جامعة سيدي محمد بن عبد الله بفاس
UNIVERSITE SIDI MOHAMED BEN ABDELLAH DE FES
كلية العلوم ظهر المهراز فاس
FACULTE DES SCIENCES DHAR EL MAHRAZ FES

BDSaS

Master Big Data Analytics & Smart Systems

RAPPORT

Twitter sentiment analysis with NLP and MongoDB

Réaliser par :

ELHAGOUCHI HALIMA

ERRAZOUKI AYA

Année universitaire 2023/2024

Table de matière

INTRODUCTION

Dans le monde d'aujourd'hui, les médias sociaux comme Twitter jouent un rôle essentiel dans la façon dont les gens partagent leurs opinions et leurs sentiments sur divers sujets.

L'analyse des sentiments sur Twitter peut fournir des informations précieuses sur les tendances, les opinions du public et les réactions aux événements actuels.

Dans ce projet, nous allons explorer comment utiliser le traitement automatique du langage naturel (NLP) pour effectuer une analyse de sentiment sur les tweets, et nous stockerons les données dans une base de données MongoDB pour une gestion efficace

Solution

1 Scraping

le scraping est un élément crucial dans un projet d'analyse sémantique des tweets car il permet de collecter efficacement les données nécessaires pour l'analyse et l'extraction de insights à partir des tweets.

1.1 Definition :

Le « scraping » fait référence à l'acte d'extraire automatiquement des données d'un site web ou d'une source en ligne. Dans le contexte de l'analyse sémantique des tweets, le scraping serait utilisé pour collecter des tweets à partir de Twitter afin de les analyser par la suite. Voici pourquoi le scraping est important dans un tel projet :

- **Collecte de données en masse :** Twitter dispose d'une immense quantité de données générées par les utilisateurs, y compris des tweets sur une variété de sujets. Le scraping permet de collecter ces données en masse pour une analyse plus approfondie.

- **Analyse de données en temps réel :** En collectant les tweets en temps réel, le scraping permet une analyse de données en temps réel, ce qui peut être crucial pour suivre les tendances actuelles ou les événements en direct.

- **Personnalisation des données :** Le scraping peut être utilisé pour extraire des tweets spécifiques en fonction de critères tels que les hashtags, les mots-clés, les utilisateurs spécifiques, etc., ce qui permet une analyse plus ciblée et personnalisée.

1.2 Application

Nous allons commencer par collecter des tweets en temps réel à l'aide de ntscraper

```
1 from ntscraper import Nitter
```

```
1 scraper = Nitter()
```

```
Testing instances:   9%|████▌                                                | 7/77 [00:06
<00:52,  1.33it/s]
```

Explication :

- importe une classe appelée Nitter depuis le module ntscraper
- **Nitter** est un service qui agit comme un proxy pour Twitter, fournissant une alternative décentralisée et sans publicité à Twitter. Par conséquent, la classe Nitter pourrait implémenter des méthodes pour interagir avec le service Nitter, telles que la récupération de tweets ou d'autres données à partir de profils Twitter.

```
1 tweets = scraper.get_tweets("elonmusk", mode = 'user', number=100)
```

```
02-Apr-24 14:38:02 - No instance specified, using random instance https://nitter.privacydev.net
02-Apr-24 14:38:07 - Current stats for elonmusk: 20 tweets, 0 threads...
02-Apr-24 14:38:12 - Current stats for elonmusk: 40 tweets, 0 threads...
02-Apr-24 14:38:16 - Current stats for elonmusk: 60 tweets, 0 threads...
02-Apr-24 14:38:20 - Current stats for elonmusk: 80 tweets, 0 threads...
02-Apr-24 14:38:24 - Current stats for elonmusk: 100 tweets, 0 threads...
```

```
1 tweets
```


```
{'tweets': [{'link': 'https://twitter.com/elonmusk/status/1774234110128468337#m',
   'text': 'At 5000 tons, Starship is the largest flying object ever made. Thrust is more than doub
le the Saturn V moon rocket.  It is the first spaceship design capable of making life multiplanetar
y.  Goal of the next mission is to make it through the meteorically extreme heat of reentry.',
   'user': {'name': 'Elon Musk',
    'username': '@elonmusk',
    'profile_id': '1683325380441128960',
    'avatar': 'https://pbs.twimg.com/profile_images/1683325380441128960/yRsRRjGO_bigger.jpg'},
```


Explication :

- **Appel à la méthode get_tweets() :** La méthode get_tweets() est appelée sur l'instance scraper, avec certains paramètres spécifiés entre parenthèses. Ces paramètres contrôlent les détails de la recherche des tweets.
- **Paramètres de la méthode get_tweets() :** "elonmusk" : C'est le nom d'utilisateur Twitter pour lequel vous souhaitez récupérer les tweets. Dans cet exemple, les tweets de l'utilisateur elonmusk seront récupérés. mode = 'user' : Ce paramètre spécifie le mode de recherche. Dans ce cas, il est défini sur 'user', ce qui signifie que nous récupérerons les tweets de l'utilisateur spécifié. number = 100 : Ce

paramètre spécifie le nombre de tweets à récupérer. Dans cet exemple, 100 tweets seront récupérés.

```
final_tweets = []
for x in tweets['tweets']:
    data = [x['link'], x['text'],x['date'],x['stats']['likes'],x['stats']['comments']]
    final_tweets.append(data)
```

```
dat = pd.DataFrame(final_tweets, columns =['twitter_link','text','date','likes','comments'])
```

```
dat
```

	twitter_link	text	date	likes	comments
0	https://twitter.com/TheRealMarroqui/status/177...	Photo d'un artisan marocain pendant la constru...	Apr 1, 2024 · 2:22 PM UTC	23	2
1	https://twitter.com/mimi_khouryy/status/177483...	Le plus beau drapeau au monde ♥ #Morocco #Ma...	Apr 1, 2024 · 4:22 PM UTC	52	5
2	https://twitter.com/MuslimSpot/status/17748498...	Stop these gender wars and fear Allah… #Musl...	Apr 1, 2024 · 5:22 PM UTC	0	0
3	https://twitter.com/MuslimSpot/status/17748498...	What is up with Feminism and Redpill in Ramada...	Apr 1, 2024 · 5:22 PM UTC	0	0

Eplication :

- **Initialisation de la liste final_tweets :** Une liste vide nommée final_tweets est créée. Cette liste sera utilisée pour stocker les données extraites des tweets.
- **Boucle à travers les tweets :** Une boucle for itère à travers chaque élément de la liste de tweets stockée dans la clé 'tweets' de la variable tweets. À chaque itération, la variable x représente un dictionnaire contenant les données d'un tweet individuel.
- **Extraction des données :** Pour chaque tweet, les données pertinentes sont extraites du dictionnaire x et stockées dans une liste appelée data. Les données extraites comprennent le lien vers le tweet, le texte du tweet, la date, le nombre de likes et le nombre de commentaires.
- **Ajout des données à la liste final_tweets :** Une fois que toutes les données pertinentes ont été extraites pour un tweet donné, la liste data est ajoutée à la liste final_tweets. À la fin de la boucle, la liste final_tweets contiendra toutes les données extraites de chaque tweet.

```
dat.to_csv("elonmusk_data.csv")
```

	A	B	C	D	E	F	G
1		twitter_link	text	date	likes	comments	
2	0	https://twitt	At 5000 tons,	Mar 31, 2024	125588	14045	
3	1	https://twitt	ðŸ˜‚,	Apr 1, 2024 Â	106748	3746	
4	2	https://twitt	Whoa, this d	Apr 1, 2024 Â	118363	8003	
5	3	https://twitt	Congratulati	Apr 1, 2024 Â	4349	342	
6	4	https://twitt	So many Apr	Apr 1, 2024 Â	147960	8543	
7	5	https://twitt	Starlink help	Apr 1, 2024 Â	36534	4841	
8	6	https://twitt	Excited to joi	Apr 1, 2024 Â	238745	15342	
9	7	https://twitt	ðŸ’¯	Apr 1, 2024 Â	69421	4427	
10	8	https://twitt	When you lo	Mar 30, 2024	23420	1407	
11	9	https://twitt	~3.5 hours la	Mar 31, 2024	13050	1090	
12	10	https://twitt	Liftoff of Falc	Mar 31, 2024	17724	1258	
13	11	https://twitt	SpaceX aimir	Mar 30, 2024	103483	7517	
14	12	https://twitt	It gets partic	Mar 30, 2024	7675	1112	
15	13	https://twitt	Everyone car	Mar 30, 2024	8451	1162	
16	14	https://twitt	The 2024 No	Mar 30, 2024	7650	620	
17	15	https://twitt	History is cor	Mar 30, 2024	37215	6446	
18	16	https://twitt	Oh hi youtub	Mar 30, 2024	10348	1625	
19	17	https://twitt	Yup	Mar 30, 2024	250549	11459	
20	18	https://twitt	Only 379,000	Mar 30, 2024	228051	18899	
21	19	https://twitt	Your subscri	Mar 29, 2024	5225	858	

Explication :

- enregistrer les données extraites dans un fichier CSV nommé "elonmusk_data.csv".

2 MongoDB

MongoDB est une solution de base de données flexible et évolutive qui peut être utilisée efficacement dans un projet d'analyse de tweets pour stocker et gérer les données extraites des tweets de manière efficace et flexible

2.1 Définition :

MongoDB est une base de données NoSQL, c'est-à-dire une base de données non relationnelle, conçue pour stocker et gérer des données de manière flexible et

évolutive. Contrairement aux bases de données relationnelles traditionnelles, MongoDB utilise un modèle de données flexible basé sur des documents au format JSON (JavaScript Object Notation).

Dans le contexte d'un projet d'analyse de tweets, MongoDB pourrait être utilisé pour stocker les données extraites des tweets pour plusieurs raisons :

- **Structure de données flexible :** Les tweets peuvent varier en longueur et en structure, et MongoDB permet de stocker des données de manière flexible sans avoir besoin d'un schéma prédéfini rigide, ce qui facilite la gestion des données non structurées.

- **Gestion des données non structurées :** Les tweets peuvent contenir une grande variété d'informations telles que du texte, des images, des liens, des données de localisation, etc. MongoDB peut stocker ces données de manière efficace sans imposer de structure fixe.

- **Évolutivité :** MongoDB est conçu pour être hautement évolutif, ce qui signifie qu'il peut gérer de grandes quantités de données et s'adapter à mesure que le volume de données augmente, ce qui est essentiel dans un projet d'analyse de tweets où les données peuvent être massives.

- **Requêtes complexes :** MongoDB offre des fonctionnalités puissantes de requête et d'agrégation qui permettent d'interroger et d'analyser les données de manière flexible, ce qui est important pour extraire des insights à partir des tweets.

- **Intégration avec d'autres technologies :** MongoDB s'intègre bien avec d'autres technologies couramment utilisées dans les projets d'analyse de données, telles que Python, R et les outils d'analyse de données comme Jupyter Notebook, ce qui en fait un choix populaire pour le stockage de données dans de tels projets.

2.2 Application

```
from pymongo import MongoClient
cluster=MongoClient("mongodb+srv://halima:halima@cluster0.wjt9qvc.mongodb.net/?retryWrites=true&w=
db=cluster['elonmusk']
collection=db['elonmusk']

collection.insert_one(tweets)
```

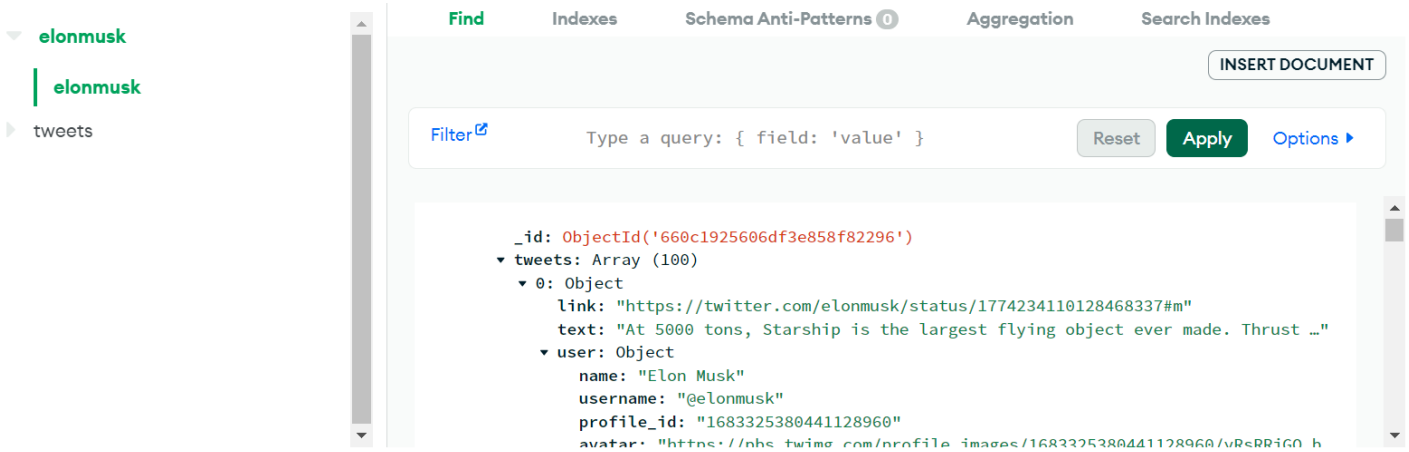
```
C:\Users\LENOVO\anaconda3\lib\site-packages\cryptography\x509\base.py:521: CryptographyDeprecationWar
ning: Parsed a negative serial number, which is disallowed by RFC 5280.
  return rust_x509.load_der_x509_certificate(data)
```

```
InsertOneResult(ObjectId('660c1925606df3e858f82296'), acknowledged=True)
```

connecte à une base de données MongoDB à l'aide de MongoClient à partir du package pymongo

Explication :

- **Importation de MongoClient :** Le code importe la classe MongoClient du module pymongo. Cette classe permet d'établir une connexion avec une instance MongoDB et d'effectuer des opérations sur la base de données.
- **Création d'une instance de MongoClient :** Une instance de MongoClient est créée en passant l'URI de connexion MongoDB comme argument. Cet URI contient des informations sur l'hôte MongoDB, les informations d'identification (nom d'utilisateur et mot de passe) et d'autres paramètres de connexion.
- **Sélection de la base de données et de la collection :** À partir de l'instance de MongoClient, une base de données nommée 'elonmusk' est sélectionnée, puis une collection également nommée 'elonmusk' est sélectionnée à l'intérieur de cette base de données. Si la base de données ou la collection n'existe pas, MongoDB les crée automatiquement lors de leur première utilisation.
- **Insertion des données :** La méthode insert_one() est appelée sur la collection sélectionnée pour insérer un document MongoDB. Dans ce cas, le dictionnaire tweets est inséré en tant que document dans la collection 'elonmusk'.



Voici un exemple de tweets :

```
1 tweets['tweets'][20]['text']
2
```

```
'I see a lot of people complaining about spam in their replies etc.  I’ve had my settings for notifi
cations set up like this for over a year now and it is SO GOOD.  Here’s how you do it:  Settings >>N
otifications >> Filters >> Quality Filter ✅ >> Muted Notifications ✅ everything.  Perhaps X shou
ld add an option to mute those who don’t subscribe to Premium.'
```

3 NLP

NLP est un outil essentiel dans un projet d'analyse sémantique des tweets car il permet de comprendre et d'analyser le langage naturel utilisé dans les tweets, ce qui permet d'extraire des insights significatifs à partir des données textuelles.

3.1 Définition :

NLP (Natural Language Processing) est une branche de l'intelligence artificielle qui se concentre sur la compréhension et la manipulation du langage humain par des ordinateurs. Dans le contexte de l'analyse sémantique des tweets, l'utilisation de NLP est cruciale pour plusieurs raisons :

- **Analyse du sentiment :** L'une des applications les plus courantes de NLP dans l'analyse de tweets est l'analyse du sentiment, qui consiste à déterminer si un tweet est positif, négatif ou neutre. Cela permet de mesurer le sentiment général autour d'un sujet ou d'un événement.

- **Extraction d'entités :** NLP peut être utilisé pour extraire des entités telles que des noms de personnes, des lieux, des organisations, des événements, etc., à partir des tweets. Cela permet d'identifier les sujets clés abordés dans les tweets et de les analyser plus en détail.

- **Classification de texte :** NLP peut être utilisé pour classer les tweets dans des catégories prédéfinies, par exemple en identifiant les tweets qui parlent de politique, de sport, de divertissement, etc. Cela permet d'organiser les tweets pour une analyse plus approfondie.

- **Extraction de thèmes et de sujets** : NLP peut être utilisé pour extraire les thèmes et les sujets principaux abordés dans les tweets, ce qui permet d'identifier les tendances et les sujets chauds sur les réseaux sociaux.

- **Correction orthographique et normalisation du texte :** NLP peut être utilisé pour corriger les fautes d'orthographe, normaliser le texte (par exemple, en convertissant le texte en minuscules) et traiter d'autres aspects du texte qui peuvent affecter la qualité de l'analyse.

3.2 Application

Ensuite, nous utiliserons des techniques de traitement automatique du langage naturel (NLP) pour prétraiter les tweets,

et appliquerons des modèles d'apprentissage automatique, tels que les classificateurs de sentiment, pour attribuer une polarité (positif, négatif ou neutre) à chaque tweet.

```
1 from transformers import AutoTokenizer, AutoModelForSequenceClassification
2 from scipy.special import softmax
3
```

Explication :

- **Importation de AutoTokenizer et AutoModelForSequenceClassification :** Ces classes sont utilisées pour charger des modèles pré-entraînés et des tokenizers à partir de Hugging Face Transformers. AutoTokenizer est utilisé pour charger le tokenizer adapté au modèle que vous souhaitez utiliser, tandis que AutoModelForSequenceClassification est utilisé pour charger le modèle pré-entraîné adapté à la classification de séquences.
- **Importation de softmax :** Cette fonction est importée à partir du module scipy.special pour calculer les valeurs softmax. La fonction softmax est utilisée pour normaliser les scores de sortie du modèle afin d'obtenir des probabilités.

```
# precprcess tweet
tweet_words = []

for word in tweet.split(' '):
    if word.startswith('@') and len(word) > 1:
        word = '@user'

    elif word.startswith('http'):
        word = "http"
    tweet_words.append(word)

tweet_proc = " ".join(tweet_words)
```

Expication :

- **Initialisation d'une liste vide tweet_words :** Une liste vide est créée pour stocker les mots du tweet après prétraitement. Boucle à travers les mots du tweet : Le code itère à travers chaque mot du tweet, qui est divisé en mots en utilisant l'espace comme délimiteur avec split(' ').
- **Remplacement des noms d'utilisateur :** Si un mot commence par "@" et a plus d'un caractère, cela signifie qu'il s'agit d'un nom d'utilisateur. Dans ce cas, le mot est remplacé par le token "@user".
- **Remplacement des liens URL :** Si un mot commence par "http", cela indique qu'il s'agit d'un lien URL. Dans ce cas, le mot est remplacé par le token "http".
- **Ajout du mot traité à la liste tweet_words :** Une fois que le mot a été prétraité, il est ajouté à la liste tweet_words.
- **Création du tweet prétraité :** Une fois que tous les mots ont été prétraités et ajoutés à la liste tweet_words, ils sont joints à l'aide de l'espace comme séparateur pour former le tweet prétraité final, stocké dans la variable tweet_proc**.**

```
# load model and tokenizer
roberta = "cardiffnlp/twitter-roberta-base-sentiment"

model = AutoModelForSequenceClassification.from_pretrained(roberta)
tokenizer = AutoTokenizer.from_pretrained(roberta)

labels = ['Negative', 'Neutral', 'Positive']

# sentiment analysis
encoded_tweet = tokenizer(tweet_proc, return_tensors='pt')
# output = model(encoded_tweet['input_ids'], encoded_tweet['attention_mask'])
output = model(**encoded_tweet)

scores = output[0][0].detach().numpy()
scores = softmax(scores)

for i in range(len(scores)):

    l = labels[i]
    s = scores[i]
    print(l,s)
```

Explication :

- **Définition du modèle et du tokenizer :** Le code spécifie le modèle cardiffnlp/twitter-roberta-base-sentiment pour la tâche d'analyse de sentiment. C'est un modèle basé sur RoBERTa pré-entraîné spécifiquement pour l'analyse de sentiment sur Twitter.
- **Chargement du modèle et du tokenizer :** Le modèle et le tokenizer sont chargés à partir de Hugging Face Transformers à l'aide de la méthode from_pretrained(). Le tokenizer est utilisé pour prétraiter le tweet, tandis que le modèle est utilisé pour prédire le sentiment du tweet.
- **Définition des étiquettes :** Une liste d'étiquettes est définie pour représenter les classes de sentiment possibles. Dans ce cas, les étiquettes sont "Negative", "Neutral" et "Positive".
- **Analyse de sentiment :** Le tweet prétraité est encodé à l'aide du tokenizer, et les tenseurs résultants sont passés au modèle pour obtenir les scores de chaque classe de sentiment. Les scores sont ensuite normalisés en utilisant la fonction softmax pour obtenir des probabilités.
- **Affichage des résultats :** Pour chaque classe de sentiment, le code affiche l'étiquette de la classe et la probabilité associée.

```
Negative 0.13389567
Neutral 0.32832897
Positive 0.53777534
```

- **résultat final de classification**

```python
# Stockage des résultats dans un nouveau DataFrame
results_df = pd.DataFrame({
    'Tweet': [tweet_text],
    'Negative': [scores[0]],
    'Neutral': [scores[1]],
    'Positive': [scores[2]]
})

# Enregistrement des résultats dans un fichier CSV
with open('sentiment_results.csv', 'a', newline='', encoding='utf-8') as csvfile:
    results_df.to_csv(csvfile, header=not csvfile.tell(), index=False, encoding='utf-8')
```

Explication :

- **Création d'un nouveau DataFrame (results_df) pour stocker les résultats de l'analyse de sentiment pour un tweet spécifique :** Le DataFrame est créé en utilisant la fonction pd.DataFrame() de pandas. Les colonnes du DataFrame sont définies comme 'Tweet', 'Negative', 'Neutral', et 'Positive'. Chaque colonne est associée à une liste contenant les valeurs correspondantes pour le tweet actuel. Par exemple, la colonne 'Tweet' contiendra le texte du tweet, tandis que les colonnes 'Negative', 'Neutral', et 'Positive' contiendront les scores associés à chaque sentiment. Chaque liste ne contient qu'une seule valeur, correspondant au tweet actuel et à ses scores associés.
- **Enregistrement des résultats dans un fichier CSV (sentiment_results.csv)** : Le DataFrame results_df est enregistré dans un fichier CSV en utilisant la méthode to_csv() de pandas. Le fichier CSV est ouvert en mode ajout ('a') pour permettre l'ajout de nouvelles lignes à un fichier existant. L'option header=not csvfile.tell() est utilisée pour déterminer si l'en-tête CSV doit être inclus ou non. Si le fichier est vide (csvfile.tell() renvoie 0), l'en-tête est inclus. Sinon, il est ignoré pour éviter la duplication de l'en-tête. L'option index=False est utilisée pour ne pas inclure les index de lignes dans le fichier CSV. L'encodage 'utf-8' est spécifié pour s'assurer que

les caractères Unicode sont correctement pris en charge lors de l'écriture dans le fichier CSV**.**

```
1 sentiment=pd.read_csv('sentiment_results.csv')
```

```
1 sentiment
```

	Tweet	Negative	Neutral	Positive
0	Good deal	0.012976	0.151712	0.835311
1	Supervised full self-driving now $99/month	0.032886	0.862123	0.104992
2	FSD Supervised continues to improve with every...	0.001460	0.083462	0.915078
3	Interesting series about a potentially good fu...	0.001773	0.023027	0.975200
4	Supervised full self-driving now $99/month	0.032886	0.862123	0.104992
...	...	...	...	...
253	NaN	0.258294	0.451272	0.290433
254	𝕏 supports the people of Brazil, without regar...	0.022530	0.617025	0.360444
255	NaN	0.258294	0.451272	0.290433
256	[Scene: @Alexandre & @ElonMusk in psychoanalys...	0.168956	0.788598	0.042446
257	This is the heart of the problem. What say you...	0.598992	0.383214	0.017794

Explication :

- **L’affichage de fichier sentiment résulte**

```
from pymongo import MongoClient
cluster=MongoClient("mongodb+srv://halima:halima@cluster0.wjt9qvc.mongodb.net/?retryWrites=true&w=majority&appName=Cluster0"
db=cluster['elonmusk']
collection=db['Nlp']

collection.insert_many(sentiment_list)
```

```
InsertManyResult([ObjectId('661e4a06fcd1a350e1d4de20'), ObjectId('661e4a06fcd1a350e1d4de21'), ObjectId('661e4a06fcd1a350e1d4de2
2'), ObjectId('661e4a06fcd1a350e1d4de23'), ObjectId('661e4a06fcd1a350e1d4de24'), ObjectId('661e4a06fcd1a350e1d4de25'), ObjectId
('661e4a06fcd1a350e1d4de26'), ObjectId('661e4a06fcd1a350e1d4de27'), ObjectId('661e4a06fcd1a350e1d4de28'), ObjectId('661e4a06fcd
1a350e1d4de29'), ObjectId('661e4a06fcd1a350e1d4de2a'), ObjectId('661e4a06fcd1a350e1d4de2b'), ObjectId('661e4a06fcd1a350e1d4de2
c'), ObjectId('661e4a06fcd1a350e1d4de2d'), ObjectId('661e4a06fcd1a350e1d4de2e'), ObjectId('661e4a06fcd1a350e1d4de2f'), ObjectId
('661e4a06fcd1a350e1d4de30'), ObjectId('661e4a06fcd1a350e1d4de31'), ObjectId('661e4a06fcd1a350e1d4de32'), ObjectId('661e4a06fcd
1a350e1d4de33'), ObjectId('661e4a06fcd1a350e1d4de34'), ObjectId('661e4a06fcd1a350e1d4de35'), ObjectId('661e4a06fcd1a350e1d4de3
6'), ObjectId('661e4a06fcd1a350e1d4de37'), ObjectId('661e4a06fcd1a350e1d4de38'), ObjectId('661e4a06fcd1a350e1d4de39'), ObjectId
('661e4a06fcd1a350e1d4de3a'), ObjectId('661e4a06fcd1a350e1d4de3b'), ObjectId('661e4a06fcd1a350e1d4de3c'), ObjectId('661e4a06fcd
1a350e1d4de3d'), ObjectId('661e4a06fcd1a350e1d4de3e'), ObjectId('661e4a06fcd1a350e1d4de3f'), ObjectId('661e4a06fcd1a350e1d4de4
0'), ObjectId('661e4a06fcd1a350e1d4de41'), ObjectId('661e4a06fcd1a350e1d4de42'), ObjectId('661e4a06fcd1a350e1d4de43'), ObjectId
```

Explication :

- Crée une instance de MongoClient en se connectant au cluster MongoDB Atlas spécifié, la base de données elonmusk et la collection Nlp dans cette base de données.
- insère plusieurs documents dans la collection sélectionnée. sentiment_list est une liste de dictionnaires, où chaque dictionnaire représente un document à insérer dans la collection

4 Power BI

Power BI offre un ensemble complet d'outils pour visualiser, analyser et partager les résultats d'un projet d'analyse sémantique des tweets, ce qui en fait un choix pertinent pour la visualisation des insights tirés des données textuelles des tweets.

4.1 Définition :

Power BI est une plateforme d'analyse commerciale (BI) développée par Microsoft, conçue pour permettre aux utilisateurs de visualiser et d'analyser leurs données de manière efficace. Voici pourquoi Power BI pourrait être utilisé dans la visualisation d'un projet d'analyse sémantique des tweets :

- **Visualisation avancée :** Power BI offre une large gamme de visualisations interactives telles que des graphiques, des cartes, des jauges, des tableaux, etc., qui peuvent être utilisées pour représenter les résultats de l'analyse sémantique des tweets de manière efficace et attrayante.

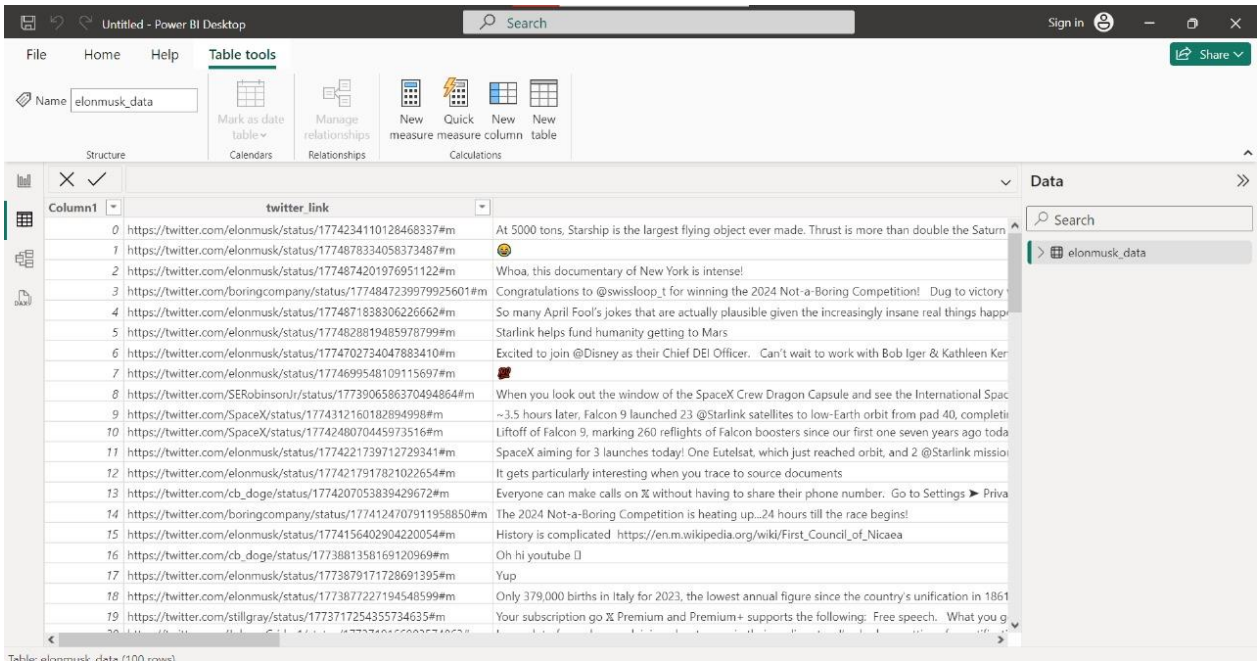
- **Intégration de données multiples :** Power BI permet l'intégration de données à partir de plusieurs sources, ce qui signifie que les résultats de l'analyse sémantique des tweets peuvent être combinés avec d'autres données pertinentes pour fournir une vue d'ensemble complète.

- **Actualisation automatique des données :** Power BI peut être configuré pour se connecter à des sources de données en ligne telles que Twitter et actualiser automatiquement les données à intervalles réguliers, garantissant que les visualisations restent à jour avec les dernières informations.

- **Analyse exploratoire :** Les fonctionnalités d'analyse de données de Power BI permettent aux utilisateurs d'explorer les données de manière interactive, en filtrant, en triant et en perforant les visualisations pour découvrir des insights cachés dans les données des tweets. Partage et collaboration : Les rapports et tableaux de bord créés dans Power BI peuvent être facilement partagés avec d'autres utilisateurs au sein de l'organisation, ce qui facilite la collaboration et la prise de décision basée sur des données.

4.2 Application

Insertion le fichier csv de elon musk tweets et sentiment result



Elon musk tableau contient liens,texts,likes,comments,date.

Et pour sentiment_result on a tweet,positive,nigative,neutre:

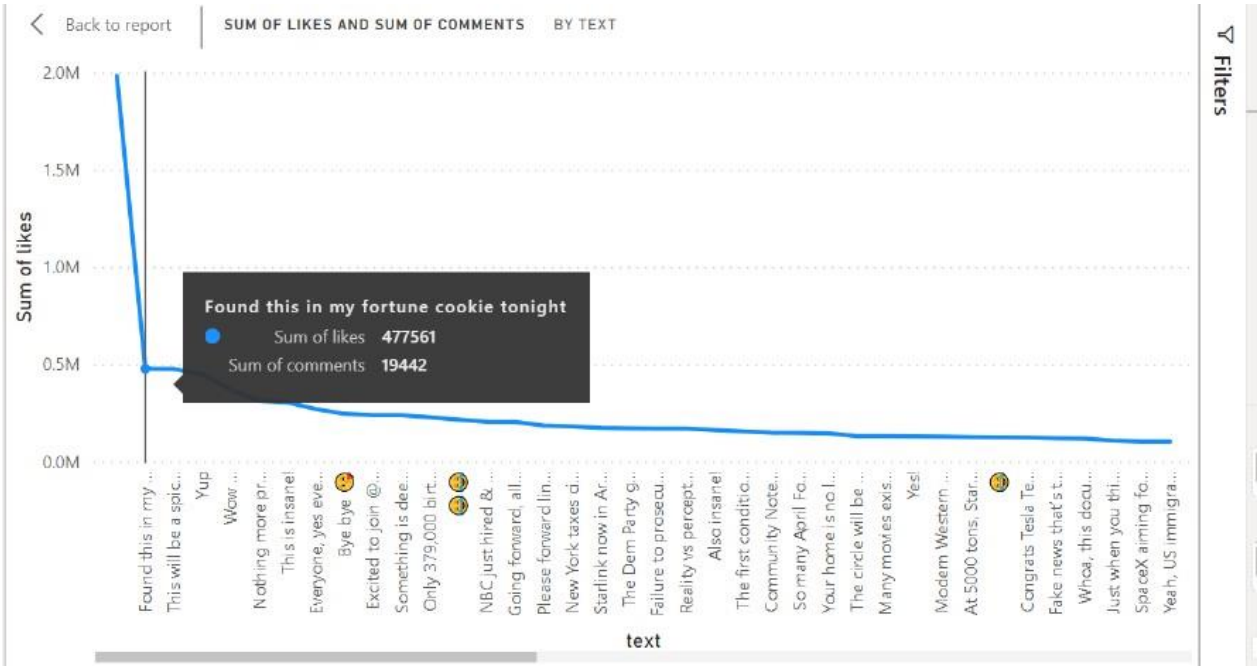
Tweet	Negative	Neutral	Positive
Good deal	0.01297645	0.15171248	0.83531106
Supervised full self-driving now $99/month	0.032885574	0.8621228	0.10499174
FSD Supervised continues to improve with every over-the-air software update Latest version 12.3.4 rollin	0.0014600678	0.08346236	0.9150776
Interesting series about a potentially good future	0.001772869	0.023027262	0.9751999
Supervised full self-driving now $99/month	0.032885574	0.8621228	0.10499174
FSD Supervised continues to improve with every over-the-air software update Latest version 12.3.4 rollin	0.0014600678	0.08346236	0.9150776
Interesting series about a potentially good future	0.001772869	0.023027262	0.9751999
United States laws prevent X from participating in corruption that violates the laws of other countries, whi	0.70466644	0.281572	0.0137615595
Interesting	0.014022414	0.25492725	0.7310503
If you're experiencing severe neck/back pain, I recommend looking into a disc replacement. If you do, er	0.18209913	0.63116217	0.1867387
	0.2582943	0.45127246	0.2904333
bon appétit 😋 🍽	0.0034331337	0.19733578	0.7992311
Worth noting that the actual brand safety score is almost perfect	0.009907305	0.12827682	0.86181587
Thanks to the X team for discovering this and @DoubleVerify for responding quickly to correct the error	0.010850002	0.12525086	0.8638992
Greece is one of dozens of countries experiencing population collapse due to low birth rates	0.80627626	0.18521148	0.008512309
Create or join X Communities!	0.015494934	0.6864768	0.2980282
	0.2582943	0.45127246	0.2904333
😕	0.16171488	0.61519814	0.22308695
The refreshing breeze of the Overton window opening	0.0020885197	0.13773227	0.86017925
	0.2582943	0.45127246	0.2904333
Always wondered what that chair was for	0.11098272	0.83275753	0.056259803

elonmusk_data.csv

File Origin: 65001: Unicode (UTF-8) | Delimiter: Comma | Data Type Detection: Based on first 200 rows

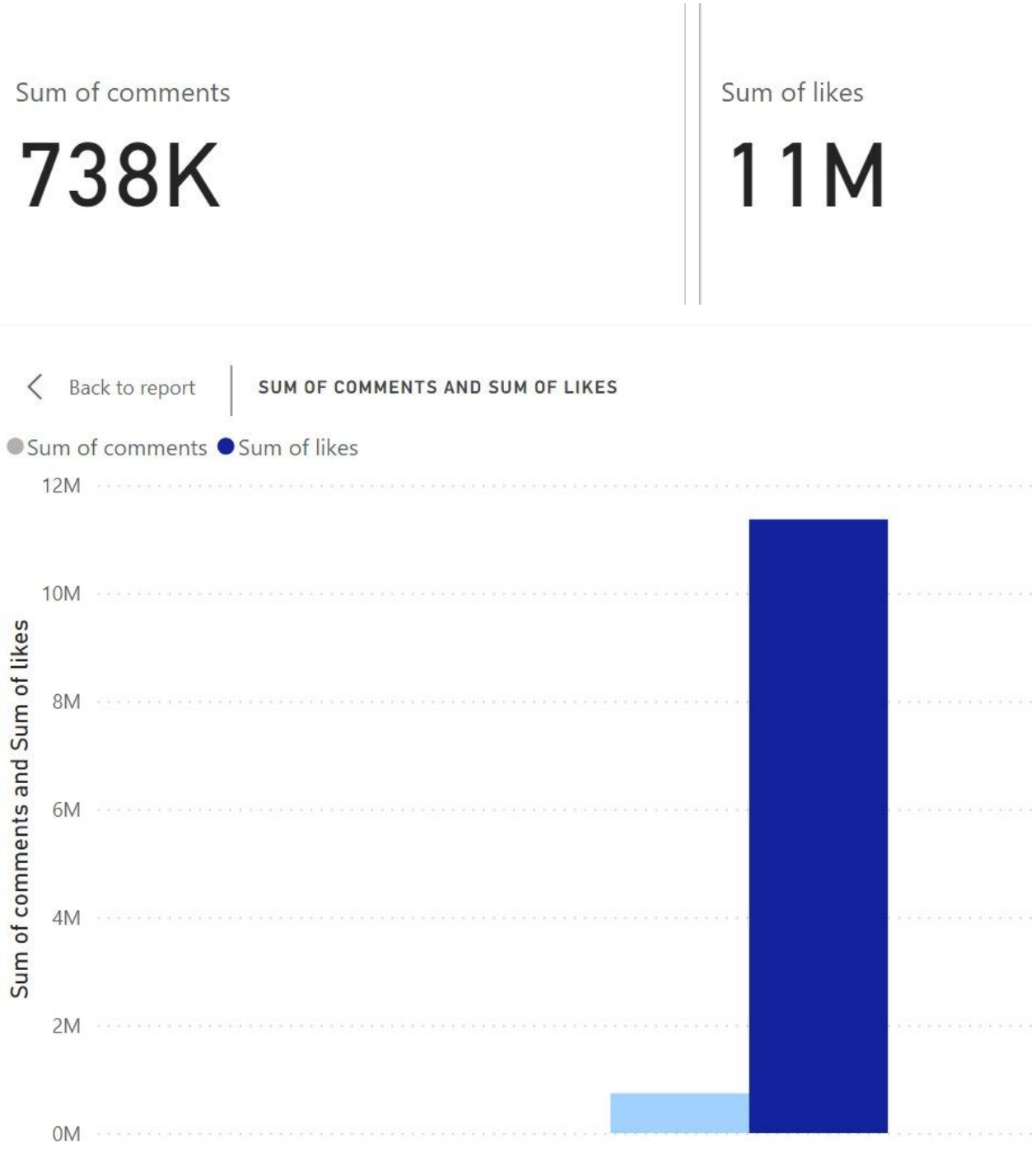
	twitter_link	text	date	likes	comments
0	https://twitter.com/elonmusk/status/17742341101284...	At 5000 tons, Starship is the largest flying object ever m...	Mar 31, 2024 · 12:35 AM UTC	125588	14045
1	https://twitter.com/elonmusk/status/17748783340583...	😊	Apr 1, 2024 · 7:15 PM UTC	106748	3746
2	https://twitter.com/elonmusk/status/17748742019769...	Whoa, this documentary of New York is intense!	Apr 1, 2024 · 6:59 PM UTC	118363	8003
3	https://twitter.com/boringcompany/status/177484723...	Congratulations to @swissloop_t for winning the 2024...	Apr 1, 2024 · 5:12 PM UTC	4349	342
4	https://twitter.com/elonmusk/status/17748718383062...	So many April Fool's jokes that are actually plausible giv...	Apr 1, 2024 · 6:49 PM UTC	147960	8543
5	https://twitter.com/elonmusk/status/17748288194859...	Starlink helps fund humanity getting to Mars	Apr 1, 2024 · 3:58 PM UTC	36534	4841
6	https://twitter.com/elonmusk/status/17747027340478...	Excited to join @Disney as their Chief DEI Officer. Can't...	Apr 1, 2024 · 7:37 AM UTC	238745	15342
7	https://twitter.com/elonmusk/status/17746995481091...	💯	Apr 1, 2024 · 7:25 AM UTC	69421	4427
8	https://twitter.com/SERobinsonJr/status/17739065863...	When you look out the window of the SpaceX Crew Dra...	Mar 30, 2024 · 2:54 AM UTC	23420	1407
9	https://twitter.com/SpaceX/status/1774312160182894...	~3.5 hours later, Falcon 9 launched 23 @Starlink satellit...	Mar 31, 2024 · 5:45 AM UTC	13050	1090
10	https://twitter.com/SpaceX/status/1774248070445973...	Liftoff of Falcon 9, marking 260 reflights of Falcon boost...	Mar 31, 2024 · 1:31 AM UTC	17724	1258
11	https://twitter.com/elonmusk/status/17742217397127...	SpaceX aiming for 3 launches today! One Eutelsat, whic...	Mar 30, 2024 · 11:46 PM UTC	103483	7517
12	https://twitter.com/elonmusk/status/17742179178210...	It gets particularly interesting when you trace to source...	Mar 30, 2024 · 11:31 PM UTC	7675	1112
13	https://twitter.com/cb_doge/status/177420705383942...	Everyone can make calls on 𝕏 without having to share t...	Mar 30, 2024 · 10:48 PM UTC	8451	1162
14	https://twitter.com/boringcompany/status/177412470...	The 2024 Not-a-Boring Competition is heating up...24 h...	Mar 30, 2024 · 5:21 PM UTC	7650	620
15	https://twitter.com/elonmusk/status/17741564029042...	History is complicated https://en.m.wikipedia.org/wiki/...	Mar 30, 2024 · 7:27 PM UTC	37215	6446
16	https://twitter.com/cb_doge/status/177388135816912...	Oh hi youtube	Mar 30, 2024 · 1:14 AM UTC	10348	1625
17	https://twitter.com/elonmusk/status/17738791717286...	Yup	Mar 30, 2024 · 1:05 AM UTC	250549	11459
18	https://twitter.com/elonmusk/status/17738772271945...	Only 379,000 births in Italy for 2023, the lowest annual...	Mar 30, 2024 · 12:57 AM UTC	228051	18899
19	https://twitter.com/stillgray/status/177371725435573...	Your subscription go 𝕏 Premium and Premium+ suppor...	Mar 29, 2024 · 2:22 PM UTC	5225	858

4.2.1 Line Chart Visualisation

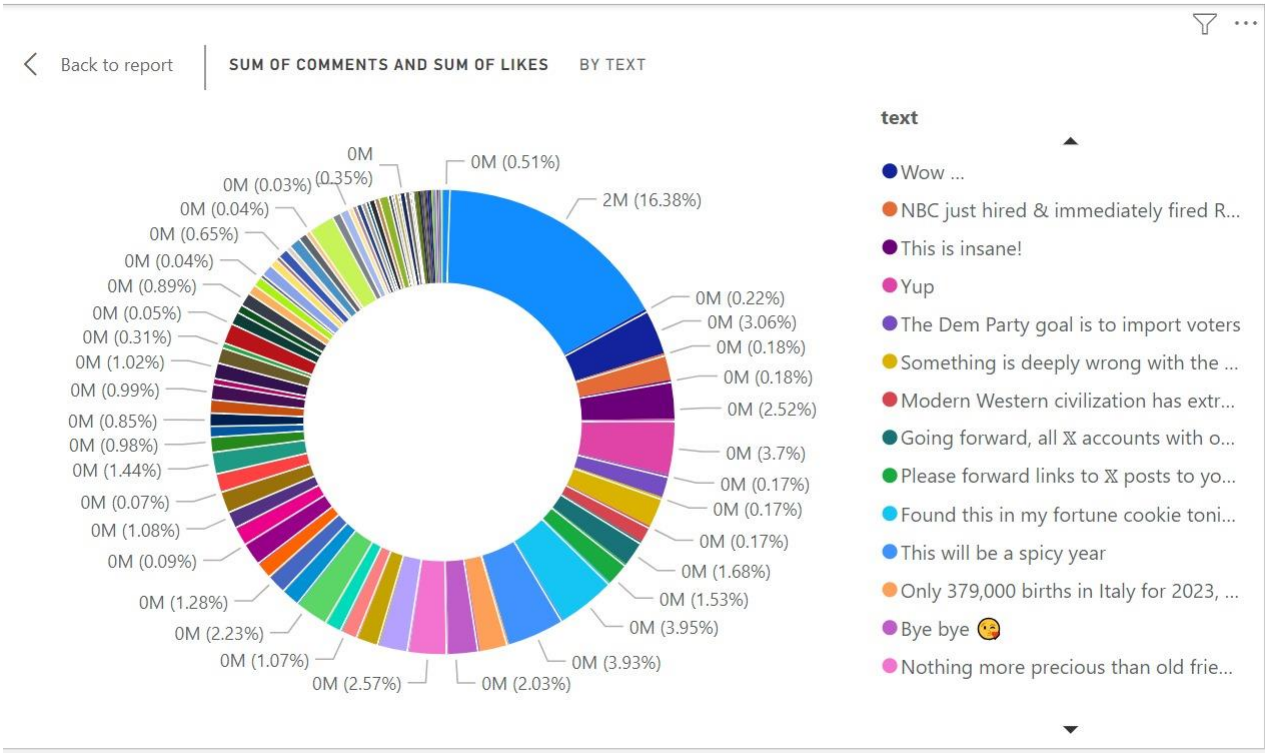


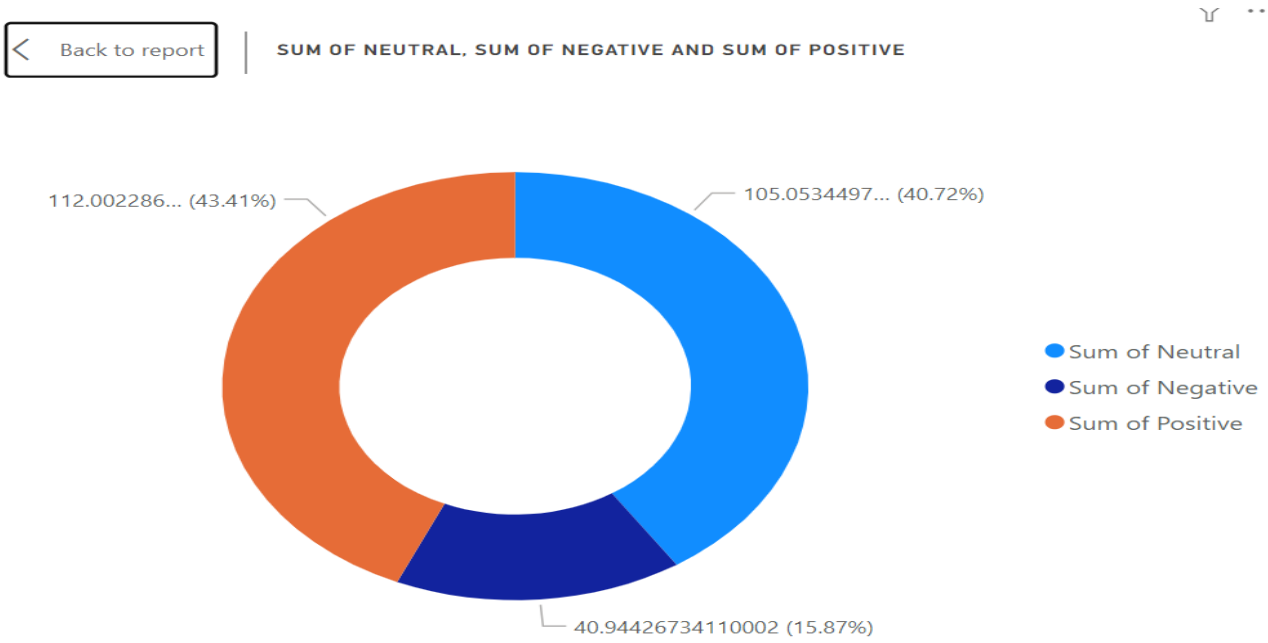
Le tweet **"found this in my fortune cookie tonight"** a le plus grand nombre de likes et comments

4.2.2 Comparison entre les comments et likes



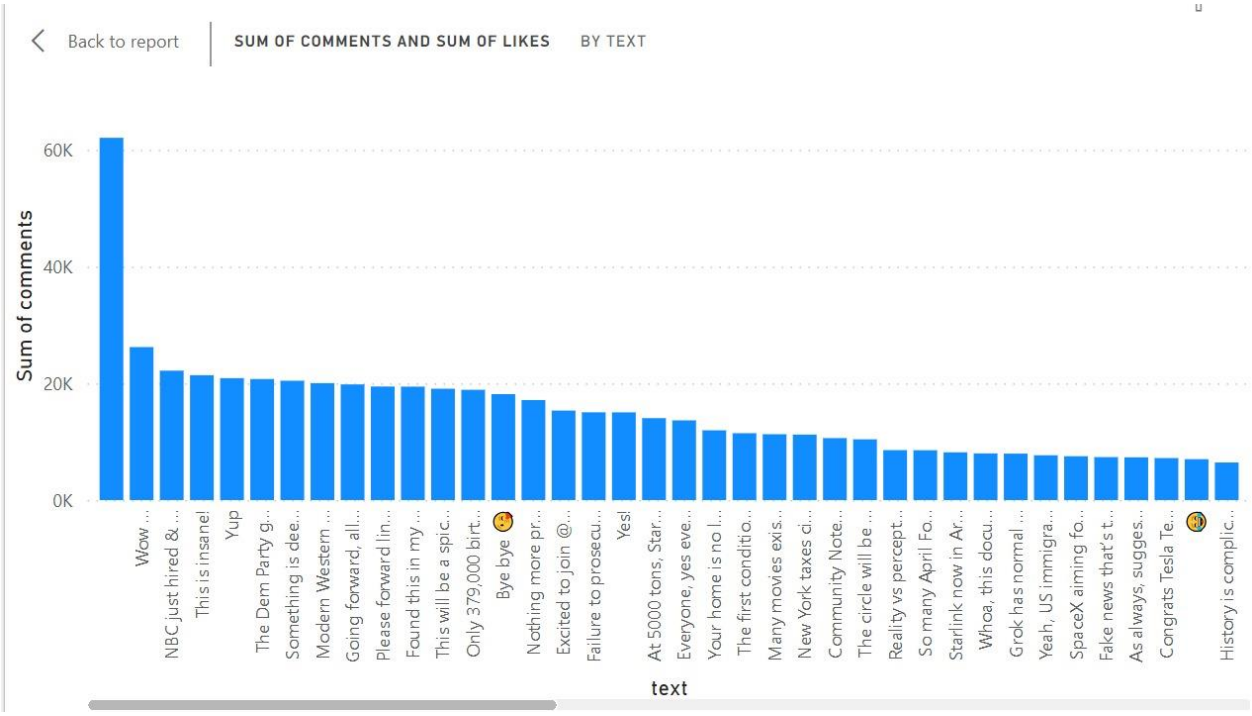
4.2.3 Donut Chart



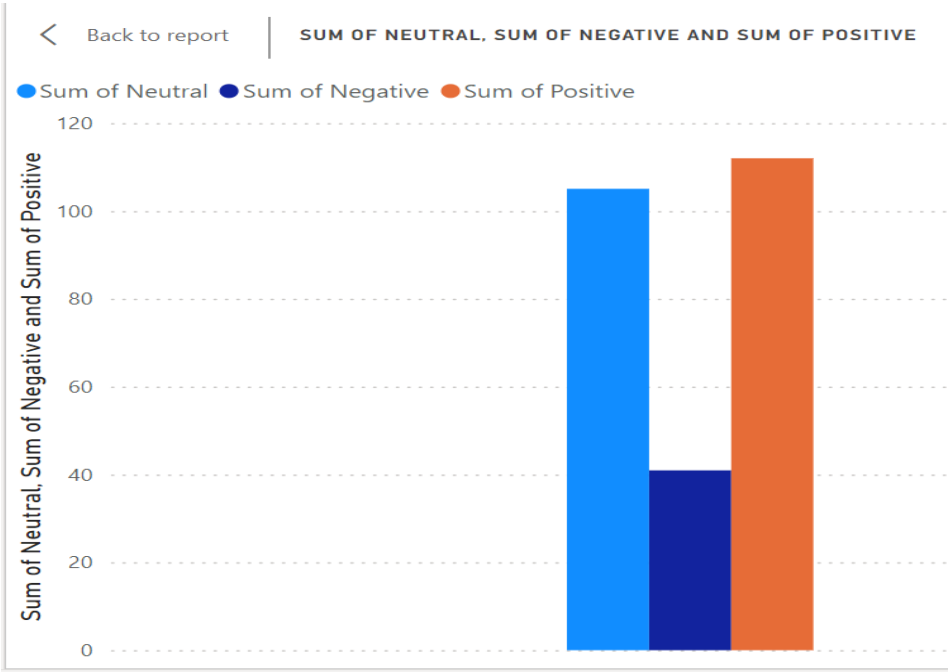


Notre dataset est contient 40.72% de neutre, 43.41% de positive et 15.87% de negative

4.2.4 Stacked Columns Chart



Comparaison de tweets par rapport les commnets et likes



Conclusion

Dans ce projet, nous avons développé une solution pour l'analyse des sentiments sur Twitter en utilisant le traitement automatique du langage naturel (NLP) et la base de données MongoDB.

Nous avons démontré comment collecter des tweets en temps réel, les prétraiter avec NLP pour l'analyse de sentiment, et stocker les données dans une base de données NoSQL pour une gestion efficace.

Cette solution peut être étendue pour surveiller les tendances, comprendre les opinions du public et prendre des décisions informées dans divers domaines tels que le marketing, la politique et la veille stratégique

Références

- https://docs.mongodb.com/
- https://datascientest.com/introduction-au-nlp-natural-language-processing
- https://medium.com/scalereal/million-tweets-and-counting-how-snscrape-can-help-you-scrape-big-data-on-twitter-5c0240cab4f3