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

27 janvier 2026

Table des matières

Partie 1

Introduction

TODO

Partie 2

État de l'art

2.1 Survey of fake news detection using machine intelligence approach (Pal et al., 2023)

Pal, A., Pranav, & Pradhan, M. (2023). Survey of fake news detection using machine intelligence approach. *Data & Knowledge Engineering*. https://doi.org/10.1016/j.datak.2022.102118

Cet article propose une vue d'ensemble complète des approches de détection automatique de fake news basées sur l'intelligence artificielle, en mettant l'accent sur la diversité des signaux exploités.

2.1.1 Les trois familles de méthodes

Les auteurs classifient les approches en trois catégories distinctes :

- **Contenu textuel** : Analyse des caractéristiques lexicales, stylistiques et des représentations vectorielles du texte.
- **Contexte social** : Exploitation des profils utilisateurs, de la structure des interactions et des graphes de diffusion.
- **Approches hybrides** : Combinaison de plusieurs sources de signaux pour une robustesse accrue.

2.1.2 Performance et Comparaison

Le survey passe en revue les principaux jeux de données publics (Kaggle et ISOT) et compare les performances des modèles. Il met en opposition les modèles d'apprentissage automatique "classiques" (SVM, régressions, arbres de décision) aux architectures d'apprentissage profond plus récentes (CNN, RNN, Transformers), démontrant la supériorité de ces dernières sur les tâches complexes.

2.1.3 Défis identifiés

Les auteurs soulignent plusieurs obstacles majeurs : la difficulté de généraliser un modèle entraîné sur un domaine spécifique (domaine shift), la sensibilité à la qualité de l'annotation des données, et surtout le **manque d'explicabilité**. Ce dernier point est crucial car l'effet "boîte noire" limite la confiance que les utilisateurs finaux peuvent accorder aux systèmes de détection.

2.2 A systematic review of multimodal fake news detection on social media using deep learning models (Nasser et al., 2025)

Nasser, M., Arshad, N. I., & Sugathan, S. K. (2025). A systematic review of multimodal fake news detection on social media using deep learning models. *Results in Engineering*. https://doi.org/10.1016/j.rineng.2025.104752

Cette revue systématique s'intéresse spécifiquement aux approches **multimodales** sur les réseaux sociaux, partant du constat que la désinformation moderne combine intrinsèquement texte et image.

2.2.1 Architectures de fusion multimodale

Les auteurs analysent comment les architectures de deep learning fusionnent les modalités :

— **Images** : Utilisation de réseaux convolutionnels (CNN) ou de Vision Transformers.
— **Texte** : Utilisation de RNN ou de Transformers (BERT, etc.).
— **Métadonnées** : Intégration de modules supplémentaires pour les signaux comportementaux.

2.2.2 Résultats et efficacité

Les résultats montrent que les modèles multimodaux surpassent systématiquement les approches uniquement textuelles, en particulier sur des plateformes riches en visuels comme Twitter, Facebook ou Weibo. L'image joue souvent un rôle de catalyseur dans la diffusion de la fausse information.

2.2.3 Limites et contraintes techniques

La revue met en évidence des freins importants à l'adoption pratique : la complexité et le coût de calcul élevés, le besoin de jeux de données annotés massifs et cohérents, et la difficulté d'aligner l'explicabilité entre le texte et l'image (savoir quelle partie de l'image a déclenché la décision). Ces contraintes valident notre choix de privilégier une solution explicable et éducative plutôt qu'une performance brute opaque.

2.3 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.

2.3.1 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 confiance (souvent classées en trois catégories : gold, silver et bronze) en rapport avec les informations à vérifier. Cette méthode interroge des moteurs de recherche et d'autres sources considérées comme 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 aux informations contextuelles ajoutées, afin d'indiquer si chaque affirmation est vraie ou non, et fournit des sources et des explications détaillées.

L'avantage de cette méthode est de réduire les hallucinations liées aux IA, en poussant le modèle à s'appuyer sur des sources vérifiables plutôt que de laisser l'IA répondre librement avec ses biais.

2.3.2 Trend Bender - L'Agent de Persuasion

Trend Bender génère des commentaires dans le but de convaincre les spectateurs, en étant doté d'un mécanisme d'auto-évaluation. L'agent génère une première réponse, qu'il réinjecte ensuite comme entrée dans un nouveau prompt afin de l'améliorer et de la 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 particulièrement pour notre projet est de constater que l'utilisation du RAG dans la détection de fake news et le fact-checking automatisé est très récente et active (article de 2025). Il met aussi en avant l'importance de la qualité et de la classification des sources externes utilisées pour la vérification, ce qui rejoint notre réflexion sur la sélection de données fiables et la façon d'expliquer les résultats à l'utilisateur.

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

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

Contrairement à l'approche avec deux agents de Logé & Ghori, cette solution propose une méthode à cinq agents autonomes, chacun se concentrant sur un aspect spécifique du cycle de vie de la désinformation.

2.4.1 Les cinq agents spécialisés

— **La Classification** : trie et catégorise 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, récupère les preuves vérifiées et remonte à 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 et suit la crédibilité des sources.

> *« 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)

2.4.2 Avantages de l'approche à 5 agents

Cette architecture possède plusieurs avantages. L'amélioration de RAG et, plus généralement, du pipeline de traitement, provient de l'ajout d'agents pour diviser la tâche, ce qui permet une optimisation indépendante de chaque étape.

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

Il est ainsi plus facile de déterminer d'où provient une erreur ou quel module doit être amélioré, puisqu'on peut associer chaque décision à un agent. La modularité permet l'amélioration indépendante de chaque composant. Enfin, la séparation des responsabilités contribue à réduire certains biais, en limitant la concentration de toutes les décisions dans un seul modèle monolithique.

2.4.3 Défis et limitations

Malgré ses avantages, le coût d'utilisation est élevé, puisque la mise en œuvre de cinq agents autonomes implique des coûts de calcul et d'infrastructure importants. La latence est également problématique, car la multiplication des interactions entre agents augmente la durée totale du traitement. Dans un environnement où la vitesse de détection est un enjeu majeur (par exemple sur les réseaux sociaux durant une campagne électorale), cette latence peut devenir un frein à l'adoption.

Partie 3

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.

3.1 Identité Visuelle et Marque

3.1.1 Nom : VerifAI

Le nom **VerifAI** est un mot-valise fusionnant « Vérification » et « AI » (Intelligence Artificielle). Il évoque instantanément notre mission centrale : utiliser la puissance de l'automatisation pour rétablir la vérité. Ce choix d'un nom court et explicite a été fait pour assurer une mémorisation facile et une portée internationale.

3.1.2 Symbolique du Logo

Notre identité visuelle a été conçue pour refléter nos valeurs technologiques et éthiques :

- **Le V et la Loupe** : La structure en "V" combinée à la loupe centrale symbolise l'investigation minutieuse et la validation de l'information (le "Fact-Checking").
- **Les Circuits Imprimés** : Intégrés dans la typographie, ils représentent le cœur technologique de notre solution et l'usage d'algorithmes avancés.
- **Le Cerveau (au centre)** : Symbolise l'intelligence (artificielle et cognitive) au service de la cybersécurité.
- **La Palette de Couleurs** : L'utilisation dominante du bleu et du blanc n'est pas anodine ; en cybersécurité et en communication, le bleu inspire la confiance, la stabilité et la sécurité technologique.

3.2 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.

FIGURE 3.1 : logo

3.3 Outils de gestion

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

Partie 4

Choix des Données

4.1 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.

4.2 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.

4.3 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 5

Expérimentations possibles

Partie 6

Conclusion

TODO

Partie 7

Références