UQAC

Université du Québec
à Chicoutimi

# Jalon 1

8INF970 - Atelier pratique en cybersécurité II

Fehmi Jaafar Samuel Desbiens

Justin Bossard Mattéo Gouhier Paul Mathé Samuel Plet Léo Raclet

26 janvier 2026

# Table des matières

# Partie 1

# Introduction

TODO

# Partie 2

# État de l'art

# Partie 3

# Truth Sleuth & Trend Bender (Logé & Ghori, 2025)

Logé, C., & Ghori, R. (2025). Truth Sleuth & Trend Bender : AI Agents to fact-check YouTube videos & influence opinions. arXiv. https://doi.org/10.48550/arXiv.2507.10577

Cette recherche se concentre sur l'automatisation du fact-checking pour les contenus multimédias (YouTube), avec deux agents spécialisés.

## 3.1 Truth Sleuth - L'Agent de Vérification des Faits

Truth Sleuth est conçu pour vérifier automatiquement le contenu de vidéos Youtube. D'abord il extrait les affirmations clés de la vidéo analysée. Ensuite, Truth Sleuth utilise la technologie **Retrieval-Augmented Generation (RAG)** pour vérifier les faits. C'est à dire qu'il va chercher des données de confiances (souvent classées en 3 catégories de confiances : Gold, silver et bronze) en rapport avec les informations à vérifier. Cette méthode utilise les moteurs de recherche et d'autres sources fiables.

> *« Truth Sleuth extracts claims from a YouTube video, uses a Retrieval-Augmented Generation (RAG) approach drawing on sources like Wikipedia, Google Search, Google FactCheck - to accurately assess their veracity »* (Logé & Ghori, 2025).

Enfin, l'agent rédige une réponse grâce à ses connaissances et les informations contextuelles ajoutées ; afin d'indiquer si chaque affirmation est vraie ou non et rajoute des sources et des explications détaillées. L'avantage de cette méthode est de réduire les hallucinations liées aux IA, en poussant le modèle à utiliser des sources vérifiables, plutôt que de laisser l'IA répondre librement avec ses biais.

## 3.2 Trend Bender - L'Agent de Persuasion

Trend Bender génère des commentaires dans le but de convaincre les spectateurs, dotés d'un mécanisme d'auto-évaluation, donc l'agent génère une réponse qu'il va lui même reprendre en entrée d'un nouveau prompt afin de l'améliorer et de le corriger.

> *« With a carefully set up self-evaluation loop, this agent is able to iteratively improve its style and refine its output. »* (Logé & Ghori, 2025).

Dans cet article, ce qui nous intéresse par rapport à notre sujet est donc de voir que l'utilisation du RAG dans la détection de fake news est d'actualité car l'article date de 2025. Et nous a fait chercher des informations sur la méthode de classification des données cherchées, afin de détecter si elles sont de qualités ou non.

---

# Partie 4

# Multi-Agent Systems for Misinformation (Gautam, 2025)

Gautam, Aditya. “Multi-agent Systems for Misinformation Lifecycle : Detection, Correction and Source Identification.” arXiv, 23 mai 2025, arxiv.org/abs/2505.17511

Contrairement à l’approche avec deux agents précédente, cette solution propose une méthode à cinq agents autonomes, chacun se concentrant sur un aspect spécifique du problème.

## 4.1 Les Cinq Agents Spécialisés

- **La Classification :** va trier et catégoriser le type de désinformation détecté.
- **L’indexation :** gère et actualise l’infrastructure de vérification en tenant une archive des données vérifiées et fiables.
- **L’Extraction :** se concentre sur la collecte et la traçabilité des preuves. Il récupère les preuves vérifiées et va chercher l’origine de la désinformation.
- **La Correction :** corrige les informations et effectue des modifications basées sur les preuves identifiées.
- **La Vérification :** vérifie la qualité finale des corrections apportées.

> *« In contrast to single-agent or monolithic architectures, our approach employs five specialized agents : an Indexer agent […], a Classifier agent […], an Extractor agent […], a Corrector agent […] and a Verification agent for validating outputs and tracking source credibility. »* (Gautam, 2025).

## 4.2 Avantages de l’approche à 5 Agents

Cette architecture possède plusieurs avantages. L’amélioration de RAG provient de l’ajout d’agents pour diviser la tâche, ce qui permet une optimisation indépendante de la tâche.

> *« By decomposing the misinformation lifecycle into specialized agents our framework enhances scalability, modularity, and explainability. »* (Gautam, 2025).

Il est aussi plus facile de déterminer d’où provient une erreur ou qu’est-ce qui devrait être amélioré, car il est plus facile de voir quel agent a pris quelle décision. La modularité permet l’amélioration indépendante de chaque agent. Enfin, la réduction des biais vient de la séparation des responsabilités.

## 4.3 Défis et Limitations

Malgré ses avantages, le coût d'utilisation est élevé, puisque la mise en marche de cinq agents autonomes implique des coûts importants. La latence est également problématique, car avoir 5 agents augmente grandement le nombre d'interactions. Dans un environnement où la vitesse de détection est un enjeu majeur, cette latence peut devenir problématique.

# Partie 5

# Notre solution

Dans cette section, nous présentons un aperçu général de notre projet, ainsi que des outils et technologies que nous utiliserons pour le réaliser.

## 5.1 Nom et logo

Notre projet s'intitulera **VerifAI** et sera représenté par le logo ci-dessous :


FIGURE 5.1 : logo

## 5.2 Outils et Technologies de développement

Pour les outils de développement, nous utiliserons :

- **Python** : Langage principal de programmation, choisi pour son adéquation avec les fonctionnalités IA, sa flexibilité et sa maîtrise par l'ensemble de l'équipe.
- **PyTorch** : Une bibliothèque Python spécialisée pour la conception, l'entraînement et l'utilisation de modèles d'IA.
- **FastAPI** : Une bibliothèque Python pour implémenter le serveur web, spécialement conçue pour gérer les requêtes asynchrones, ce qui est idéal pour l'inférence de modèles d'IA.

## 5.3 Outils de gestion

Pour la gestion de projet, nous communiquons via **WhatsApp**. Le code et les sources des rapports sont partagées sur GitHub.

# Partie 6

# Choix des Données

## 6.1 Stratégie de Sélection des Données

L'entraînement d'une IA à pour but de donner de nombreux exemples de ce que l'on cherche afin quelle apprenne à le reconnaître. Dans notre cas, les fake news et les informations vérifiées. Pour permettre à une IA de fonctionner aux mieux dans un cas précis, il faut donc l'entraîner à l'aide d'un jeu de données divers et adapté à son utilisation. Nous avons identifié deux approches pour constituer une base de données la plus adaptées : le scraping de sources existantes et l'utilisation d'un jeu de données publics.

## 6.2 Approche par Scraping

Le scraping est une méthode qui consiste à parcourir automatiquement des pages web et en extraire ses données. Dans notre cas, deux sources pourraient s'appliquer :

**Sources de Contenu Vérifié et Fiable :**

— Regrouper plusieurs sites d'informations reconnus : Ces sources constituent une base de données d'articles vérifiés, permettant d'établir des références que l'on sait justes pour un modèle d'apprentissage. Permettant d'avoir un grand nombre de données tout en étant certains de l »exactitude des informations. Par exemple, c'est ce que fait la plateforme de vérification Vera, qui regroupe environ 300 sites de confiance.

**Sources de Contenu Trompeur :**

— Site regroupant des fake News : Afin d'avoir des informations vérifiées, nous pourrions directement aller sur une plateformes qui références les fake news ou les informations qui peuvent prêter à la confusion, le but de ces plateformes/rubriques étant de valider la véracité ou non des informations. Par exemple, la rubrique Vrai/Faux de FranceInfo. La plateforme de FranceInfo maintient une rubrique spécialisée qui analyse les affirmations fausses ou trompeuses. Cette source est très intéressante car elle fournit l'identification des fausses informations, mais aussi une analyse et des explications détaillées des raisons pour lesquelles ces affirmations sont erronées, tout en étant rattachée à un organisme fiable.

## 6.3 Jeux de Données Publics Reconnus

Au-delà du scraping, permettant de se créer un jeu de données, nous pourrions aussi nous appuyer sur des jeux de données publics existants, largement utilisés, testés et reconnus :

— **Webz.io :** Une base de données indexant des millions d'articles web en anglais, avec des métadonnées riches, permettant une analyse du contexte.

— **ISOT Fake News Dataset :** Un jeu de données spécialisé contenant plusieurs milliers d'articles validés comme vrais ou faux, avec des étiquetages de très bonne qualité.
— **FakeNewsNet :** Un ensemble de données complet incluant non seulement le contenu des articles.

Ces jeux de données permettent d'avoir des articles vérifiés ainsi que des informations supplémentaires permettant d'avoir du contexte ou de distinguer certains articles, qui pourront permettre à l'IA de gagner en précision.

# Partie 7

# Expérimentations possibles

# Partie 8

# Conclusion

TODO

# Partie 9

# Références