Université Internationale de Casablanca

UNIVERSITÉ RECONNUE PAR L'ÉTAT

Rapport de Projet

Sous le thème

Analyse de Tweets secteur Telecom Maroc à l'aide de bert

Réalisé par : Youssef Yaslane

Encadré par : M. Karim ardi

Introduction :

1. Contexte et Objectifs du Projet :

Dans un monde de plus en plus connecté, les réseaux sociaux sont devenus une source précieuse d'informations pour les entreprises. Les utilisateurs partagent leurs opinions, expériences et critiques sur divers services, y compris ceux des opérateurs télécoms. Pour les entreprises du secteur des télécommunications, ces retours en temps réel constituent une opportunité stratégique d'améliorer leurs offres, d'anticiper les crises et de mieux comprendre les attentes des clients.

Ce projet vise à **analyser les tweets des utilisateurs concernant les services télécoms**, en particulier ceux liés à l'opérateur **Inwi** et aux télécommunications marocaines.
L'objectif est de :

- **Extraire et analyser des tweets mentionnant des services télécoms.**
- **Identifier les sentiments dominants (positif, neutre, négatif) à partir des tweets.**
- **Distinguer les points de frustration des clients à travers l'analyse des tweets négatifs.**

2. Importance de l'Analyse de Sentiment dans le Secteur des Télécommunications :

Les entreprises de télécommunications sont souvent confrontées à des défis liés à :

- **La qualité du service client** (réseaux, internet, forfaits mobiles).
- **La gestion des pannes** et des interruptions de service.
- **La perception de la marque** par les utilisateurs.
- **La concurrence intense** entre opérateurs, rendant chaque retour client crucial pour améliorer l'offre.

L'analyse de sentiment permet de :

- **Anticiper les problèmes** en détectant rapidement une augmentation des tweets négatifs liés à un service.
- **Surveiller la satisfaction client** en temps réel.
- **Adapter la communication marketing** en fonction de l'évolution de l'opinion publique.
- **Réduire le taux de churn** (désabonnement) en identifiant et en traitant les problèmes majeurs avant qu'ils ne deviennent critiques.

Exemple concret :
Si plusieurs tweets signalent des pannes récurrentes dans une région spécifique, l'entreprise

peut prioriser les interventions techniques dans cette zone, améliorant ainsi la satisfaction des clients et préservant sa réputation.

3. Présentation des Outils et Technologies Utilisés :

Python est le langage principal de ce projet en raison de sa flexibilité et de la richesse de ses bibliothèques dédiées au traitement des données et au Machine Learning. Il offre une large gamme d'outils pour l'analyse de texte, la manipulation de données et l'intégration avec des API.

Pandas est une bibliothèque de manipulation et d'analyse de données, essentielle pour :

- Charger les tweets à partir de fichiers CSV.
- Nettoyer et prétraiter les données textuelles.
- Appliquer des transformations (comme extraire des sentiments pour chaque tweet).
- Créer des DataFrames pour stocker les résultats de l'analyse.

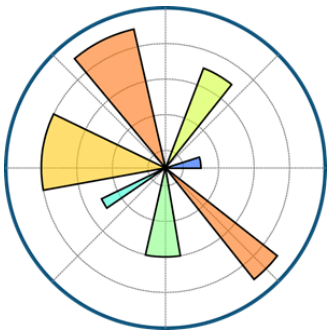
La bibliothèque **Transformers** de Hugging Face est au cœur de ce projet pour effectuer l'analyse de sentiment à l'aide de **modèles BERT pré-entraînés**. Hugging Face simplifie l'intégration des modèles de Machine Learning de pointe grâce à des **pipelines prêts à l'emploi**.

- **BERT (Bidirectional Encoder Representations from Transformers)** est un modèle de traitement du langage naturel (NLP) capable de comprendre le contexte d'un mot en fonction de son environnement dans une phrase.
- **DistilBERT** est une version plus légère de BERT, utilisée pour améliorer la vitesse d'exécution tout en conservant de bonnes performances.

MongoDB est une base de données NoSQL qui permet de **stocker les tweets analysés** directement sous forme de documents JSON. Cette approche est particulièrement utile pour manipuler de grandes quantités de données textuelles non structurées.

Pourquoi MongoDB ?

- **Flexibilité** : Permet de stocker des documents de formats variés (tweets avec sentiments).
- **Scalabilité** : Adapté à la gestion de gros volumes de données (utile pour des millions de tweets).
- **Simplicité d'intégration avec Python** grâce à la bibliothèque `pymongo`.



Matplotlib est une bibliothèque de visualisation de données largement utilisée pour créer des graphiques statiques, interactifs et publiables.

Dans ce projet, **Matplotlib** est utilisé pour :

- **Visualiser la répartition des sentiments** (graphiques à barres, camemberts).
- Créer des **histogrammes** des tweets positifs, neutres et négatifs.
- Générer des **courbes temporelles** montrant l'évolution des sentiments au fil du temps.

4. Résumé de l'Environnement Technique

- **Langage :** Python
- **Manipulation de données :** Pandas
- **Modèle NLP :** BERT (via Hugging Face Transformers)
- **Visualisation :** Matplotlib
- **Stockage des résultats :** MongoDB (optionnel)

Description des Données

1. Source des Données : Collecte de Tweets :

Les données utilisées pour ce projet proviennent de **tweets collectés via l'API Twitter**. Twitter est une plateforme de microblogging largement utilisée par les utilisateurs pour partager des expériences, donner leur avis et interagir avec des marques.

Méthode de Collecte :

- **API Twitter v2** : La collecte des tweets est réalisée grâce à l'API Twitter, en utilisant la bibliothèque tweepy (Python).
- **Filtres de recherche** : Les tweets sont collectés en fonction de mots-clés spécifiques tels que :
    - forfait Inwi : pour capturer des avis sur les forfaits de l'opérateur Inwi.
    - télécom Maroc : pour suivre les tendances générales dans le secteur des télécommunications marocaines.
- **Période de collecte** : La collecte peut être réalisée en temps réel ou sur une période définie.
- **Langue** : Les tweets sont principalement en français et en arabe, reflétant la diversité linguistique des utilisateurs marocains.

2. Description du Fichier CSV (`tweets.csv`) :

Les tweets collectés sont stockés dans un fichier au format CSV (Comma Separated Values). Ce fichier contient plusieurs colonnes représentant les caractéristiques essentielles de chaque tweet.
Ce format est simple à manipuler avec `pandas` pour effectuer des analyses de données.

```
tweets.describe()
✓ 0.0s
```

	tweet_id	nbr_likes	nbr_retweets	nbr_caracters	author_id
count	1.000000e+02	100.000000	100.000000	100.00000	1.000000e+02
mean	1.871206e+18	0.820000	0.100000	159.99000	8.697558e+17
std	5.400058e+14	1.659865	0.333333	86.96098	8.288053e+17
min	1.869654e+18	0.000000	0.000000	20.00000	4.188782e+07
25%	1.870986e+18	0.000000	0.000000	85.50000	4.213655e+08
50%	1.871129e+18	0.000000	0.000000	149.50000	1.071537e+18
75%	1.871679e+18	1.000000	0.000000	213.00000	1.716783e+18
max	1.871896e+18	8.000000	2.000000	347.00000	1.869382e+18

3. Présentation des Colonnes du CSV :

```
tweets.columns
✓ 0.0s

Index(['_id', 'tweet_id', 'tweet_date', 'nbr_likes', 'nbr_retweets',
       'nbr_caracters', 'author_id', 'text', 'comments'],
      dtype='object')
```

4. Exemple de Données (Aperçu du CSV) :

```
tweets.head(4)
✓ 0.0s                                                    Python
```

_id	tweet_id	tweet_date	nbr_likes	nbr_retweets	nbr_caracters	author_id	text	comments
6bea28f79f312fd703d50e	1871877916285706650	2024-12-25 11:17:27+00:00	0	0	230	1849839193763012608	مباريات اليوم من مؤجل الجولة 14 من البطولة الإ...	[{'comment_order': 0, 'comment_date': None, 'c...
'6bedc9f79f312fd703d50f	1871878320314634276	2024-12-25 11:19:04+00:00	0	1	139	1455257774	RT @LNFP_Officiel: مباريات اليوم من مؤجل الجول...	[{'comment_order': 0, 'comment_date': None, 'c...
'6bedfcf79f312fd703d511	1871870247172780236	2024-12-25 10:46:59+00:00	1	0	52	1342619772308115456	Xmas Botola Pro Inwi 🇲🇦🎄!!!! https://t.co/xWRQ...	[{'comment_order': 0, 'comment_date': None, 'c...
6bee16f79f312fd703d512	1871847495346815252	2024-12-25 09:16:34+00:00	0	0	70	1086724355801956352	@NdouK1997 @InnocentKhabz Nia mudivha Pelepel...	[{'comment_order': 0, 'comment_date': None, 'c...

5. Prétraitement des Données :

Avant d'effectuer l'analyse de sentiment, une étape essentielle consiste à **nettoyer et prétraiter les tweets**.
Les étapes typiques du prétraitement incluent :

- **Suppression des stopwords (mots vides)** : "le", "de", "et", etc.
- **Tokenization** : Diviser le texte en mots ou sous-mots.
- **Suppression des liens et mentions** : Elimination des @user et des URLs.
- **Suppression des répétitions** : Réduire des mots comme "looooooool" en "lol".

6. Conclusion :

Les données collectées à partir de Twitter fournissent des informations riches pour comprendre les avis des utilisateurs sur les services télécoms. Le prétraitement rigoureux et la bonne structuration du fichier CSV permettent de garantir des résultats précis lors de l'analyse de sentiment.

Environnement et Configuration :

1. Présentation de l'Environnement :

Le projet d'analyse de sentiment est développé dans un environnement **Python**, qui offre une large gamme de bibliothèques adaptées à la manipulation des données, au traitement du langage naturel (NLP) et à la visualisation.
L'environnement peut être configuré localement (sur votre machine) ou dans un notebook Jupyter (Google Colab ou JupyterLab).

2. Installation et Configuration des Bibliothèques :

Les principales bibliothèques utilisées pour ce projet sont :

- **pandas** – Manipulation des données (chargement, nettoyage et transformation).
- **transformers** – Utilisation de modèles pré-entraînés pour l'analyse de sentiment (BERT).
- **pymongo** (optionnel) – Stockage des résultats dans une base de données MongoDB.

2.1 Installation des Bibliothèques :

L'installation peut être réalisée via `pip` ou directement dans un notebook.

```
# Installer pandas pour manipuler les données
pip install pandas

# Installer transformers pour utiliser des modèles NLP (BERT)
pip install transformers

# Installer pymongo pour interagir avec MongoDB (optionnel)
pip install pymongo

# Installer matplotlib pour visualiser les résultats (facultatif mais recommandé)
pip install matplotlib
```

2.2 Importation des Bibliothèques :

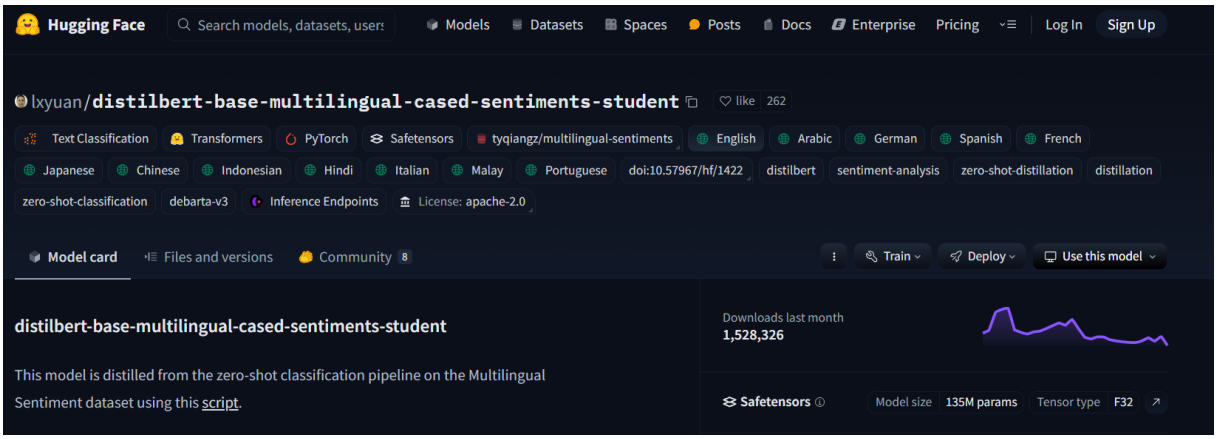
```
import pandas as pd
from transformers import pipeline
from pymongo import MongoClient
import matplotlib.pyplot as plt
```

3. Téléchargement et Installation du Modèle BERT :

Pour effectuer l'analyse de sentiment, nous utilisons le modèle **distilbert-base-multilingual-cased-sentiments-student**.
Ce modèle est pré-entraîné sur des données multilingues et est optimisé pour détecter des sentiments dans plusieurs langues, y compris le français et l'arabe.

Pourquoi ce Modèle ?

- **Multilingue** : Il peut analyser des tweets en plusieurs langues.
- **Léger et Rapide** : DistilBERT est une version plus compacte de BERT, offrant des performances similaires avec une vitesse d'exécution accrue.
- **Fine-tuned** pour l'analyse de sentiment, réduisant le besoin de pré-entraîner un modèle à partir de zéro.



3.1 Chargement Automatique du Modèle avec Transformers :

Le modèle peut être téléchargé directement depuis Hugging Face et utilisé avec un simple pipeline.

```python
from transformers import pipeline

distilled_student_sentiment_classifier = pipeline(
    model="lxyuan/distilbert-base-multilingual-cased-sentiments-stude
    return_all_scores=True
)

# english
distilled_student_sentiment_classifier ("I love this movie and i woul
>> [[{'label': 'positive', 'score': 0.9731044769287109},
  {'label': 'neutral', 'score': 0.016910076141357422},
  {'label': 'negative', 'score': 0.009985478594899178}]]

# malay
distilled_student_sentiment_classifier("Saya suka filem ini dan saya
[[{'label': 'positive', 'score': 0.9760093688964844},
  {'label': 'neutral', 'score': 0.018045164644471817},
  {'label': 'negative', 'score': 0.005945465061813593}]]

# japanese
distilled_student_sentiment_classifier("私はこの映画が大好きで、何度も見ま
>> [[{'label': 'positive', 'score': 0.9342429041862488},
  {'label': 'neutral', 'score': 0.040193185210227966},
  {'label': 'negative', 'score': 0.025563929229974747}]]
```

Architecture du Projet :

1. Schéma Global du Projet : De la Collecte à l'Analyse :

L'architecture globale du projet suit un pipeline clair et structuré, qui passe par plusieurs étapes, de la collecte des tweets jusqu'à l'analyse de sentiment et l'export des résultats. Ce pipeline est automatisé pour traiter de grands volumes de données en temps réel ou en mode batch (par lots).

Flux de Travail du Projet :

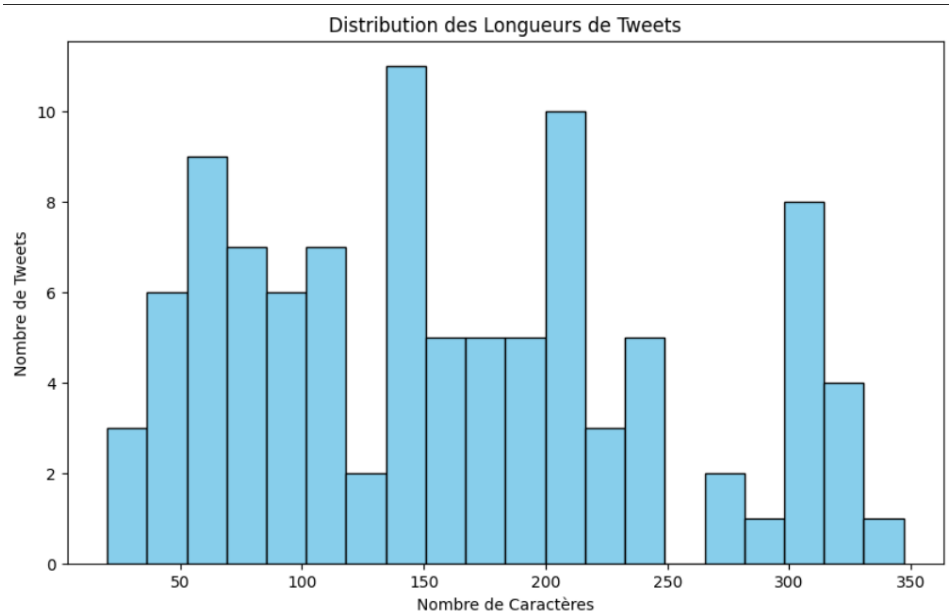
- **Collecte des tweets** via l'API Twitter (à l'aide de Tweepy).
- **Stockage des tweets** dans un fichier CSV (`tweets.csv`).

- **Chargement des données** dans un DataFrame `pandas` pour prétraitement.
- **Nettoyage des textes** pour enlever les éléments inutiles (hashtags, mentions, liens).
- **Application du modèle BERT** à chaque tweet pour obtenir un score de sentiment (positif, neutre, négatif).
- **Filtrage des tweets négatifs** afin de détecter les problèmes récurrents.

Visualisation des Données :

La visualisation des données est une étape essentielle pour interpréter les résultats de l'analyse de sentiment et identifier rapidement les tendances émergentes. Grâce aux graphiques générés, il est possible de mieux comprendre les comportements des utilisateurs, les plaintes fréquentes et les périodes où l'activité est plus intense. Cette section présente et explique plusieurs visualisations clés issues de l'analyse des tweets.

1. Histogramme de la Longueur des Tweets



Description :

Cet histogramme montre la distribution des tweets en fonction de leur longueur (nombre de caractères). L'analyse de la longueur des tweets permet d'identifier si des corrélations existent entre la longueur et le sentiment exprimé.

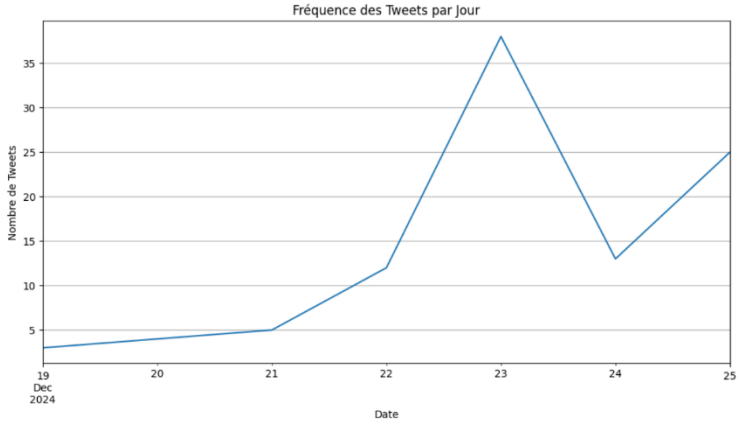
Analyse :

- La longueur des tweets semble variée, mais la plupart des tweets sont concentrés autour de 50 à 150 caractères.
- Des pics de longueur peuvent refléter des tweets plus détaillés, souvent négatifs ou expliquant des problèmes complexes.

Interprétation :

- Les tweets négatifs peuvent avoir une longueur plus importante car les utilisateurs décrivent en détail leurs plaintes.
- Les tweets positifs ou neutres sont souvent plus courts et directs.

2. Fréquence des Tweets par Jour :



Description :

Ce graphique temporel montre l'évolution du nombre de tweets par jour. Il met en lumière les périodes d'activité accrue, permettant de repérer des événements clés qui influencent les interactions des utilisateurs.

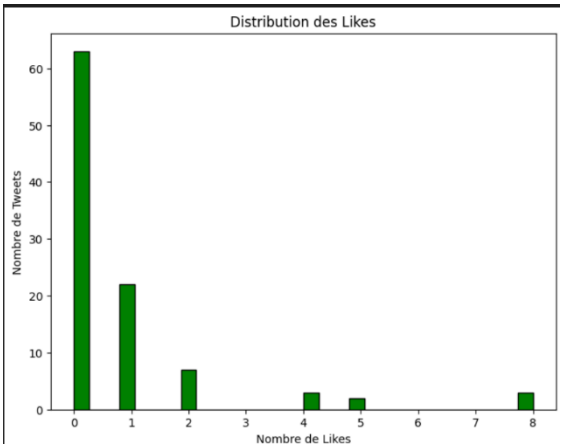
Analyse :

- Un pic de tweets est observé à certaines dates, ce qui peut indiquer des incidents ou des annonces ayant suscité des réactions en chaîne.
- Les jours où l'activité est faible peuvent refléter des périodes de stabilité sans incidents majeurs.

Interprétation :

- L'analyse temporelle est cruciale pour identifier les jours où des problèmes de service (ex : pannes de réseau) ont été fortement signalés.
- Ces pics peuvent également coïncider avec des campagnes marketing ou des lancements de nouveaux produits.

3. Distribution des Likes :



Description :

Cet histogramme montre la distribution du nombre de likes reçus par les tweets. Il aide à identifier quels tweets ont généré de l'engagement positif ou négatif.

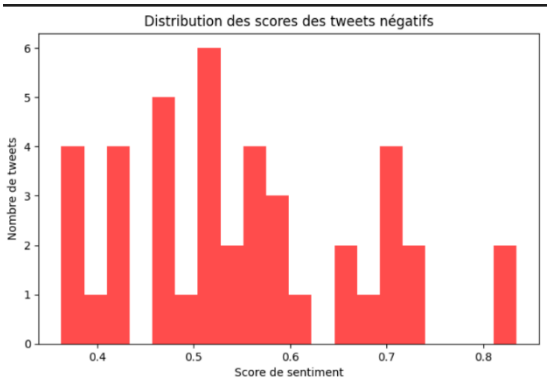
Analyse :

- La majorité des tweets reçoivent peu de likes (moins de 10).
- Certains tweets deviennent viraux avec un nombre de likes très élevé. Ces tweets peuvent contenir des critiques virulentes ou des expériences fortement partagées.

Interprétation :

- Les tweets les plus populaires (likes élevés) attirent généralement plus l'attention des autres utilisateurs.
- Les tweets négatifs avec un grand nombre de likes peuvent révéler des problèmes systémiques partagés par une large partie des utilisateurs.

4. Distribution des Scores des Tweets Négatifs :



Description :

Cet histogramme représente la distribution des scores de sentiment des tweets négatifs. Les scores indiquent la confiance du modèle BERT dans la classification des tweets comme négatifs.

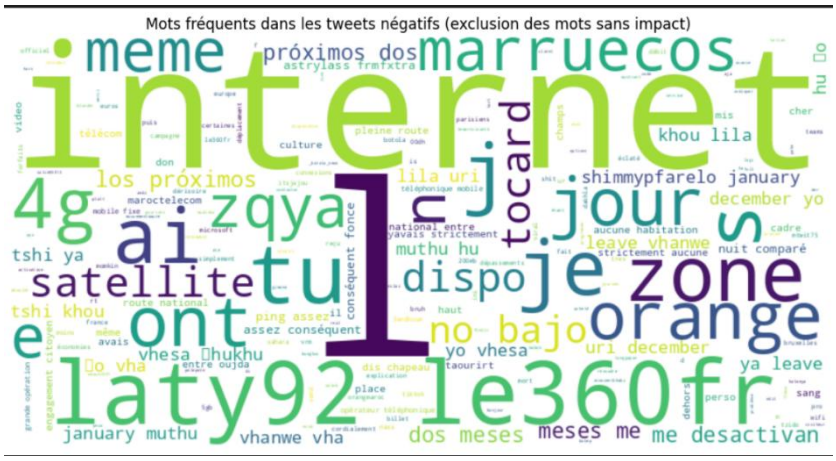
Analyse :

- La majorité des tweets négatifs présentent des scores situés entre **0,5 et 0,7**, indiquant que le modèle est relativement sûr de ses prédictions.
- Quelques tweets ont des scores très élevés (> 0,8), suggérant une forte certitude dans leur classification.

Interprétation :

- Un score élevé de tweets négatifs reflète des plaintes claires et directes, tandis que des scores proches de 0,5 indiquent une certaine ambiguïté dans le texte du tweet.
- L'analyse de ces scores permet de prioriser les tweets les plus critiques et de détecter les problèmes urgents.

5. Nuage de Mots des Tweets Négatifs :



Description :

Le nuage de mots représente les mots les plus fréquents dans les tweets négatifs. Plus un mot apparaît fréquemment, plus il est affiché en gros dans le nuage.

Analyse :

- Des termes comme **"internet"**, **"maroc"**, et **"orange"** sont fortement présents, indiquant que les problèmes de connectivité sont des préoccupations majeures.
- Des termes spécifiques comme **"panne"**, **"lent"** ou **"service"** ressortent, mettant en évidence les plaintes récurrentes.

Interprétation :

- Ce nuage de mots permet d'identifier rapidement les termes associés aux problèmes fréquents rencontrés par les utilisateurs.
- Les opérateurs télécoms peuvent s'en servir pour prioriser des actions correctives sur les problèmes les plus mentionnés.

6. Synthèse des Visualisations :

Les visualisations présentées fournissent des insights précieux sur :

- La **fréquence et la nature des tweets** (positifs, négatifs, neutres).
- Les **problèmes récurrents** identifiés grâce au nuage de mots.
- L'**évolution temporelle** des sentiments, permettant de détecter des événements spécifiques ayant généré des réactions massives.
- Les **tweets viraux** qui nécessitent une attention particulière, notamment ceux qui reçoivent beaucoup de retweets et de likes.