

CHALLENGE MÉDIA-SCAN

Système Intelligent d'Observation et d'Analyse des Médias au Burkina Faso

1. CONTEXTE ET PROBLÉMATIQUE

Contexte institutionnel : Le Conseil Supérieur de la Communication (CSC) régule plus de 100 médias au Burkina Faso. Chaque jour, des milliers d'articles, posts et émissions sont diffusés sur différentes plateformes (web, réseaux sociaux). Le CSC a pour mission de surveiller le paysage médiatique, d'assurer le pluralisme de l'information et de contrôler le respect des obligations réglementaires.

Problèmes rencontrés :

- Surveillance manuelle impossible à l'échelle (100+ médias, milliers de contenus/jour)
- Pas de données objectives sur l'audience et l'influence réelle des médias
- Difficulté à mesurer le pluralisme et la diversité thématique
- Détection tardive des contenus sensibles (haine, désinformation)
- Contrôle manuel des grilles de programmes et obligations (temps, ressources)

Bénéficiaires cibles : CSC (régulation), MTDPC (politique médias), médias eux-mêmes (feedback), citoyens (transparence), chercheurs (études sur les médias).

2. OBJECTIF DU DÉFI

Développer une **plateforme d'intelligence artificielle** capable de **collecter automatiquement** les contenus des médias burkinabè en ligne, de **mesurer leur audience et influence**, **d'analyser les thèmes traités**, de **contrôler le respect des obligations**, et de **déetecter les contenus sensibles**, le tout via un **dashboard interactif** générant des rapports pour le CSC.

Objectifs spécifiques :

1. Collecter automatiquement les contenus d'au moins 5 médias en ligne (articles, posts)
2. Classifier automatiquement les contenus par thématique (politique, économie, sécurité, santé, culture, sport)
3. Mesurer l'audience et calculer un score d'influence composite pour chaque média
4. Contrôler le respect des grilles de programmes et détecter les inactivités (bonus)
5. Déetecter les contenus sensibles (discours de haine, désinformation, toxicité)
6. Créer un dashboard web avec visualisations, statistiques et export de rapports

3. MODULES FONCTIONNELS

MODULES OBLIGATOIRES

MODULE 1 - Collecte Automatique : Web scraping des médias burkinabè (Lefaso.net, FasoPresse, Sidwaya, médias sociaux publics). Extraire : titre, date, contenu, source, URL, engagement (likes, partages, commentaires). Technologies : BeautifulSoup, Scrapy, Selenium.

MODULE 2 - Analyse Thématique : Classification automatique par machine learning. Catégories : Politique, Économie, Sécurité, Santé, Culture, Sport, Autres. Technologies : CamemBERT, FlauBERT, scikit-learn, spaCy.

MODULE 3 - Audience et Influence : Calculer pour chaque média : volume de publications, engagement total (likes + partages + commentaires), portée estimée, score d'influence composite. Générer un classement des médias par impact réel.

MODULE 6 - Dashboard Interactif : Interface web simple affichant toutes les statistiques, classements, graphiques d'évolution, alertes, et permettant l'export de rapports PDF/Excel. Technologies : Streamlit, Dash, Plotly.

MODULES BONUS (Recommandés)

MODULE 4 - Contrôle des Grilles : Vérifier respect des obligations : régularité de publication (activité sur 90 jours minimum), détection de la cessation imminente, alertes sur inactivité prolongée, écarts entre grille déclarée et réalité.

MODULE 5 - Détection Contenus Sensibles : Identifier automatiquement les posts à risque : incitation à la haine, fake news, discours toxique. Utiliser des modèles de détection de toxicité et fact-checking basique.

4. LIVRABLES ATTENDUS

Niveau	Requis minimum
Niveau 1 POC/Potentiel	<ul style="list-style-type: none">• Code Python fonctionnel (collecte + analyse)• Au moins 100 articles collectés de plusieurs médias• Collecte automatique de 10+ médias• Classification ML performance (>80% accuracy)• Système de scores d'influence robuste• Dashboard interactif complet• Base de données structurée• Export de rapports• Module 4 OU 5 implémenté• Tests sur 30 jours de données• Documentation README

5. CRITÈRES D'ÉVALUATION (100 points)

Innovation Technique (30 pts)	Qualité du scraping, robustesse classification ML, pertinence scores influence, détection contenus sensibles, qualité code
Impact et Utilité (25 pts)	Utilité réelle pour CSC, insights pertinents, déployabilité opérationnelle, scalabilité
Viabilité (20 pts)	Coûts infrastructure, maintenabilité, documentation
Approche Frugale (15 pts)	Open source, optimisation ressources, hébergement local possible, pas de dépendances payantes
Présentation (10 pts)	Clarté démo, qualité visuels dashboard, réponses aux questions

6. RESSOURCES ET TECHNOLOGIES

Données disponibles : Médias burkinabè en ligne (Lefaso.net, FasoPresse, Sidwaya, L'Observateur Paalga, AIB, etc.), réseaux sociaux publics (Facebook, X/Twitter pages officielles), liste des médias déclarés au CSC, exemples de grilles de programmes.

Stack recommandé : Python 3.9+, BeautifulSoup/Scrapy (scraping), Selenium (sites dynamiques), CamemBERT/FlauBERT (NLP français), scikit-learn/spaCy (ML), pandas/numpy (data), SQLite/PostgreSQL (DB), Streamlit/Dash (dashboard), Plotly/Seaborn (viz), FastAPI (API bonus).

Support : Mentorat technique MTDPCE et CITADEL pendant le hackathon, accès experts CSC pour questions réglementaires.

7. MÉTRIQUES DE SUCCÈS TECHNIQUES

Composante	Objectif minimum	Objectif excellence
Collecte données	5 médias, 100+ articles	10+ médias, 1000+ articles
Classification thématique	Accuracy >60%	Accuracy >80%
Détection contenus sensibles	Détection basique	Précision >75%
Dashboard	Visualisations basiques	Interactif + export + alertes

8. FORMAT DONNÉES ATTENDU

Exemple de structure pour base de données :

```
{
  "articles": [
    {
      "id": "123",
      "media": "Lefaso.net",
      "titre": "...",
      "date": "2025-11-14",
      "url": "https://...",
      "contenu": "...",
      "categorie": "Politique",
      "engagement": {
        "likes": 45,
        "partages": 12,
        "commentaires": 8
      },
      "sensible": false,
      "toxicite_score": 0.05
    }
  ],
  "medias": [
    {
      "nom": "Lefaso.net",
      "nb_articles": 156,
      "engagement_total": 12450,
      "score_influence": 8.7,
      "rang": 1,
      "actif_90j": true
    }
  ]
}
```

9. PLANNING INDICATIF (72 heures)

Jour 1 (24h) : Développer le scraper, collecter les premières données (100+ articles de 5 médias), commencer l'analyse d'audience.

Jour 2 (24h) : Implémenter l'intelligence artificielle : classification thématique ML, calcul scores d'influence, contrôle conformité (si bonus).

Jour 3 (24h) : Créer le dashboard complet avec toutes les visualisations, intégrer tous les modules, tester, préparer la démo et la documentation.

10. INFORMATIONS PRATIQUES

Durée du hackathon : 72 heures

Partenaires : CSC (Conseil Supérieur de la Communication)

Approche : Focus sur données publiques accessibles (pas de partenariats complexes ou accès payants requis)

Support technique : Équipes MTDPCE et CITADEL disponibles pendant le hackathon, experts CSC pour questions réglementaires

Modalités de soumission : Code source (GitHub public), au moins 100 articles collectés et analysés, dashboard fonctionnel avec visualisations, documentation d'utilisation, vidéo démo (5 min max), présentation finale devant jury.

11. IMPACT ATTENDU

Pour le CSC : Gain de temps considérable dans la surveillance, statistiques objectives sur le paysage médiatique, identification rapide des médias influents, détection précoce de problèmes.

Pour les médias : Feedback sur leur influence réelle, comparaison avec leurs pairs, alertes sur conformité réglementaire.

Pour la démocratie : Meilleure surveillance du pluralisme, transparence sur l'influence médiatique, détection rapide de la désinformation, données objectives pour recherche académique.