Analyse et Stratégie XAI pour le Système de Vérification

1. Introduction

Ce document propose une stratégie pour faire évoluer le système de vérification de crédibilité vers un modèle centré sur l'explicabilité (XAI), en se basant sur les excellents travaux de modélisation UML et ontologique déjà réalisés.

2. Analyse du Système Actuel

Forces :

- **Approche Hybride :** La combinaison de règles logiques et de modèles d'IA est une approche robuste qui reflète l'état de l'art. Les règles sont excellentes pour les cas déterministes (ex: source sur liste noire) et l'IA pour les nuances sémantiques.
- **Fondation Ontologique :** L'ontologie OWL que tu as développée est un atout majeur. Elle formalise le domaine de manière rigoureuse et ouvre la voie à des inférences et des explications encore plus riches à l'avenir.
- **Utilisation de LIME :** L'intégration de LIME pour l'analyse de sentiment est un excellent premier pas vers l'XAI. Tu as identifié le besoin d'expliquer les prédictions des modèles "boîte noire".

Axes d'Amélioration pour l'XAI :

- **La "Boîte Noire" du Score Global :** Le principal obstacle à l'explicabilité se trouve dans la fonction calculate_overall_score. C'est une fonction complexe avec des poids et des heuristiques codés en dur. Si un utilisateur demande "Pourquoi ai-je obtenu 7/10 ?", la seule réponse est "Parce que la somme pondérée des facteurs X, Y et Z a donné ce résultat". Ce n'est pas une explication, c'est une description d'un calcul.
- **Manque de Transparence sur l'Agrégation :** Le système ne montre pas clairement *comment* les différents modules (règles, sentiment, biais, etc.) contribuent au score final. Un marqueur de sensationnalisme a-t-il eu plus d'impact qu'un score de sentiment négatif ? L'utilisateur ne peut pas le savoir.
- **Interface Utilisateur Limitée :** L'interface actuelle est un point d'entrée, mais elle n'est pas conçue pour *présenter* des explications. Une véritable interface XAI doit visualiser le raisonnement.

3. Ma Recommandation : Une Architecture Basée sur les "Constats" (Findings)

Je propose de changer de paradigme. Au lieu de voir le système comme une machine qui produit *un score*, nous allons le voir comme un système qui produit **une liste de**

constats sur l'information soumise.

Un **constat** est un objet structuré qui contient :

- source: Quelle partie du système a fait ce constat ? (Ex: "Règle : Âge du Domaine", "Modèle IA : Analyse de Biais").
- description: Une description lisible du constat. (Ex: "Le domaine du site web est très récent (moins de 90 jours).").
- impact: Un score numérique représentant l'influence de ce constat sur la crédibilité finale (positif ou négatif).
- explanation (optionnel): Des détails supplémentaires, comme les mots-clés qui ont déclenché une règle ou une explication LIME pour un modèle.

Le flux devient :

1. L'utilisateur soumet une information.
2. Chaque module d'analyse (règles, NLP) ne retourne plus de scores partiels, mais une **liste de constats**.
3. Le système agrège tous les constats de tous les modules.
4. Le **score global** est simplement la **somme des impacts de tous les constats**, normalisée sur une échelle (ex: 0 à 10).
5. Le **backend retourne à l'interface** le score global ET la liste complète des constats.

Avantages de cette approche :

- **Transparence Totale :** Pour expliquer le score final, il suffit de montrer la liste des constats. L'utilisateur voit exactement ce qui a contribué positivement et négativement.
- **Modularité :** Ajouter une nouvelle règle ou un nouveau modèle d'IA revient à créer une nouvelle fonction qui génère des constats. Cela n'impacte pas une complexe fonction de calcul de score.
- **Explicabilité Native :** Le système est conçu autour de l'explication. L'explication n'est pas une couche ajoutée après coup, elle est le cœur du résultat.

4. Plan d'Action

1. **Remanier le backend Flask (app.py)** pour implémenter cette nouvelle architecture à base de constats.
2. **Créer une nouvelle interface web (index.html)** capable de recevoir cette structure de données et de la visualiser de manière claire et interactive.

Nous allons maintenant mettre en œuvre ce plan.