Projet d'Analyse de Sentiment des Tweets sur les JO 2024 :

Exploration des Opinions des Utilisateurs de Twitter

1. Introduction

Ce projet vise à analyser le sentiment des utilisateurs de Twitter concernant les Jeux Olympiques de 2024 (JO 2024).
Twitter est une plateforme où les individus expriment leurs opinions et partagent leurs réactions en temps réel. En exploitant ces données, j'ai souhaité explorer les sentiments et les opinions des utilisateurs de Twitter à l'égard des JO 2024, afin de répondre à une question :

Quel est le sentiment des français au sujet des Jeux Olympique 2024 à Paris ?

L'objectif principal de ce projet est de comprendre les sentiments généraux et les opinions des utilisateurs de Twitter concernant les JO 2024. Nous cherchons à identifier les tendances positives, négatives et neutres associées à cet événement sportif mondial. Cela nous permettra de mieux appréhender l'opinion publique et les réactions des internautes vis-à-vis des JO 2024.

2.Méthodologie :

A. Création du dataframe

Les données à utiliser sont les tweets des utilisateurs de Twitter qui concernent le sujet des Jeux Olympiques.
Pour extraire ces données, il existe plusieurs possibilités.

- **L'extraction des données selon les hashtags identifiés**

Depuis février 2023, l'API "Twitter API v2" a mis en place de nombreuses restrictions. L'accès gratuit permet de rechercher et extraire 1 500 tweets par mois. Ces données ne seront pas suffisantes pour la recherche voulue.

De ce fait, j'ai utilisé snscrape, une librairie de python permettant la recherche et la récupération de tweets sans limitation. J'ai souhaité récupérer les tweets à partir de la date d'annonce des JO de Paris, le 13 septembre 2017, jusqu'au 05 juin 2023.

Les hashtags identifiés étaient les suivants :

```
#JO2024 - #JeuxOlympiques - #jeuxolympiques - #PARIS2024 - #Paris2024 - #JO - #jo - jeux olympiques
```

Les variables extraites sont les suivantes :

```
TweetURL
User
Verified
Source
Location
Language
Tweet
Likes_Count
Retweet_Count
Quote_Count
Reply_Count
```

J'ai obtenu un dataset de 53202 entrées et 12 colonnes, sauvegardé dans un fichier csv " Tweets_df".

- **Nettoyage du dataset**

Plusieurs étapes ont été mise en oeuvre pour obtenir un dataset pertinents :

- Gestion des valeurs manquantes :

```
Observation des valeurs manquantes :
Date                 0
TweetURL             1
User                 1
Verified             1
Source           53202
Location         19356
Language             4
Tweet                4
Likes_Count          7
Retweet_Count        7
Quote_Count          7
Reply_Count          7
dtype: int64
```

Les variables “Source” et “Location” contiennent de nombreuses valeurs manquantes.

La variable “Source” qui contient uniquement des valeurs manquantes sera supprimée.

De même pour les variables inutiles : "TweetURL", "Language".

La variable “ Location” se verra appliquer plusieurs traitements afin de pouvoir utiliser les informations de localisation des tweets ultérieurement.

Ainsi:

- les emojis (de nombreux drapeaux) sont supprimés

- les valeurs manquantes sont remplacées par “No Location”

- Les localisations comportant les valeurs “ Paris” ou “75” sont remplacées par “Paris, France”

- Les localisations comportant les noms de grandes villes de France sont remplacées par “France”

- Les localisations n’ayant ni Paris, ni 75, ni France, ni une grande ville de France dans leur valeur est remplacée par “Other”.

Les valeurs manquantes restantes sont supprimées. Le dataframe ne contient plus de valeurs manquantes, il est composé de 53198 entrée et 8 variables.

- **Autres manipulations**

La variable "Date" comme son nom l'indique contient des dates et heures. Il faut donc convertir son type en " Datetime".

A ce stade, nous constatons que l'extraction n'a pas permis de collecter les tweets antérieurs au 13 avril 2022. Cela est un manque à gagner car il ne sera pas possible d'analyser les tweets au moment de l'annonce des Jeux olympiques à Paris, en 2017.

Récupération des Hashtags

Afin de pouvoir analyser les hashtags les plus utilisés et leur nombre, il a fallu isoler ces hashtags de la variable “ Tweets”.

Ainsi, une variable “ Hashtags” a été ajoutée au dataframe, contenant les Hashtags utilisés dans chaque tweets.

Les valeurs vides sont remplacés par “ No Hashtag used”

Un second dataframe a été créé afin d’y stocker chaque hashtags différents et le nombre de fois qu’ils ont été utilisés.

B. Analyse de sentiments

L'analyse de sentiment est le processus d'évaluation des opinions, des émotions et des attitudes exprimées dans un texte pour déterminer si elles sont positives, négatives ou neutres. Elle est utilisée pour extraire des informations significatives à partir de grandes quantités de données textuelles, permettant de comprendre les sentiments du public, d'évaluer la réputation d'une marque, de détecter les tendances émergentes et d'orienter les décisions commerciales.

- Pré-processing

Comme toute technique de machine learning, l'analyse de sentiments se fait sur des données préparées en amont.

Le pré-processing a été mené à partir de la variable "Tweets".

Les actions suivantes ont été appliquées grâce à une fonction :

- Suppressions des caractères spéciaux
- Suppressions de la ponctuation
- Suppression des emoji
- Tokenisation et lemmatisation
- Suppression des chiffres
- Suppression des stop words
- Suppression des liens

Ainsi, les tweets ayant subi ce traitement ont été stocké dans une variable “ data_clean”

- Analyse avec Textblob

Dans ce projet, l'analyse de sentiment a été faite grâce au machine learning : le langage naturel (NLP) a été utilisé pour évaluer les tweets relatifs aux JO 2024, via la librairie Textblob.

La variable “data_clean” a été soumise à une pipeline de traitement NLP, qui extrait la polarité du sentiment de chaque tweet et établit le sentiment global (positif, négatif ou neutre) associé à chaque tweet.

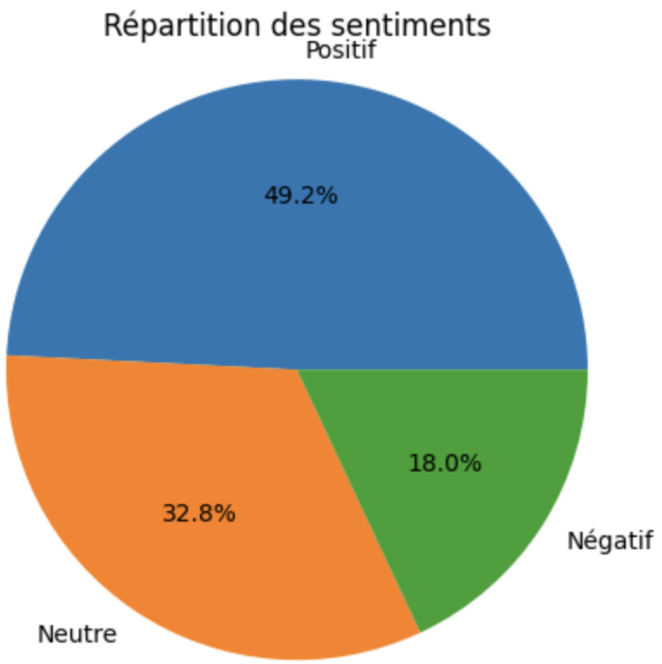
Ainsi :

Si la polarité est supérieure à 0, le sentiment est positif.

Si la polarité est inférieur à 0, le sentiment est négatif.

Les autres valeurs de polarité donnent un sentiment neutre.

La distribution de la variable ‘sentiment’ est la suivante :



C. Word Cloud

La méthodologie du wordcloud consiste à représenter graphiquement les mots les plus fréquents dans un texte. Cette visualisation permet d'identifier rapidement les termes les plus saillants et les plus utilisés dans un texte.

Le wordcloud permet de saisir rapidement les thèmes principaux, les préoccupations ou les tendances dominantes dans un corpus.

L'analyse des résultats du wordcloud se fait en examinant les mots les plus fréquents et leur contexte. Cela peut aider à comprendre les sujets qui reviennent le plus souvent, les opinions ou les émotions prédominantes, ou encore à détecter des tendances ou des mots clés pertinents.

Ici, nous constatons que les mots les plus utilisés sont à première vue “ Jeux olympiques”, “monde”, “paris”, “paralympique”.

Notons que, bien qu'il fournisse une représentation visuelle intéressante, le worcloud ne fournit pas de mesure quantitative précise. Il est important de l'interpréter avec prudence.

3. Visualisation sur Power BI

Après avoir recueilli les données puis les avoir modélisées, nous pouvons désormais les visualiser.

Pour cela, j'ai choisi Power BI, un outil puissant et populaire pour la visualisation des données.

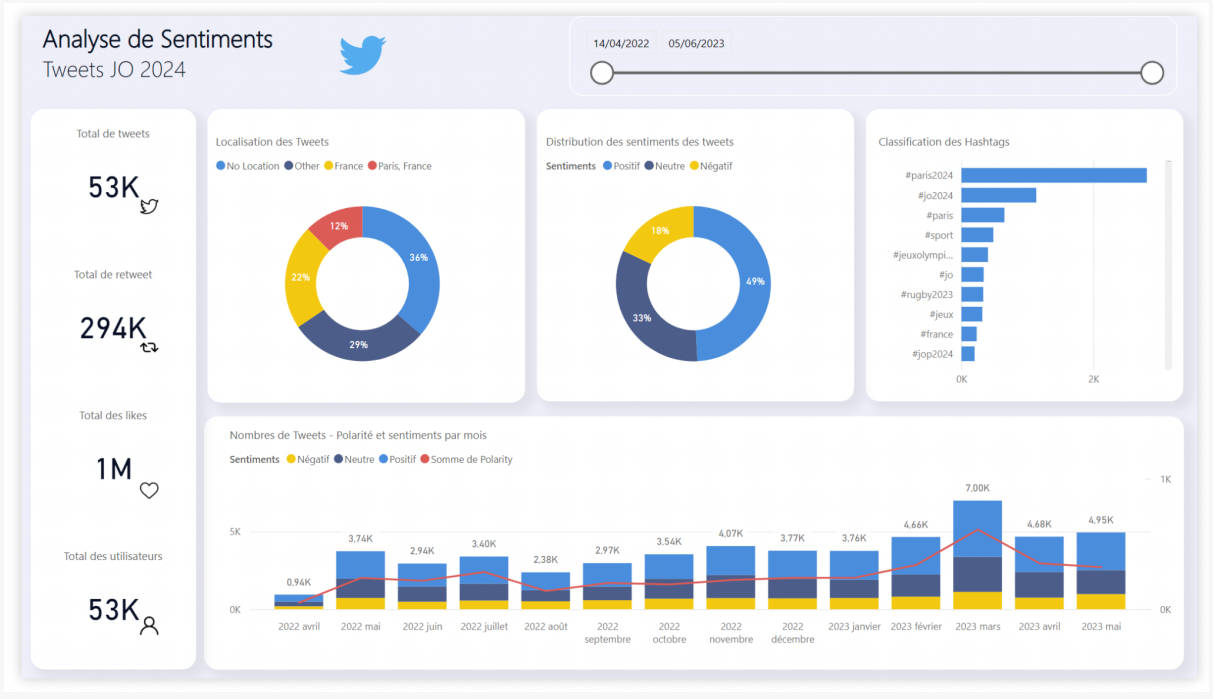
Dans un premier temps, j'ai importé le dataframe incluant l'analyse de sentiment précédemment enregistré, puis le dataframe contenant les hashtags et leurs comptes.

Quelques manipulations ont été faites sur Power Query, notamment changer les types des variables, et remplacer les "." par des "," dans les valeurs décimales, afin de changer le type en nombre décimale.

Les KPI's souhaitez pour répondre à notre problématique étaient les suivantes :

- Nombre total de tweets
- Nombre total de likes
- Nombre total de retweets
- Nombre d'utilisateurs
- Localisation des tweets
- Distribution des sentiments des tweets
- Classification des Hshtags
- Nombre de tweets par mois - Evolution de la polirité et compte des sentiments

Voici le dashboard réalisé :



Ce dashboard est interactif et il est possible de trier les informations par période choisie.

4. Résultats

Hypothèse à vérifier :

De mon point de vue, les parisiens étaient mécontents de l'arrivée des JO et de ce qui en découle : recrutement de bénévoles de masse, prix des entrées jugées exorbitants, augmentation des prix des locations courtes durées, travaux interminables en petite couronne .
J'ai souhaité obtenir les sentiments des utilisateurs de Twitter concernant les JO 2024 afin de confirmer ou non mon hypothèse.
Les résultats nous permettent de visualiser les tendances et les schémas émotionnels à travers des graphiques et des analyses statistiques.

Insights :

Il se trouve que

- Les sentiments des utilisateurs de Twitter sont en majorité (49%) positifs à l'égard des JO 2024. 18% des tweets sont négatifs et 33% sont neutres.
- La plupart des tweets n'ont pas de localisation, mais 12% proviennent de Paris, 22% proviennent de France et 29% d'autres localisations

- 17% des tweets sont positifs et proviennent de France, 15% sont positifs et proviennent d’autres localisations
- 5% des tweets sont négatifs et proviennent de France, 2%sont positifs et proviennent d’autres localisations
- Suites aux annonces récentes (prix, recrutement de bénévoles) entre mars et mai 2023, ces actualités ont fait augmenter le nombre de tweets au sujet des JO, néanmoins, malgré une augmentation des tweets au sentiments négatifs à ce moment-là, les sentiments positifs restent majoritaires.

5. Conclusion :

Les résultats de ce projet indiquent un accueil favorable de la population à l'égard des Jeux Olympiques 2024. Avec une majorité de sentiments positifs exprimés dans les tweets analysés, il semble que les utilisateurs de Twitter aient une attitude positive et enthousiaste envers cet événement sportif majeur à venir. Cela suggère un fort intérêt et un soutien de la part du public.

Ces conclusions sont renforcées par le fait que la majorité des tweets sont neutres, ce qui indique que de nombreux utilisateurs ont adopté une attitude objective et impartiale dans leurs discussions sur les Jeux Olympiques. De plus, malgré la présence de certains tweets négatifs, leur proportion reste relativement faible par rapport aux sentiments positifs et neutres exprimés.

Il est intéressant de noter que la localisation des tweets varie, avec un pourcentage significatif provenant de Paris, ainsi que d'autres régions de la France. Cela témoigne de l'intérêt national pour les Jeux Olympiques 2024 et de l'engagement de différentes régions dans l'événement.

Les résultats de cette analyse de sentiment peuvent être utiles pour les organisateurs des JO 2024, les médias, les sponsors et d'autres parties prenantes intéressées par l'opinion publique sur cet événement. Ces informations peuvent les aider à ajuster leurs stratégies de communication, à répondre aux préoccupations du public et à mieux comprendre les réactions des utilisateurs de Twitter vis-à-vis des JO 2024.