



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

1	Introduction	3
2	État de l'art	4
2.1	Survey of fake news detection using machine intelligence approach (Pal et al. 2023)	4
2.1.1	Les trois familles de méthodes	4
2.1.2	Performance et Comparaison	4
2.1.3	Défis identifiés	4
2.2	A systematic review of multimodal fake news detection on social media using deep learning models (Nasser et al. 2025)	4
2.2.1	Architectures de fusion multimodale	4
2.2.2	Résultats et efficacité	5
2.2.3	Limites et contraintes techniques	5
2.3	Truth Sleuth & Trend Bender (Logé et Ghori 2025)	5
2.3.1	Truth Sleuth - L'Agent de Vérification des Faits	5
2.3.2	Trend Bender - L'Agent de Persuasion	5
2.4	Multi-Agent Systems for Misinformation (Gautam 2025)	6
2.4.1	Les cinq agents spécialisés	6
2.4.2	Avantages de l'approche à 5 agents	6
2.4.3	Défis et limitations	6
2.5	Automatic deception detection : Methods for finding fake news (Conroy et al. 2016)	6
2.5.1	Typologie des méthodes de détection	6
2.5.2	Hybridation des méthodes pour une détection robuste	7
2.6	Beyond News Contents : The Role of Social Context for Fake News Detection (Shu et al. 2019)	7
2.6.1	Le Framework TriFN : Une approche tripartite	7
2.6.2	Résultats et détection précoce	7
2.6.3	Pertinence pour VerifAI	8
2.7	Fake news detection : A survey of graph neural network methods (Phan et al. 2023)	8
2.7.1	La domination des GCN	8
2.7.2	Quatre approches de détection	8
2.7.3	Défis pour la détection précoce	8
3	Notre solution	9
3.1	Identité Visuelle et Marque	9
3.1.1	Nom : VerifAI	9
3.1.2	Symbolique du Logo	9
3.2	Outils et Technologies de développement	10
3.3	Outils de gestion	10
4	Choix des Données	11
4.1	Stratégie de Sélection des Données	11
4.2	Approche par Scraping	11

4.3 Jeux de Données Publics Reconnus	11
5 Expérimentations possibles	13
6 Conclusion	14
Références	15

Partie 1

Introduction

La désinformation, ou fake news, est une menace croissante pour la démocratie, notamment lors des élections. Au Québec, un tiers des électeurs y ont été exposés en 2022, et les journalistes en font une priorité d'inquiétude (Gouvernement du Québec, 2019 ; Sauvé, 2019). Face à ce défi, notre projet propose une solution innovante en deux volets :

1. **Détection** automatisée par IA, combinant des algorithmes récents et des données temporelles pour une identification plus rapide et précise.
2. **Explication** transparente des résultats via des modèles de langage (LLM), afin d'éduquer le public et de renforcer sa capacité à distinguer le vrai du faux.

L'objectif de ce projet est de proposer une solution technologique et innovante pour lutter efficacement contre la désinformation tout en responsabilisant les citoyens.

Partie 2

État de l'art

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

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)

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 et Ghori 2025)

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)

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.

2.5 Automatic deception detection : : Methods for finding fake news (Conroy et al. 2016)

Cette recherche passe en revue les technologies actuelles qui jouent un rôle clé dans l'adoption et le développement de la détection des fake news.

2.5.1 Typologie des méthodes de détection

L'article ([AutomaticDeceptionDetection2015 ?](#)) propose une classification claire des approches existantes en deux grandes catégories :

- **Approches linguistiques** : Analyse du contenu textuel pour identifier des “fuites linguistiques” (mots, syntaxe, sémantique, discours) associées à la tromperie, comme par exemple l’usage excessif de pronoms, de négations, ou d’émotions extrêmes pouvant trahir un texte mensonger.
- **Approches par réseaux** : Exploitation des métadonnées, des comportements sur les réseaux sociaux, ou des bases de connaissances structurées (comme DBpedia) pour évaluer la crédibilité d’une information, notamment par la vérification de faits via des graphes de connaissances ou l’analyse des profils d’utilisateurs suspects.

2.5.2 Hybridation des méthodes pour une détection robuste

Les auteurs plaident pour un système hybride intégrant :

1. **Linguistique** (analyse de surface, syntaxe profonde, sémantique, rhétorique) pour capter les indices textuels de tromperie.
2. **Réseaux** (comportements en ligne, liens entre sources, vérification par bases de données) pour ajouter une dimension contextuelle et sociale.
3. **Machine Learning** (SVM, Naïve Bayes) pour entraîner des classificateurs à partir de données annotées.

Les auteurs pointent d’ailleurs que la combinaison de l’analyse syntaxique, pour détecter des incohérences, et des graphes de connaissances, pour vérifier des faits, a montré des taux de précision allant jusqu’à 91% dans certains cas.

2.6 Beyond News Contents : The Role of Social Context for Fake News Detection (Shu et al. 2019)

Cette recherche part du constat fondamental que la détection basée uniquement sur le contenu textuel est souvent inefficace, car les fausses nouvelles sont intentionnellement rédigées pour imiter le style des vraies nouvelles afin de tromper les lecteurs.

2.6.1 Le Framework TriFN : Une approche tripartite

Les auteurs proposent le modèle TriFN (Tri-Relationship Fake News detection) qui modélise simultanément trois entités et leurs interactions :

- Les Éditeurs (Publishers) : Le modèle analyse le biais partisan des éditeurs. Un éditeur ayant un fort biais partisan (extrême gauche ou droite) est statistiquement plus enclin à publier des fausses nouvelles qu’un média grand public neutre.
- Les Utilisateurs : Le modèle intègre les interactions sociales et le score de crédibilité des utilisateurs. Les utilisateurs peu crédibles ou malveillants tendent à partager davantage de désinformation que les utilisateurs fiables.
- Le Contenu : L’analyse textuelle classique reste présente mais est enrichie par les deux vecteurs précédents.

2.6.2 Résultats et détection précoce

Les expérimentations menées sur les jeux de données FakeNewsNet (BuzzFeed et PolitiFact) démontrent deux points cruciaux pour notre projet :

- Supériorité du contexte social : Les fonctionnalités basées sur le contexte social (qui publie, qui partage) se révèlent plus performantes que celles basées uniquement sur le texte (comme l’analyse linguistique LIWC).
- Détection précoce (Early Detection) : Le modèle TriFN parvient à atteindre un score F1 supérieur à 80% moins de 48 heures après la publication d’une nouvelle, prouvant qu’il est possible de détecter une fake news tôt dans son cycle de diffusion, même avec des interactions limitées.

2.6.3 Pertinence pour VerifAI

Cette étude valide notre hypothèse selon laquelle l'inclusion de données temporelles et contextuelles est indispensable. Elle suggère que pour dépasser l'état de l'art, notre solution ne doit pas analyser le texte en seulement, mais prendre en compte l'écosystème de sa diffusion (source et partages).

2.7 Fake news detection : A survey of graph neural network methods (Phan et al. 2023)

Ce survey récent justifie l'utilisation des structures de graphes pour modéliser la complexité des réseaux sociaux, là où les méthodes d'apprentissage classiques échouent souvent à capturer les relations d'interdépendance entre les utilisateurs et les contenus.

2.7.1 La domination des GCN

L'étude analyse 27 articles majeurs et révèle que les Graph Convolutional Networks (GCN) sont la technique dominante (utilisée dans plus de 74% des cas étudiés). Contrairement à l'analyse de texte isolée, les GCN permettent de propager l'information à travers le réseau : si un utilisateur malveillant partage une news, cette information "contamine" le score de fiabilité de la news via les arêtes du graphe.

2.7.2 Quatre approches de détection

Les auteurs classifient les méthodes de détection via GNN en quatre catégories :

- Basée sur la connaissance (Knowledge-based) : Vérification des faits via des graphes de connaissances externes (Fact-checking).
- Basée sur le style (Style-based) : Analyse des intentions trompeuses via le style d'écriture.
- Basée sur le contexte (Context-based) : Analyse de la crédibilité des utilisateurs et des éditeurs.
- Basée sur la propagation (Propagation-based) : Analyse de la cascade de diffusion de l'information (qui partage quoi et quand).

2.7.3 Défis pour la détection précoce

L'article soulève un point critique pour notre projet : la détection précoce (Early Detection). Les méthodes basées sur la propagation sont très performantes mais nécessitent que la news ait déjà été partagée massivement. Pour une détection rapide (avant la viralité), les auteurs suggèrent de privilégier les approches basées sur le contexte (analyser la source dès la publication) ou d'utiliser des GNN hétérogènes capables de traiter simultanément le texte, l'image et le profil utilisateur dès les premiers instants.

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

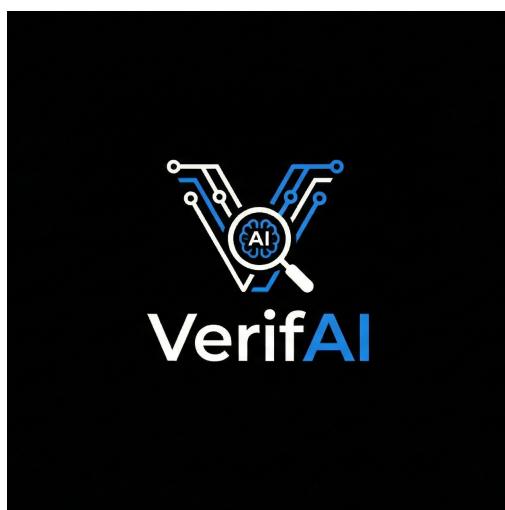


FIGURE 3.1 : logo

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.
- **HuggingFace** - Une plateformes permettant l'accès à de nombreux modèles IA, ainsi que leur entraînement et utilisation.

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

Ce Jalon 1 a permis de définir l'architecture technique et l'identité de VerifAI. Les prochaines étapes (Jalon 2) se concentreront sur le développement du prototype et l'entraînement des premiers modèles sur les données identifiées.

Références

- Conroy, Nadia K., Victoria L. Rubin, et Yimin Chen. 2016. « Automatic Deception Detection : Methods for Finding Fake News ». *Proceedings of the Association for Information Science and Technology* 52 (1) : 1-4. <https://doi.org/10.1002/pra2.2015.145052010082>.
- Gautam, Aditya. 2025. « Multi-Agent Systems for Misinformation Lifecycle : Detection, Correction And Source Identification ». Prépublié mai 23. <https://doi.org/10.48550/arXiv.2505.17511>.
- Logé, Cécile, et Rehan Ghori. 2025. « Truth Sleuth and Trend Bender : AI Agents to Fact-Check YouTube Videos and Influence Opinions ». Prépublié juillet 16. <https://doi.org/10.48550/arXiv.2507.10577>.
- Nasser, Maged, Noreen Izza Arshad, Abdulalem Ali, et al. 2025. « A Systematic Review of Multimodal Fake News Detection on Social Media Using Deep Learning Models ». *Results in Engineering* 26 (juin) : 104752. <https://doi.org/10.1016/j.rineng.2025.104752>.
- Pal, Aishika, Pranav, et Moumita Pradhan. 2023. « Survey of Fake News Detection Using Machine Intelligence Approach ». *Data & Knowledge Engineering* 144 (mars) : 102118. <https://doi.org/10.1016/j.datak.2022.102118>.
- Phan, Huyen Trang, Ngoc Thanh Nguyen, et Dosam Hwang. 2023. « Fake News Detection : A Survey of Graph Neural Network Methods ». *Applied Soft Computing* 139 (mai) : 110235. <https://doi.org/10.1016/j.asoc.2023.110235>.
- Shu, Kai, Suhang Wang, et Huan Liu. 2019. « Beyond News Contents : The Role of Social Context for Fake News Detection ». *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (New York, NY, USA), WSDM '19, janvier 30, 312-20. <https://doi.org/10.1145/3289600.3290994>.