ENSIAS

جامعة محمد الخامس بالرباط
Université Mohammed V de Rabat

ANALYSE TEMPS RÉEL DES DISCUSSIONS REDDIT CAN 2025

Architecture Big Data & Machine Learning

Présentée par:

BOUTANFIT Salma
ELIDRISSI Asma
JENNANE Salma
Ouahib Salma

Jury:

Yasser EL MADANI
EL ALAMI

Plan

1 Contexte & Problématique

2 Objectifs du projet

3 Architecture du système

4 Réalisation

5 Diffucltés rencontrés et solutions

6 Conclusion et perspectives

1 Contexte & problématique

1.Contexte

- Événement sportif majeur africain
- Millions d'interactions sur réseaux sociaux

- Plateforme de discussions thématiques
- Communautés (subreddits) engagées
- Source riche d'opinions et analyses

DÉFIS BIG DATA

Volume massif • Centaines de posts par heure

Flux continu • Données en temps réel

Données non structurées • Texte libre, langues multiples

Traitement temps réel • Analyses instantanées

QUESTION CENTRALE

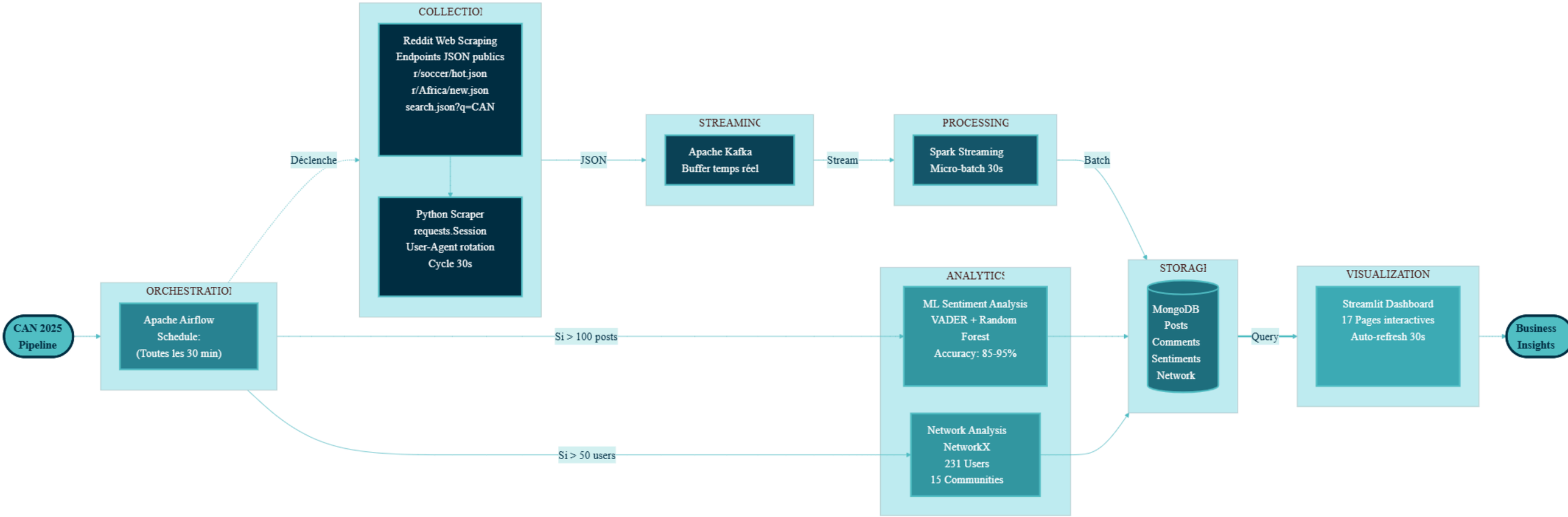
▶ ***Comment concevoir une architecture Big Data capable** de collecter, traiter et analyser en **temps quasi-réel** les discussions Reddit sur la CAN 2025 ?*

2 Objectifs du projet

Objectifs du projet

- ✓ PIPELINE TEMPS RÉEL
- ✓ ANALYSE DE SENTIMENTS (ML)
- ✓ ANALYSE DE RÉSEAU SOCIAL
- ✓ DASHBOARD INTERACTIF

4 Architecture du système


CAN 2025 Pipeline
ORCHESTRATIO
Apache Airflow
Schedule:
(Toutes les 30 min)
Déclenche
COLLECTIO
Reddit Web Scraping
Endpoints JSON publics
r/soccer/hot.json
r/Africa/new.json
search.json?q=CAN
Python Scraper
requests.Session
User-Agent rotation
Cycle 30s
JSON
STREAMINC
Apache Kafka
Buffer temps réel
Stream
PROCESSING
Spark Streaming
Micro-batch 30s
Batch
Si > 100 posts
Si > 50 users
ANALYTICS
ML Sentiment Analysis
VADER + Random
Forest
Accuracy: 85-95%
Network Analysis
NetworkX
231 Users
15 Communities
STORAGI
MongoDB
Posts
Comments
Sentiments
Network
Query
VISUALIZATION
Streamlit Dashboard
17 Pages interactives
Auto-refresh 30s
Business Insights

POURQUOI KAPPA ?

✓ **SIMPLICITÉ**

Une seule couche de traitement

✓ **LATENCE FAIBLE**

Données affichées en < 2 minutes

✓ **REJOUABILITÉ**

Kafka permet retraitement historique

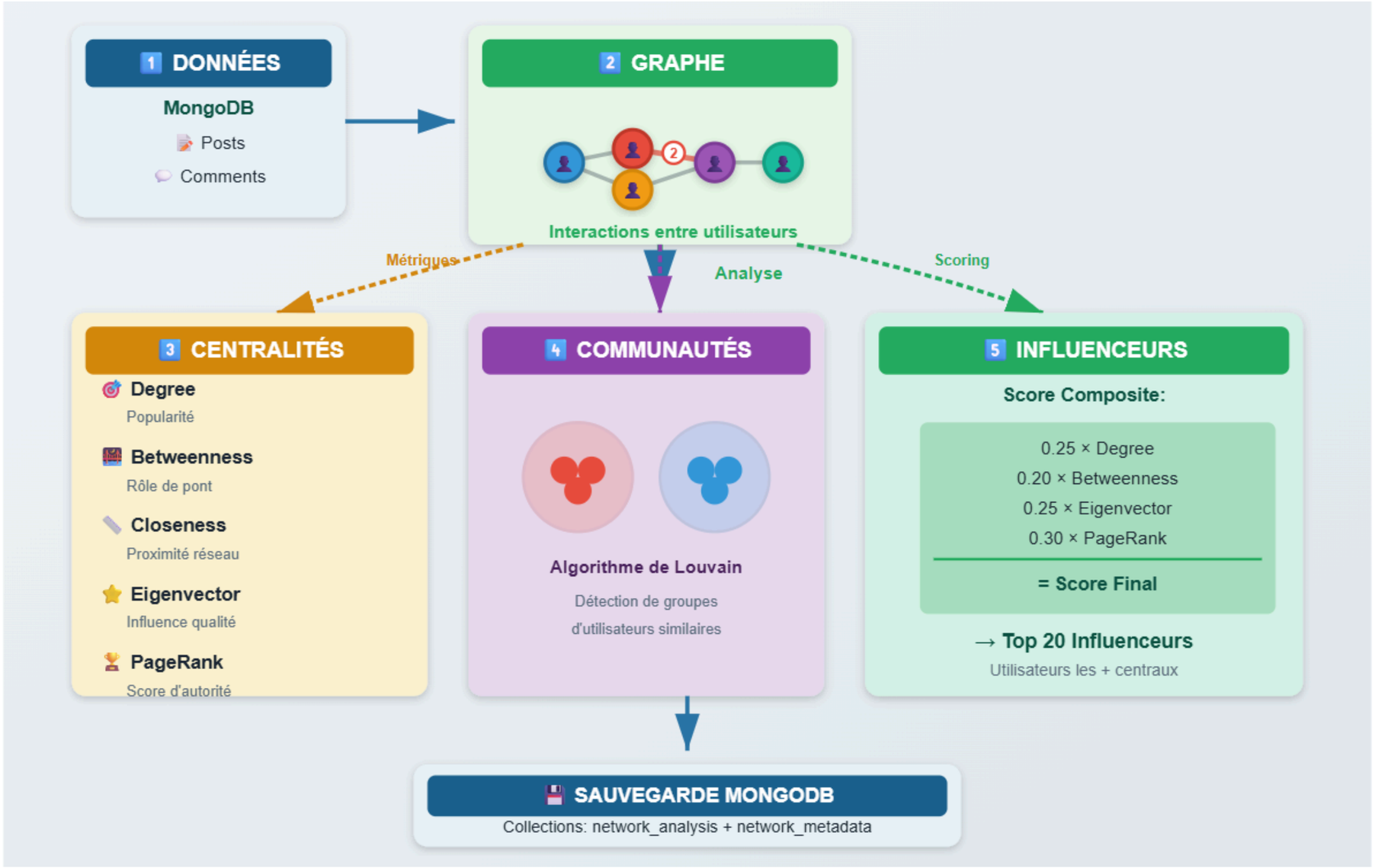
✓ **SCALABILITÉ**

Ajout de consumers facilité

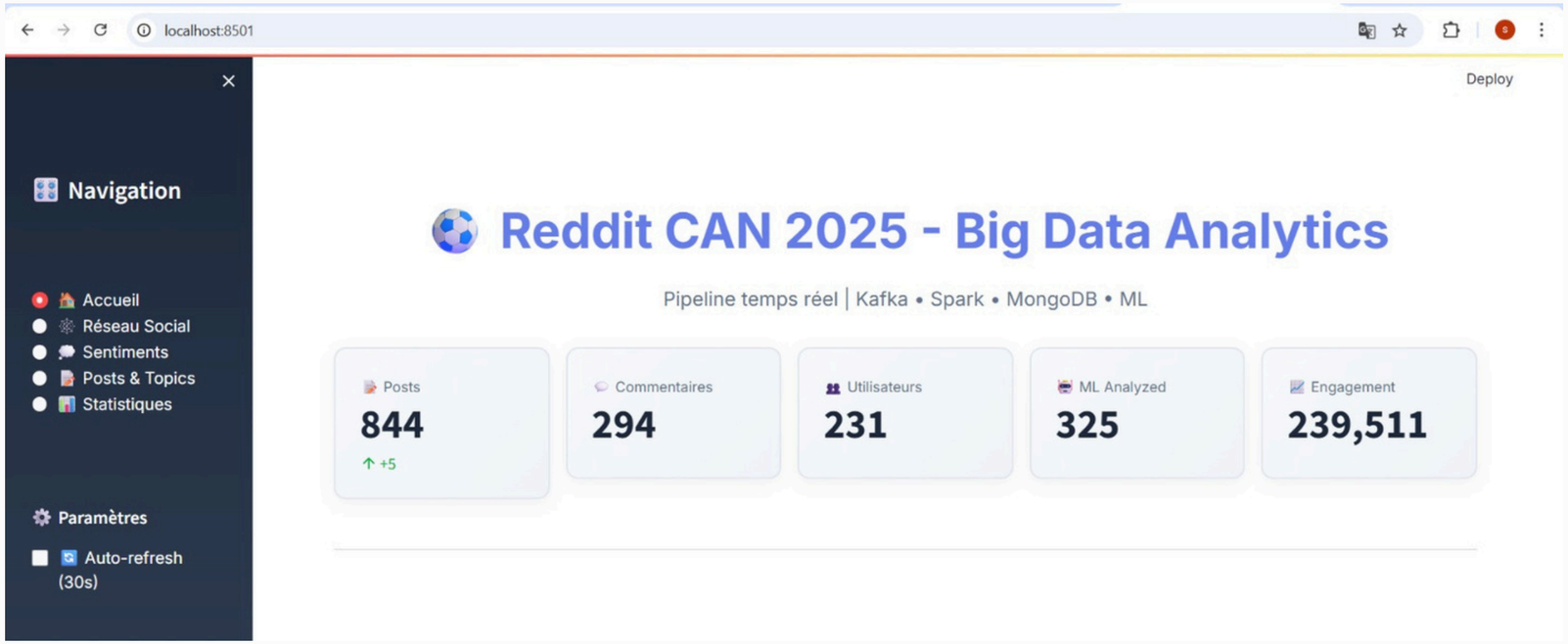
Composant	Latence	Fonction
Scraping	30 s	Collecte Reddit
Kafka	< 1 s	Message broker
Spark	30 s	Traitement
MongoDB	Instantané	Écriture
ML Analysis	Variable	Périodique
Network Analysis	Variable	Périodique
Dashboard	30 s	Refresh cache

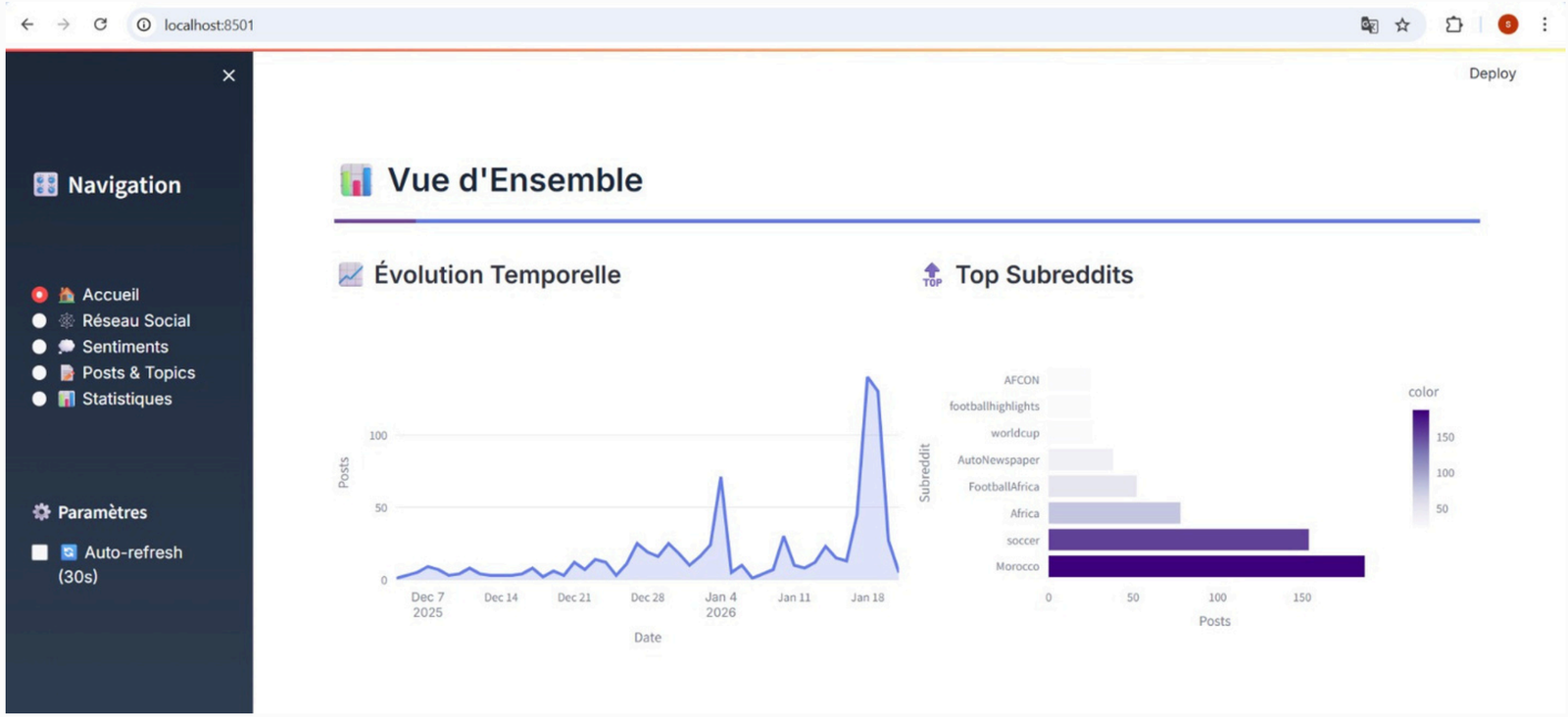
Métrique	Valeur
THROUGHPUT	
Posts/batch Spark	30 posts
Capacité théorique	1,000+ posts/heure
Posts collectés (réels)	844 posts
DISPONIBILITÉ	
Pipeline	24/7
Healthchecks	Docker
Retry automatique	3 tentatives
Monitoring	Airflow centralisé
STOCKAGE	
Taille moyenne	~1.2 MB / 100 posts
Collections MongoDB	5 collections
Backup	Quotidien possible

Analyse de Réseau Social

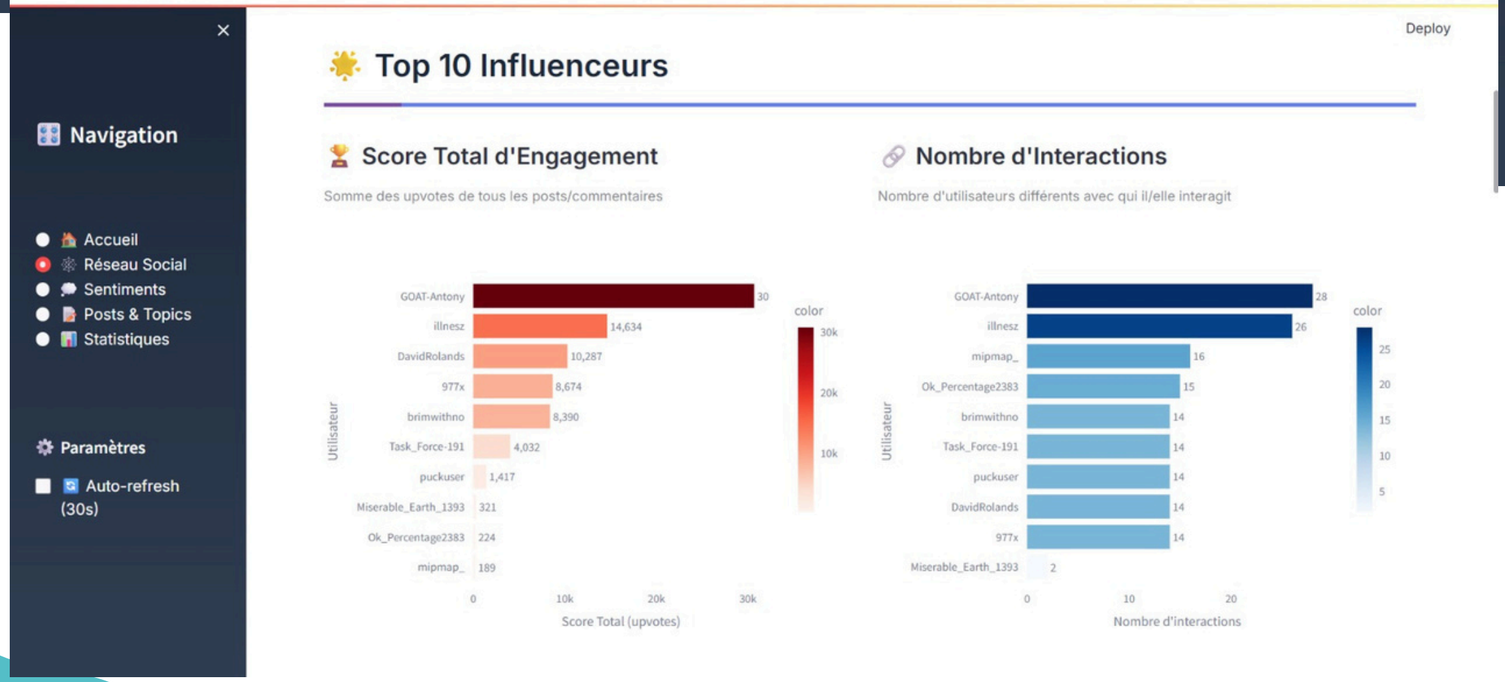


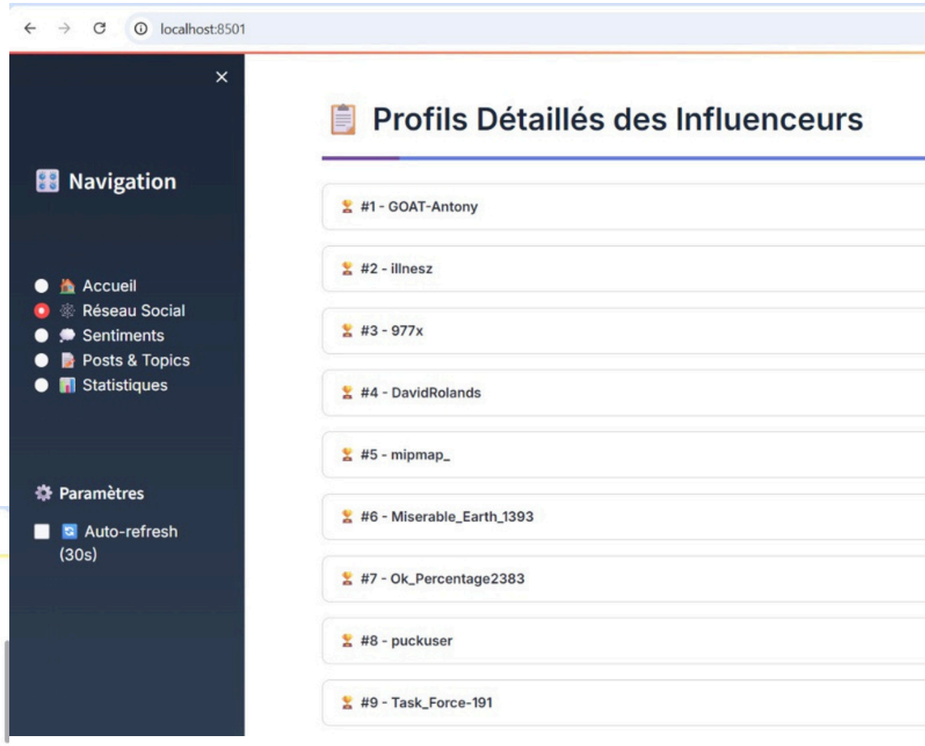
6 Résultats et démonstration


localhost:8501
Deploy
Navigation
Accueil
Réseau Social
Sentiments
Posts & Topics
Statistiques
Paramètres
Auto-refresh (30s)
Reddit CAN 2025 - Big Data Analytics
Pipeline temps réel | Kafka • Spark • MongoDB • ML
Posts
844
+5
Commentaires
294
Utilisateurs
231
ML Analyzed
325
Engagement
239,511


localhost:8501
Deploy
Navigation
Accueil
Réseau Social
Sentiments
Posts & Topics
Statistiques
Paramètres
Auto-refresh (30s)
Vue d'Ensemble
Évolution Temporelle
Posts
100
50
0
Dec 7 2025
Dec 14
Dec 21
Dec 28
Jan 4 2026
Jan 11
Jan 18
Date
Top Subreddits
Subreddit
AFCON
footballhighlights
worldcup
AutoNewspaper
FootballAfrica
Africa
soccer
Morocco
0
50
100
150
Posts
color
150
100
50


localhost:8501
Deploy
Navigation
Accueil
Réseau Social
Sentiments
Posts & Topics
Statistiques
Paramètres
Auto-refresh (30s)
Analyse du Réseau Social
Qui sont les influenceurs ? Quelles sont les communautés ?
Vue d'Ensemble du Réseau
Utilisateurs
231
Interactions
218
Densité
0.82%
% connexions possibles
Communautés
15
Groupes détectés
Clustering
0.0%
Tendance groupes


localhost:8501
Deploy
Navigation
Accueil
Réseau Social
Sentiments
Posts & Topics
Statistiques
Paramètres
Auto-refresh (30s)
Top 10 Influenceurs
Score Total d'Engagement
Somme des upvotes de tous les posts/commentaires
GOAT-Antony 30
illnesz 14,634
DavidRolands 10,287
977x 8,674
brimwithno 8,390
Task_Force-191 4,032
puckuser 1,417
Miserable_Earth_1393 321
Ok_Percentage2383 224
mipmap_ 189
Utilisateur
color
30k
20k
10k
0
10k
20k
30k
Score Total (upvotes)
Nombre d'Interactions
Nombre d'utilisateurs différents avec qui il/elle interagit
GOAT-Antony 28
illnesz 26
mipmap_ 16
Ok_Percentage2383 15
brimwithno 14
Task_Force-191 14
puckuser 14
DavidRolands 14
977x 14
Miserable_Earth_1393 2
Utilisateur
color
25
20
15
10
5
0
10
20
Nombre d'interactions


localhost:8501
Navigation
Accueil
Réseau Social
Sentiments
Posts & Topics
Statistiques
Paramètres
Auto-refresh (30s)
Profils Détaillés des Influenceurs
#1 - GOAT-Antony
#2 - illnesz
#3 - 977x
#4 - DavidRolands
#5 - mipmap_
#6 - Miserable_Earth_1393
#7 - Ok_Percentage2383
#8 - puckuser
#9 - Task_Force-191

Deploy

🏘️ Communautés Détectées

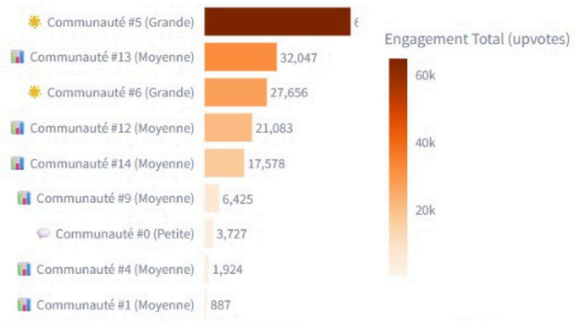
Algorithme de Louvain : Détecte automatiquement les groupes d'utilisateurs qui interagissent fréquemment ensemble.

💡 **Interprétation** :

- Grandes communautés (>25 membres) = Groupes de fans très actifs
- Moyennes (15-25 membres) = Discussions thématiques
- Petites (<15 membres) = Niches spécialisées

👥 Taille des Communautés (Top 10)

📈 Engagement Total par Communauté



Deploy

📋 Tableau Récapitulatif des Communautés

Communauté	Membres	Interactions Moy.	Engagement Total
Communauté #5 (Grande)	28	1.96	64876 🔥
Communauté #6 (Grande)	26	1.96	27656 🔥
Communauté #1 (Moyenne)	17	1.88	887 🔥
Communauté #2 (Moyenne)	16	1.88	598 🔥
Communauté #4 (Moyenne)	15	1.87	1924 🔥
Communauté #9 (Moyenne)	15	1.87	6425 🔥
Communauté #12 (Moyenne)	15	1.87	21083 🔥
Communauté #13 (Moyenne)	15	1.93	32047 🔥
Communauté #14 (Moyenne)	15	1.93	17578 🔥
Communauté #0 (Petite)	14	1.86	3727 🔥

🧭 Navigation

- 🏠 Accueil
- 🕸️ Réseau Social
- 💬 Sentiments
- 📝 Posts & Topics
- 📊 Statistiques

⚙️ Paramètres

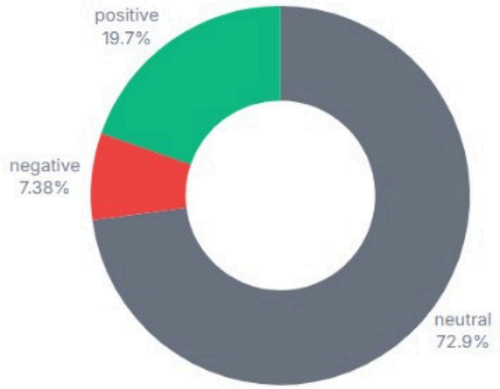
- 🔄 Auto-refresh (30s)

Analyse des Sentiments

Classification ML des opinions

Modèle	Accuracy	Baseline	Coverage
Random F...	85-95%	VADER	38.7%

Distribution



Exemples

Positifs Neutres Négatifs

Malian fans trolling a Tunisian fan following Mali's win vs Tunisia th...

Ismail Jakobs assists Nicolas Jackson for Senegal's first goal in AFCO...

Today's MotD: Uganda vs Nigeria (AFCON, Morocco)...

18 | 5 | r/soccer

Oya! How far?

Today's Match of the Day is a series covering a fixture happening today, somewhere in the world, between historical rivals, across

The point is simple: you don't need a bet to care about a football match; All you really need is context.

You can follow the game on [QFAX](https://qfax.football) or your preferred scores app (Fotmo

Exemples

Positifs Neutres Négatifs

[BBC] 'I can kill you right now' - Sudan's footballers on civil war...

2054 | 92 | r/soccer

"They didn't even give him a chance. They shot him more than 20 or 25 times.

"One of our childhood friends was also with them, but he couldn't say anything. So he just saw our friend die in front of his eyes

The matter of fact way in which Sudan forward John Mano recounts the death of his best friend Medo is at odds with the intense loo

Med

After poor results at Afcon (africa), the Gabon foopball team has been...

Sudan's warriors fall fighting as teenage Mbaye seals Senegal's place ...

Posts & Topics

Exploration des discussions

Subreddit: Tous

Score Min: 0 — 0 — 11530

Trier par: Score

- Score
- Date
- Commentaires

844 posts trouvés

Posts Détaillés

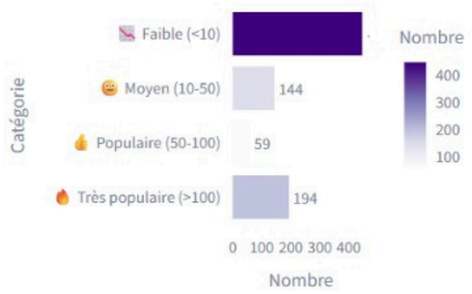
- A DR Congo fan stood for 90+ minutes during the AFCON game vs Senegal, honoring ...
- Zinedine Zidane at AFCON watching his son Luca Zidane play for Algeria...
- A fan in the Senegal vs Botswana game (no idea for which team)...

Statistiques Avancées

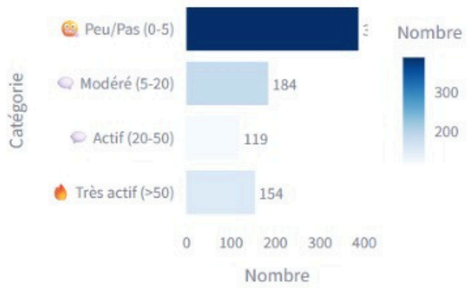
Analyses statistiques approfondies avec insights visuels

Analyse des Données Clés

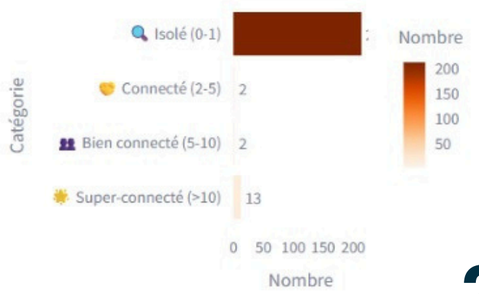
Répartition des Scores



Répartition des Commentaires



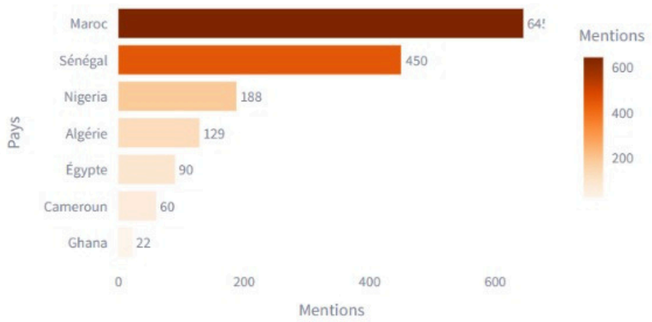
Répartition des Connexions



🌎 Analyses Temporelles & Géographiques

🗓️ Activité par Jour de la Semaine

🌎 Pays Africains Mentionnés



⏳ **Dimanche** est le jour le plus actif !

⏳ **Maroc** domine avec 645 mentions !

🎯 Insights Statistiques

📊 Répartition :

- 🔥 **Top 25%** (score > 80) : **211 posts** (25.0%)
- 📉 **Bottom 25%** (score < 1) : **37 posts** (4.4%)
- 💡 Les posts très populaires (>100 upvotes) sont **rares** mais génèrent beaucoup d'attention

💬 Engagement :

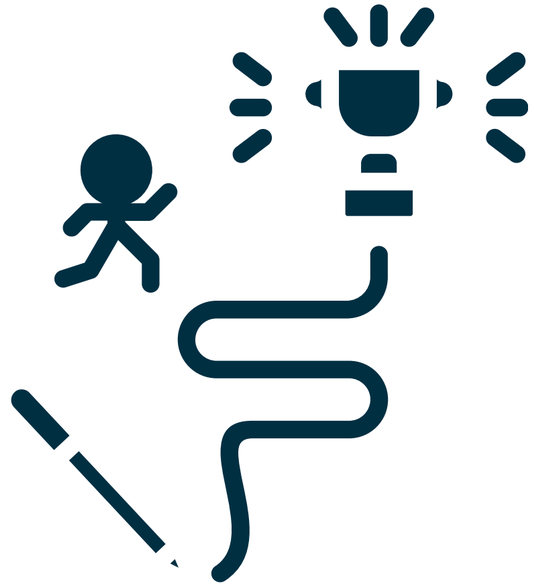
- 🔥 **Top 10%** (>81 comments) : **84 posts**
- 💬 Ces posts génèrent **28,933** commentaires (72.4% du total)
- 💡 **10.0% des posts → 72.4% des discussions**

🌟 Connectivité :

- 🌟 **Top 10%** (>1 connexions) : **17 utilisateurs**
- 💡 Ces **super-connecteurs** sont les influenceurs du réseau
- 📈 Le reste (214 users) a **<1 connexions**

7

DIFFICULTÉS RENCONTRÉES & SOLUTIONS



RATE LIMITING REDDIT

Problème :
Blocage de l'API Reddit
Risque de blocage du scraper

Solution :
- Scraping espacé (30 secondes)
- Rotation User-Agent
- Gestion erreurs 429 (Too Many Req.)

VOLUME INITIAL LIMITÉ

Problème :
Peu de posts en début de projet
Dataset insuffisant pour ML

Solution :
- Élargissement des mots-clés
- Collecte historique (7 jours)
- Extraction des commentaires
- Multi-subreddits

LABELLISATION ML

Problème :
Pas de labels humains disponibles
Annotation manuelle trop coûteuse

Solution :
- VADER pour génération auto labels
- Validation sur échantillon manuel
- Approche semi-supervisée

7 Conclusion et perspective

Objectifs atteints

- Architecture Kappa fonctionnelle et performante
- Pipeline temps réel avec latence 1–2 minutes
- ML fiable (85–95% accuracy)
- Réseau social cartographié (15 communautés, 20 influenceurs)
- Dashboard interactif opérationnel
- Résultats concrets sur données réelles
- Projet applicable à des cas métiers réels

PERSPECTIVES D'ÉVOLUTION

Topic Modeling (LDA/BERTopic)

Alertes Temps Réel

API Reddit Officielle

Deep Learning (BERT/RoBERTa)

Détection Anomalies

Merci pour votre attention