 the world doesn't just n...	0.000000	0.000000	Neutral

Figure 9. Ajout des sentiments

On peut donc ajouter les sentiments comme attribut à notre jeu de donnée.

Néanmoins les modèles de Machine Learning ne sont pas très compatible avec les données sous forme de texte, nous devons donc les convertir en format numérique.

En assignant à chaque mot une valeur unique. Chaque document que nous voyons peut être codé comme un vecteur de longueur fixe avec la longueur du vocabulaire des « Known words ». La valeur à chaque position dans le vecteur pourrait être remplie par un nombre ou une fréquence de chaque mot dans le document codé.

On importe alors la fonction CountVectorizer du package scikitlearn.

La fonction CountVectorizer fournit un moyen simple à la fois de tokeniser une collection de documents et de créer un vocabulaire de knows words, mais aussi d'encoder de nouveaux documents en utilisant ce vocabulaire.

En utilisant ces vecteurs on a alors la fréquence d’apparition des mots

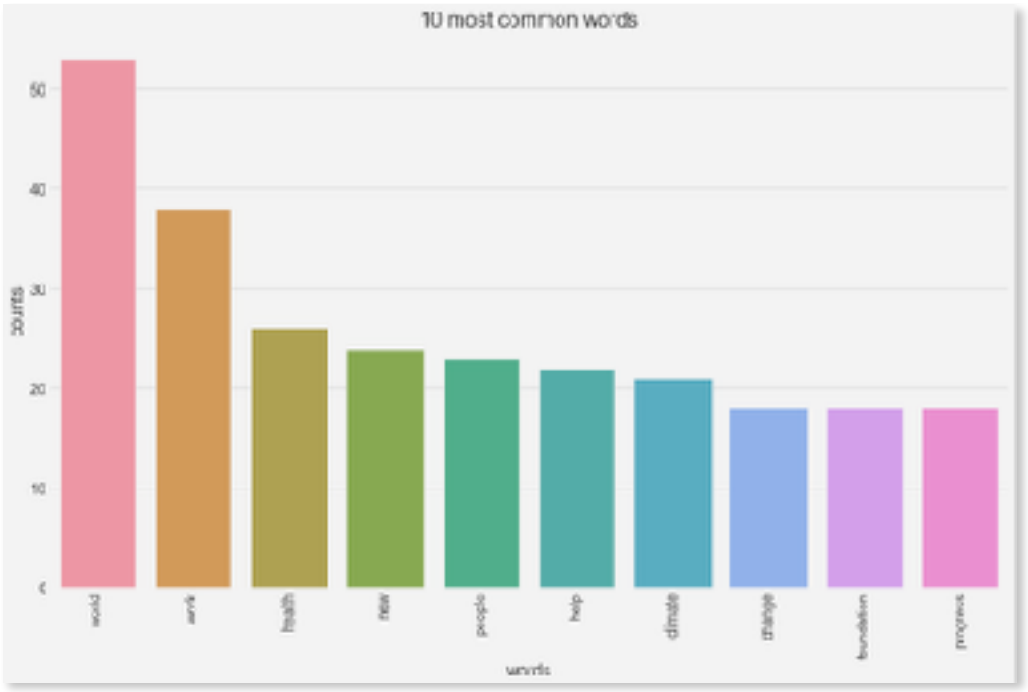

Figure 10. Barplot des 10 mots les plus utilisés

Les mots les plus fréquents font ressortir les thèmes citée précédemment, c'est-à-dire le climat et les actions de la Gates Foundation.

Cependant nous ne pouvons pas seulement déterminer les thèmes à partir de graphique, donc pour analyser les différents thèmes présent dans les tweets de Bill Gates, on utilise **Latent Dirichlet Allocation** (LDA).

La LDA (Latent Dirichlet Allocation) est une méthode non-supervisée générative qui se base sur les hypothèses suivantes :

- Chaque document du corpus est un ensemble de mots sans ordre (bag-of-words)
- Chaque document mm aborde un certain nombre de thèmes dans différentes proportions qui lui sont propres p(θm)
- Chaque mot possède une distribution associée à chaque thème p(ϕk). On peut ainsi représenter chaque thème par une probabilité sur chaque mot.
- En représente le thème du mot wn

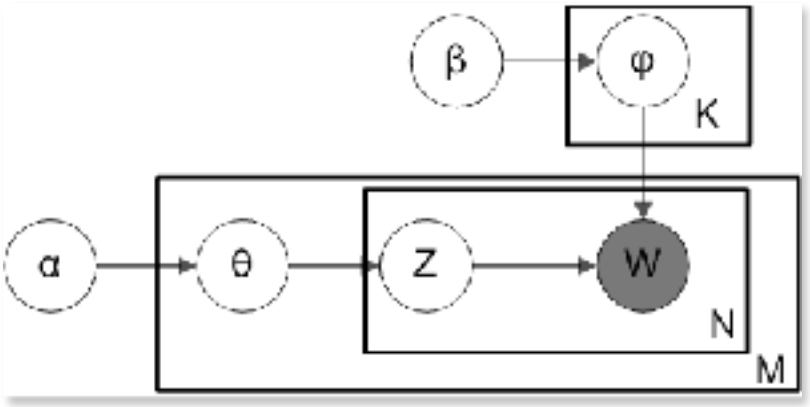

Figure 11. LDA Schema

```
Topics found via LDA:

Topic #0:
world coronavirus like need really help new work opportunity potential

Topic #1:
work world foundation change climate vaccines health excited working people

Topic #2:
people covid19 help need vaccine excited new world today billion

Topic #3:
world years progress work health melinda countries optimistic incredible new

Topic #4:
work world health life read disease lives important new climate
```

Figure 12. Affichage des topics trouvé via LDA

On remarque que les thèmes de la santé, du climat sont omniprésent, ceci confirme les hypothèses émises précédemment en ce qui concerne les situation actuelle et les actions de la Gates Foundation.

Sur l'aspect analyse de sentiment, nous avons décidé de modéliser nos topics à l'aide du package PyLDavis.

L'interface nous fourni:

- Un panel à gauche fournissant une vue global du model, quelle importance à un topic et à quel point sont-ils reliés les uns aux autres.
- Un panel à droite contenant un diagramme en bâton, indiquant quels mots ont été les plus utilisés.

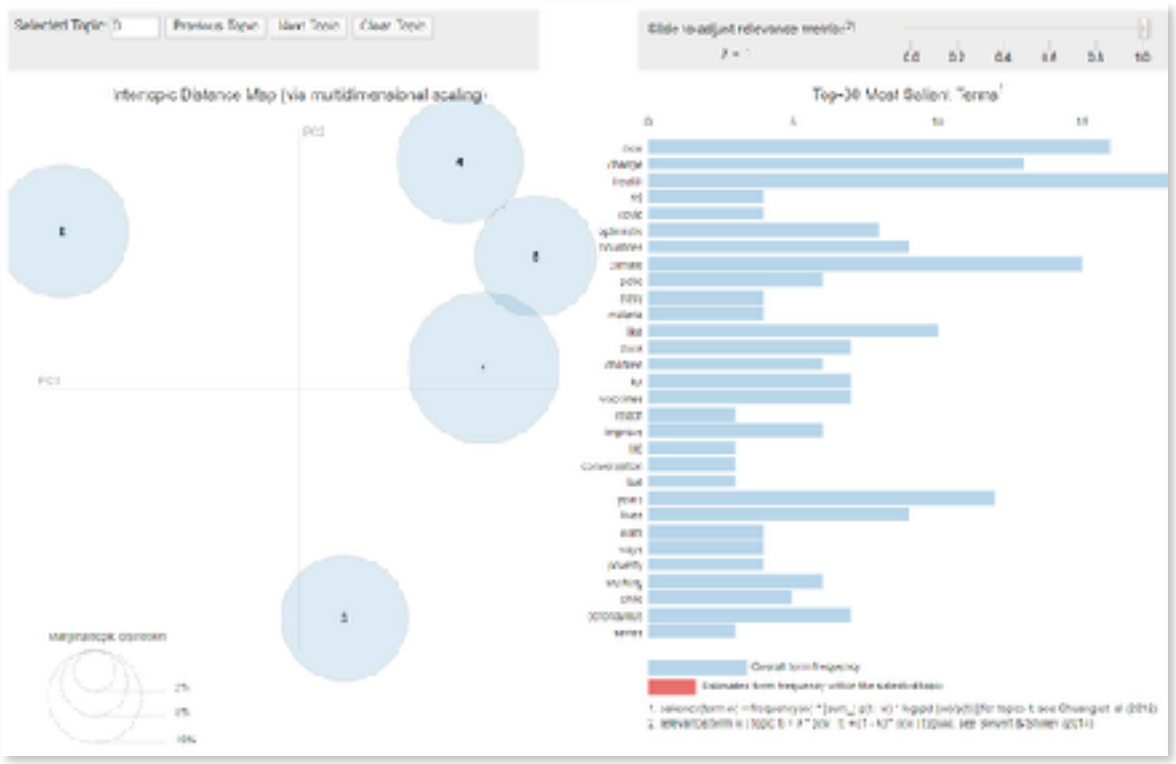

Figure 13. Ensemble global de **PyLDAvis**

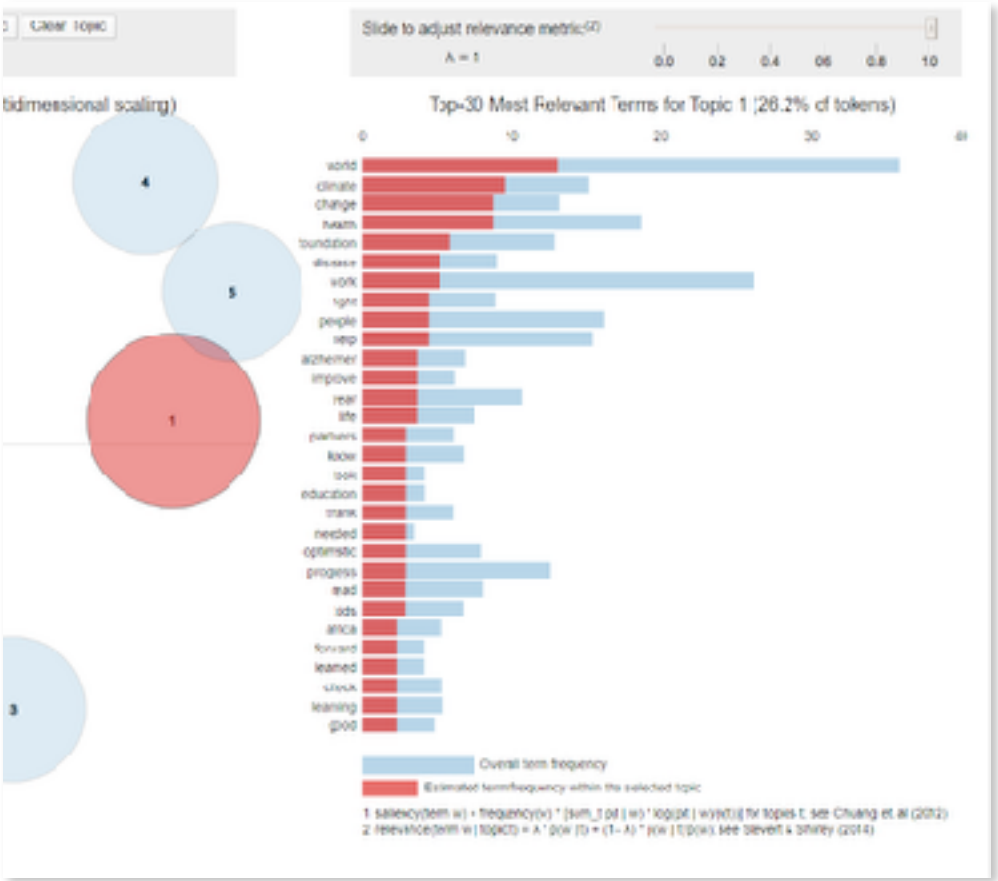

Figure 14. Exemple de topic sélectionner (ici topic 1)

III.Analyse des données

1. Analyse en Composantes Principales

L'analyse en composante principale est une méthode permettant de réduire le nombre de variable et de rendre l'information moins redondante.

Elle consiste à transformer des variables corrélées en nouvelles variables decorrélées les unes des autres. Ces nouvelles variables sont nommées 'composantes principales'.

Elle vise à construire un modèle permettant de prédire/expliquer les valeurs prises par une variable cible qualitative à partir d'un ensemble de variables explicatives quantitatives ou qualitatives.

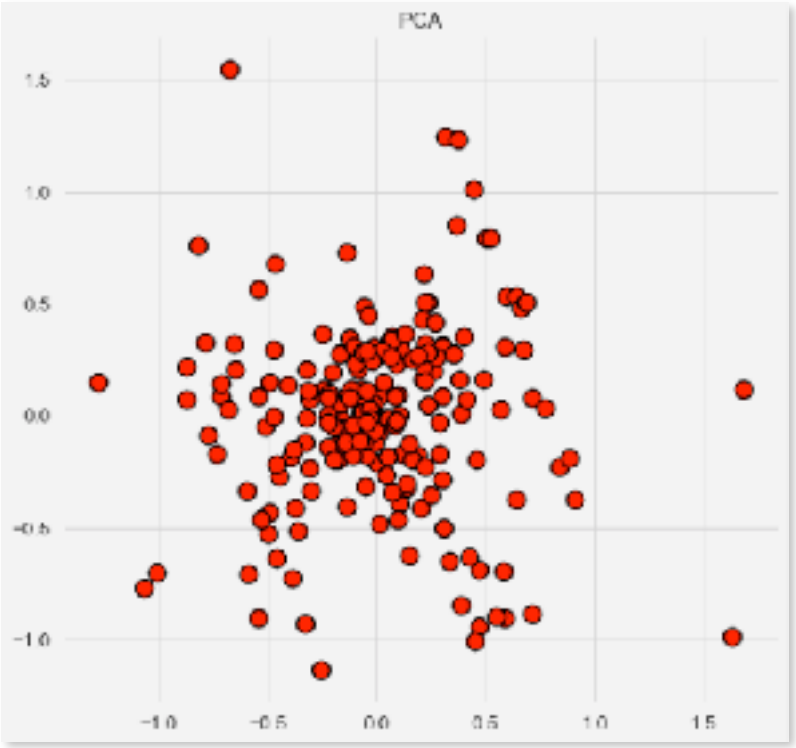

Figure 15. Représentation des individus sur le plan factoriel.

2. K-Means

Le partitionnement en K-moyennes (ou k-means en anglais) est une méthode de partitionnement de données et un problème d'optimisation combinatoire.

Étant donnés des points et un entier *k*, le problème est de diviser les points en *k* groupes, souvent appelés *clusters*, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster, la fonction à minimiser est la somme des carrés de ces distances.

Une question qui doit se poser est quelle valeur doit prendre k, pour répondre à cette question, on utilise l**e coefficient de silhouette** comme mesure de la qualité d'une partition.

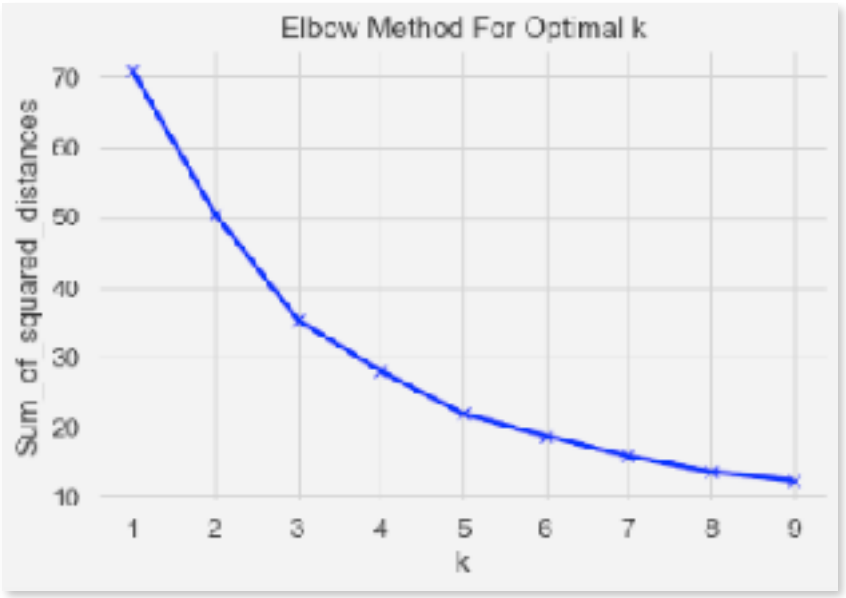

Figure 16. Expression du coefficient de silhouette en fonction du nombre de cluster

La partition en k = 3 clusters, semble être la meilleure au sens de la métrique 'silhouette', ce qui est cohérent avec notre étude des tweets positifs, neutres et négatifs.

Figure 17. Application de l'algorithme K-Means avec k=3.

3. Classification Ascendante Hiérarchique

La classification ascendante hiérarchique est une méthode de classification automatique, à partir d'un ensemble de *n* individus, son but est de répartir ces individus dans un certain nombre de classes.

Initialement, chaque individu forme une classe, soit *n* classes. On cherche à réduire le nombre de classes, ceci se fait itérativement. À chaque étape, on fusionne deux classes, réduisant ainsi le nombre de classes. Les deux classes choisies pour être fusionnées sont celles qui sont les plus « proches », en d'autres termes, celles dont la dissimilarité entre elles est minimale, cette valeur de dissimilarité est appelée *indice d'agrégation*. Comme on rassemble d'abord les individus les plus proches, la première itération a un indice d'agrégation faible, mais celui-ci va croître d'itération en itération.

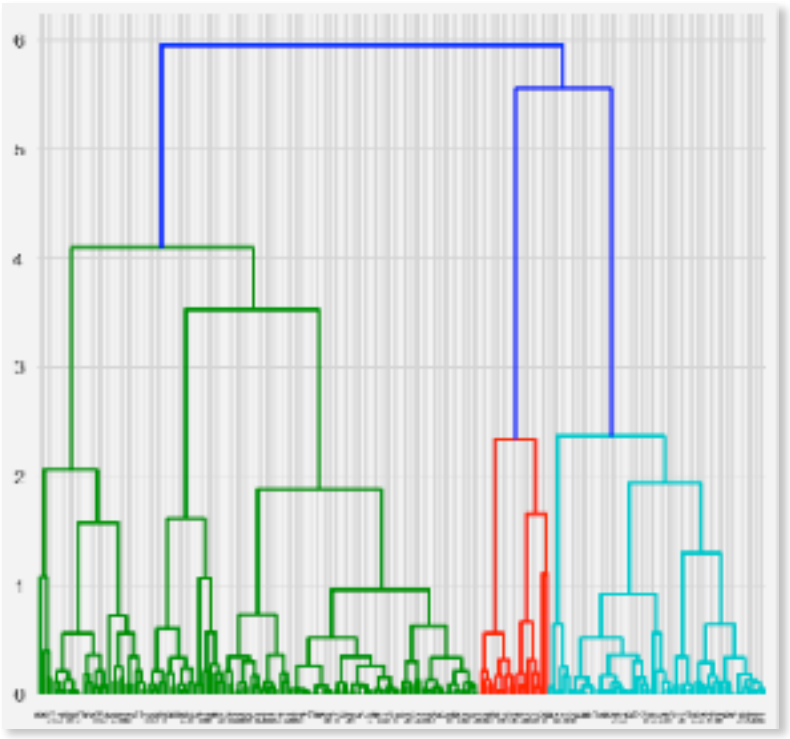

Figure 18. Application de l'algorithme de CAH

Le dendogramme suggère un découpage en 3 classes. On obtient les mêmes résultats de K-Means.

4. DBScan

DBSCAN (density-based spatial clustering of applications with noise) est un algorithme de partitionnement de données fondé sur la densité dans la mesure qui s'appuie sur la densité estimée des clusters pour effectuer le partitionnement.

L'algorithme DBSCAN utilise 2 paramètres : la distance et le nombre minimum de points devant se trouver dans un rayon pour que ces points soient considérés comme un cluster.

Les paramètres d'entrées sont donc une estimation de la densité de points des clusters. L'idée de base de l'algorithme est ensuite, pour un point donné, de récupérer son voisinage et de vérifier qu'il contient bien le nombre minimum de points le nombre en voisinage ou plus.

Ce point est alors considéré comme faisant partie d'un cluster. On parcourt ensuite l'autre voisinage de proche en proche afin de trouver l'ensemble des points du cluster.

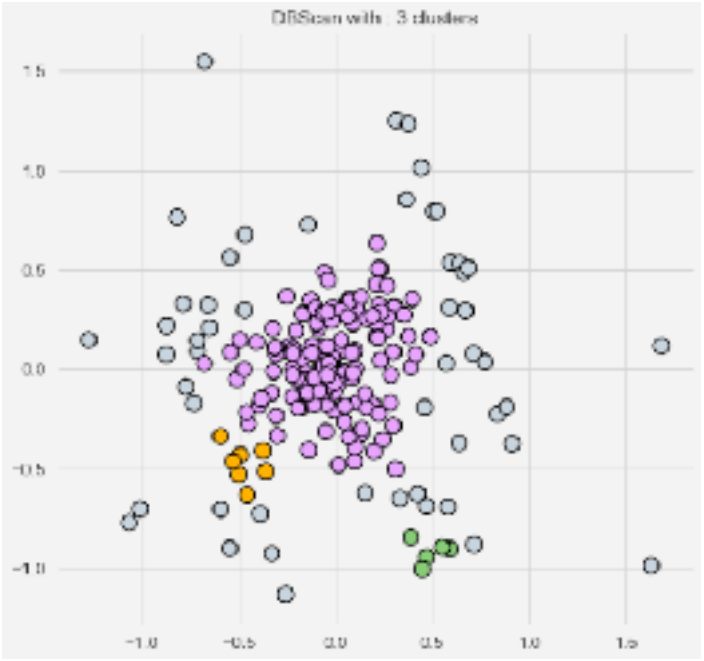

Figure 19. Application de l'algorithme DBScan

En ce qui concerne le paramétrage de l'algorithme, en modifiant Epsilon et le nombre minimum de point pour former un cluster nous obtenons 3 clusters colorés distinguable sur la figure. Et les points en gris sont des outliers.

IV.Conclusion et Perspective

Les données extraites de commentaires ou de posts sur les réseaux sociaux peuvent être révélateurs mais celles-ci parviennent souvent de manière non labélisée.

Dans le cadre de ce projet notre regard s'est porté sur l'analyse de sentiment à travers les tweets d'un utilisateur ou d'une thématique recherchée.

Après le pré-traitrement, le texte brut est converti en valeurs numériques qui permettent d'obtenir les différents topics dégagés mais aussi les catégories de sentiments obtenus à partir de leur polarité et de leur subjectivité.

Les 3 différentes catégories obtenues sont alors : positif, négatif et neutre. Elles furent donc utilisées en tant que labels pour tester divers algorithmes de clustering.

Ainsi, 3 différents clusters sont dégagés avec une précision de 78% pour le K-Means, 3 clusters agrégés sur le CAH et 3 clusters avec 50 outliers pour DBScan (sur un dataset de 200 tweets).

Parmi les différentes perspectives possible pour ce projet, nous pouvons citer différentes technique de Machine Learning comme la méthode des **facteurs naturel d'hydratation (NMF).**

De plus, K-Means est base sur la distance euclidienne, il serait intéressant d'étudier l'angle entre les individus et donc d'utilisé **spherical K-Means.**