Détection automatique de *fake news*
Apprentissage Automatique et ChatGPT

État de l'Art

CADET Florent

15 mars 2025

RÉSUMÉ

L'augmentation rapide des contenus trompeurs en ligne, souvent désignés par le terme *fake news*, représente un défi majeur pour la société contemporaine. Cette étude examine les diverses approches de détection automatique des *fake news* à travers l'apprentissage automatique et les grands modèles de langage (LLMs). Les méthodes classiques telles que les Machines à Vecteurs de Support (SVM) et les classificateurs Naïve Bayes sont comparées à des techniques de deep learning, comme les ; réseaux de neurones convolutifs (CNN) et les transformers comme BERT et FakeBERT. Les LLMs, montrent des limites en tant que détecteurs autonomes mais offrent des contributions significatives lorsqu'ils sont intégrés avec des modèles spécialisés. Cet état de l'art cherche a mettre en lumière les défis techniques et éthiques liés à la détection des *fake news*, tels que les biais des données d'apprentissage et la robustesse des modèles face aux contenus évolutifs. Des axes de recherche futurs incluent l'intégration de méthodes multi-modales et de systèmes hybrides combinant IA et vérification humaine. Cette revue de la littérature propose des solutions pour améliorer la précision, la transparence et l'éthique des systèmes de détection automatiques, soulignant l'importance de collaborations interdisciplinaires pour lutter efficacement contre la désinformation.

ABSTRACT

The rapid increase in misleading online content, often referred to as *fake news*, represents a major challenge for contemporary society. This study examines various approaches to the automatic detection of *fake news* through machine learning and large language models (LLMs). Classical methods such as Support Vector Machines (SVMs) and Naïve Bayes classifiers are compared with deep learning techniques, such as ; convolutional neural networks (CNNs) and transformers like BERT and FakeBERT. LLMs, show limitations as stand-alone detectors but offer significant contributions when integrated with specialized models. This state-of-the-art paper seeks to highlight the technical and ethical challenges associated with the detection of *fake news*, such as the biases of training data and the robustness of models in the face of evolving content. Future research directions include the integration of multi-modal methods and hybrid systems combining AI and human verification. This literature review proposes solutions to improve the accuracy, transparency and ethics of automatic detection systems, underlining the importance of interdisciplinary collaborations to effectively combat misinformation.

1 Introduction

1.1 Contexte et Motivation

L'essor des plateformes numériques et des réseaux sociaux a profondément modifié la manière dont l'information est diffusée et consommée. Toutefois, cette démocratisation de la publication en ligne s'accompagne d'une prolifération de contenus trompeurs, intentionnellement erronés ou biaisés, souvent désignés sous le terme de *fake news*. Ces fausses informations ont un impact significatif sur les opinions publiques, les processus politiques et la confiance dans les médias.

En France, les fake news posent un défi particulier. Selon une étude, un site de fake news en France a généré plus de 11 millions d'interactions par mois, soit cinq fois plus que certaines marques de presse bien établies [7]. De plus, 75% des jeunes Français estiment que la politique est le sujet le plus susceptible d'être ciblé par des fake news, suivi par la politique internationale et les potins [11]. Les réseaux sociaux comme Facebook et Twitter sont principalement responsables de la propagation des fake news en France, avec de nombreux Français admettant avoir cru ou partagé des informations fausses [12, 7].

Ces chiffres illustrent l'ampleur du problème des fake news en France et leur capacité à influencer les perceptions et les comportements à l'échelle mondiale.

Les premières approches de détection des *fake news* reposaient sur des techniques classiques d'apprentissage automatique, telles que les machines à vecteurs de support (SVM) et les classificateurs bayésiens naïfs, qui cherchaient à identifier des motifs linguistiques et des indices sémantiques révélateurs de tromperie [4]. Avec les avancées en intelligence artificielle, des modèles basés sur l'apprentissage profond, notamment les réseaux neuronaux hiérarchiques attentionnels (3HAN) [10] et les modèles de type BERT [6], ont permis de capturer des caractéristiques linguistiques plus sophistiquées.

Récemment, l'impact des grands modèles de langage (LLMs) sur la détection des *fake news* a été étudié. Des recherches montrent que, bien que les LLMs puissent générer du texte convaincant et offrir des analyses riches en contexte, ils ne surpassent pas toujours les modèles spécialisés entraînés sur des jeux de données annotés [5]. Ces études suggèrent que les LLMs peuvent servir d'outils d'assistance aux détecteurs de *fake news* plutôt que d'être utilisés comme classificateurs autonomes.

1.2 Objectifs de l'État de l'Art

Ce document vise à fournir une revue des travaux de recherche consacrés à la détection automatique des *fake news*, en mettant l'accent sur les méthodes d'apprentissage automatique et l'apport des LLMs comme BERT. Plus précisément, nous nous fixons les objectifs suivants :

- *Présenter les approches classiques et modernes :* Examiner l'évolution des techniques de classification, en comparant les modèles traditionnels (SVM, Naïve Bayes) aux approches plus récentes basées sur l'apprentissage profond.
- *Évaluer l'apport des LLMs dans la détection des fake news :* Analyser comment ces modèles peuvent compléter les méthodes existantes et quelles sont leurs forces et leurs limitations.
- *Analyser les méthodes d'évaluation :* Passer en revue les jeux de données disponibles, les métriques utilisées et les protocoles d'évaluation appliqués dans la littérature.
- *Discuter des défis actuels et des perspectives futures :* Identifier les limitations des méthodes actuelles et proposer des pistes d'amélioration pour renforcer la fiabilité et l'éthique des systèmes de détection.

1.3 Structure du Document

Ce document est structuré comme suit :

- *Chapitre 2 : Modèles d'apprentissage automatique pour la détection de fake news* — Présentation des différentes approches utilisées dans la détection des *fake news*, des modèles classiques aux techniques de deep learning.
- *Chapitre 3 : Applications des LLMs*—Analyse des capacités des modèles génératifs dans la détection des *fake news*, accompagnée d'études de cas et d'évaluations expérimentales.
- *Chapitre 4 : Collection de test et méthodologies d'évaluation*—Présentation des jeux de données utilisés dans les recherches en détection des *fake news* et des critères de performance des modèles.
- *Chapitre 5 : Défis et limitation*—Discussion des principaux enjeux techniques et sociétaux liés à la détection automatique des *fake news* et des limitations liées.
- *Chapitre 6 : Perspectives et Travaux Futurs*—Analyse des axes de recherche émergents pour améliorer la détection des *fake news*.
- *Chapitre 7 : Conclusion*—Synthèse des résultats, implications théoriques et pratiques, et recommandations pour les recherches futures.

1.4 Contribution et Importance

Cette revue contribue à l'avancement des connaissances sur la détection des *fake news* en retraçant l'évolution des approches et en mettant en lumière les avancées en intelligence artificielle. Elle évalue l'impact des grands modèles de langage et identifie leurs limites. Enfin, elle propose des perspectives d'amélioration pour développer des systèmes de détection plus fiables et éthiques.

2 Etat de l'art en apprentissage automatique

La détection automatique des *fake news* a considérablement évolué, passant de simples méthodes de classification de texte à des architectures complexes de deep learning. Ce chapitre explore les différentes techniques utilisées, depuis les approches classiques comme les modèles bayésiens et les SVM jusqu'aux modèles avancés tels que BERT et les architectures de type Transformer.

2.1 Méthodes classiques

Les premières approches reposaient sur des modèles d'apprentissage automatique qui analysaient les caractéristiques textuelles des articles.

2.1.1 N-grams et TF-IDF

Les représentations basées sur les N-grams (unigrammes, bigrammes) et les vecteurs TF-IDF ont été largement utilisées pour modéliser la structure textuelle des *fake news*. Ces techniques permettent d'identifier les séquences de mots les plus caractéristiques des articles trompeurs et ont constitué une base essentielle pour les premiers systèmes de détection [4].

2.1.2 Machines à Vecteurs de Support (SVM)

Les SVM ont montré leur efficacité en classifiant les articles en fonction de leur contenu linguistique. En combinaison avec des représentations TF-IDF, ces modèles ont permis d'atteindre des performances notables dans les premières approches de détection automatique des *fake news*. Par exemple, selon Conroy, Rubin et Chen [4], des techniques de syntaxe profonde et d'analyse sémantique ont été combinées avec succès pour atteindre une précision de 85-91 % dans la détection de *fake news*. De plus, l'utilisation de classificateurs SVM a

démontré une performance de 86 % dans la détection des spams d'opinion négatifs, soulignant la robustesse de cette approche dans des contextes variés.[4].

2.1.3 Naïve Bayes

Le classificateur Naïve Bayes exploite la probabilité conditionnelle des mots pour déterminer si un article est faux ou authentique [2]. Bien qu'efficace pour des données textuelles bien structurées, il est limité face à des *fake news* utilisant des constructions syntaxiques plus complexes [4, 16].

2.2 Deep Learning et Réseaux Neuronaux

L'émergence du deep learning a permis d'améliorer considérablement la précision des modèles de détection.

2.2.1 Réseaux de Neurones Convolutionnels (CNN)

Les CNN, bien que principalement utilisés en vision par ordinateur, ont été adaptés à l'analyse de texte en identifiant des motifs textuels complexes. Cette approche a démontré des performances prometteuses dans la classification des *fake news* en exploitant des caractéristiques locales du texte [6].

2.2.2 Réseaux de Neurones Récurrents (RNN) et LSTM

Les RNN, notamment les LSTM, sont utilisés pour capturer les dépendances contextuelles et temporelles dans les articles. Ces modèles sont capables d'analyser la structure narrative des *fake news* et d'identifier des incohérences dans le discours [8].

2.2.3 Transformers : BERT et autres architectures basée sur les LLM

Les modèles basés sur les transformers, comme BERT et FakeBERT, ont révolutionné la détection de *fake news* grâce à leur apprentissage bidirectionnel et à la combinaison avec les réseaux de neurones convolutifs, permettant ainsi une meilleure compréhension contextuelle et une précision accrue. [6]. Des modèles spécifiques comme WELFake exploitent des représentations linguistiques enrichies pour améliorer la classification des *fake news* [14].

Récemment, des recherches ont examiné l'efficacité des LLMs tels que GPT-3.5 dans la détection des *fake news*. Bien que ces modèles puissent générer des analyses pertinentes, ils ne surpassent pas nécessairement les modèles spécialisés complété d'un apprentissage fini sur des jeux de données annotés. Ils peuvent toutefois être exploités comme outils d'assistance pour améliorer l'interprétabilité des résultats, en fournissant des analyses multi-perspectives instructives. Ces modèles apportent des explications raisonnables et informatives à partir de diverses perspectives, aidant ainsi à affiner et contextualiser les résultats obtenus par des modèles de plus petite taille. De plus, ils permettent aux utilisateurs de bénéficier de nouvelles perspectives pour une meilleure compréhension des nouvelles [5].

2.3 Comparaison des méthodes

Les méthodes classiques, telles que l'utilisation du "sac de mots" (bag of words) et des n-grams pour analyser les fréquences de mots ou de groupes de mots, sont rapides et peu coûteuses en ressources. Cependant, elles manquent de compréhension contextuelle et peinent face aux articles trompeurs utilisant des tournures plus élaborées [4]. En revanche, les modèles de *deep learning* et les *transformers*, tels que BERT [6] ou GPT [9], offrent une meilleure précision en capturant les subtilités du langage, au prix d'une complexité accrue et d'un besoin de calcul plus important [10].

Le chapitre suivant présente comment les LLMs peuvent être intégrés dans ces systèmes pour renforcer la détection automa-

tique et améliorer l'interprétabilité des modèles existants.

3 Applications des LLMs dans la Détection des *fake news*

Les grands modèles de langage (LLMs) comme ceux utiliser par ChatGPT ont suscité un intérêt croissant dans la lutte contre la désinformation. Leur capacité à analyser, comprendre et générer du texte leur permet de jouer un rôle clé dans la détection des *fake news*, tout en nécessitant une évaluation de leur efficacité en fonction de leur complémentarité avec d'autres approches [5]. Ce chapitre examine les applications des LLMs, ses atouts, ses limites et des études de cas démontrant son efficacité lorsqu'il est intégré dans des systèmes de détection.

3.1 Capacités des LLMs dans la Détection des fake news

Un LLM peut être utilisé comme un outil d'aide à la classification et à l'analyse des *fake news* grâce à plusieurs fonctionnalités :

- *Analyse contextuelle et sémantique* : Un LLM, grâce à son apprentissage sur de vastes corpus textuels, est capable d'identifier des incohérences sémantiques et des structures linguistiques typiques des *fake news* [13].
- *Détection assistée par des bases de connaissances* : En intégrant des bases de faits structurées comme Politifact, un LLM peut évaluer la véracité des affirmations en croisant les informations [1].
- *Génération de raisonnements explicatifs* : Contrairement aux modèles boîte noire, un LLM peut fournir des explications détaillées sur son processus de classification, améliorant ainsi l'interprétabilité des résultats [5].

3.2 Études de Cas et Expérimentations

Des études récentes ont exploré l'application d'LLM dans la détection des *fake news*, mettant en évidence ses forces et ses limites.

3.2.1 Comparaison avec les modèles traditionnels

Une analyse comparative entre des LLMs (comme BERT) et des modèles classiques comme Naïve Bayes a révélé qu'un LLM seul obtient des performances inférieures aux modèles spécifiquement entraînés sur des jeux de données annotés. Toutefois, lorsqu'il est utilisé en complément d'un modèle supervisé, le LLM apporte une plus-value en générant des raisonnements explicatifs détaillés, ce qui aide à mieux comprendre les motifs derrière les décisions prises. Cela permet non seulement d'affiner l'évaluation des textes suspects, mais aussi de réduire les faux positifs de manière significative, avec une amélioration de 2,50 % et 1,10 % sur les jeux de données PolitiFact et GossipCop respectivement en termes de précision et de rappel [8].

3.2.2 Détection des *fake news* sur les réseaux sociaux

Dans le cadre de la désinformation liée à la COVID-19, une expérimentation utilisant un LLM sur des jeux de données de tweets et d'articles a montré que le modèle pouvait identifier certaines fausses informations, mais rencontrait des difficultés face aux contenus nuancés et à la désinformation partielle [1]. Cette limitation met en évidence la nécessité d'une validation croisée avec des sources factuelles externes.

3.3 Défis et Limites

Bien que prometteur, un LLM présente plusieurs limitations :

- *Biais des données d'entraînement* : Les LLMs sont influencés par les biais

présents dans leurs données d'apprentissage, ce qui peut affecter la fiabilité de leurs classifications [5].
— *Absence de vérification en temps réel* : un LLM n'a pas un accès direct à des bases de données constamment mises à jour, ce qui peut entraîner des erreurs lorsqu'il évalue des informations récentes [13].
— *Vulnérabilité à la manipulation* : Les *fake news* conçues pour contourner les détecteurs automatiques peuvent exploiter les failles des LLMs, ce qui nécessite l'intégration de techniques de défense supplémentaires [8].

Les LLMs représentent une avancée significative dans la détection des *fake news*. L'avenir de la lutte contre la désinformation repose sur des approches hybrides combinant intelligence artificielle et validation humaine [5]. Le chapitre suivant présente les collections et les méthodologies d'évaluation utilisées pour mesurer l'efficacité des systèmes de détection des *fake news*.

4 *Dataset* et méthodologies d'évaluation

L'évaluation des systèmes de détection des *fake news* repose sur l'utilisation de *Dataset* annotées et de méthodologies rigoureuses. Ces *Dataset* permettent d'entraîner et de tester les modèles en fournissant des échantillons de textes classifiés selon leur véracité. De plus, des métriques et des protocoles d'évaluation standardisés permettent de comparer les performances des différents algorithmes.

4.1 Dataset couramment utilisées

Différentes *Dataset* ont été développées pour répondre aux besoins spécifiques des recherches sur la détection des *fake news*. Ces jeux de données sont construits à partir de sources journalistiques, de vérifications factuelles ou de réseaux sociaux.

4.1.1 LIAR

LIAR est un *Dataset* introduite par William Yang Wang contenant plus de 12 000 déclarations politiques annotées extraites du site *Politifact*. Chaque déclaration est classée selon six niveaux de véracité : "*pants-fire*", "*false*", "*barely-true*", "*half-true*", "*mostly-true*" et "*true*". Cette granularité permet aux modèles d'apprendre à détecter différents degrés de désinformation [15].

4.1.2 FakeNewsNet

FakeNewsNet [8] est un *Dataset* combinant deux ensembles bien connus :
— PolitiFact : Contient des articles fact-checkés et annotés en fonction de leur fiabilité.
— GossipCop : Se concentre sur les rumeurs et *fake news* circulant dans le domaine du divertissement.
Ces jeux de données sont souvent utilisés pour comparer les performances des modèles de détection [8, 5, 15].

4.1.3 WELFake

WELFake [14] est un *Dataset* de 72 000 articles annotée vraies et fausses, intégrant des caractéristiques linguistiques extraites automatiquement ainsi que des représentations vectorielles de texte basées sur des embeddings de mots *(Word2Vec, TF-IDF)*. Les caractéristiques linguistiques comprennent des patterns lexicaux et syntaxiques, qui permettent d'enrichir les données collectées pour une analyse plus fine des *fake news* [14, 13].

4.1.4 COVID-19 *fake news Dataset*

Avec l'augmentation de la désinformation liée à la pandémie, plusieurs *Dataset* spécifiques ont été créées. Alghamdi et al. ont développé un jeu de données combinant des articles de presse et des tweets vérifiés autour de la pandémie, permettant aux cher-

cheurs d'évaluer l'efficacité des modèles sur des *fake news* liée à la COVID-19 [1].

4.2 Métriques d'évaluation

L'évaluation des modèles de détection des *fake news* repose sur des métriques issues de la classification supervisée, permettant d'estimer leur fiabilité et leur performance.

Ces trois métriques sont essentielles pour évaluer un modèle :

- *Précision* : Proportion d'articles correctement identifiés comme *fake news* parmi toutes les prédictions positives.
- *Rappel* : Proportion de *fake news* effectivement détectées parmi l'ensemble des *fake news* présentes dans le corpus.
- *F1-score* : Moyenne harmonique de la précision et du rappel, permettant une mesure équilibrée entre ces deux indicateurs.

4.3 Protocoles d'évaluation

Les protocoles d'évaluation sont essentiels pour assurer la robustesse et la reproductibilité des modèles. Plusieurs méthodologies sont couramment utilisées dans la recherche.

4.3.1 Validation croisée

La validation croisée permet de mesurer la capacité de généralisation des modèles [13]. En divisant l'ensemble des données en plusieurs sous-ensembles et en alternant les parties utilisées pour l'entraînement et le test, on obtient une estimation plus fiable de la performance réelle d'un modèle.

4.3.2 Comparaison avec des modèles de référence

Pour évaluer un modèle, ses performances sont comparées à des modèles de référence tels que :

- *Naïve Bayes et SVM* : Utilisés comme référence dans les premières approches de classification des *fake news* [4].
- *BERT et FakeBERT* : Ces modèles basés sur les transformers offrent des performances élevées en raison de leur capacité à capturer le contexte [6].
- *LLMs comme GPT-3.5* : Bien que les LLMs comme GPT-3.5 aient montré une capacité d'analyse multi-perspective, ils ne surclassent pas toujours les SLMs entraînés spécifiquement pour la tâche. Cependant, leur intégration en tant que conseillers pour les SLMs peut améliorer les performances globales de détection des fake news [5].

4.3.3 Évaluation sur des scénarios réalistes

Les modèles sont également évalués sur des scénarios réels en intégrant des tests avec des *fake news* récemment diffusées sur les réseaux sociaux. Cette approche permet d'identifier les limitations des modèles et leur robustesse face à de nouvelles formes de désinformation [1].

L'évaluation rigoureuse des systèmes de détection des *fake news* repose sur l'utilisation de *Dataset* diversifiées et de méthodologies éprouvées. L'amélioration des modèles nécessite une prise en compte des biais présents dans les jeux de données, ainsi qu'une combinaison de plusieurs métriques d'évaluation pour garantir la fiabilité des résultats. Le chapitre suivant présente les défis et limitations actuels de ces approches et propose des perspectives pour améliorer la robustesse des modèles face aux nouvelles formes de désinformation.

5 Défis et limitations actuels

Bien que les avancées en intelligence artificielle aient significativement amélioré la détection des *fake news*, plusieurs défis et limitations demeurent. Ces obstacles concernent à la fois des aspects techniques, éthiques et sociaux, entravant l'efficacité et l'adoption généralisée de ces technologies.

Sur le plan technique, les méthodes actuelles reposent sur des approches linguistiques et de réseau qui, bien que prometteuses, présentent des limites en termes de précision et de robustesse face à la diversité et à l'évolution constante des fausses nouvelles. Par exemple, les méthodes linguistiques telles que les "*bag of words*" ou l'analyse syntaxique peuvent manquer de contexte, tandis que les approches de réseau dépendent fortement de la qualité et de la disponibilité des données liées. D'un point de vue éthique, l'utilisation de ces technologies soulève des questions importantes concernant la confidentialité et la surveillance. Les biais algorithmiques peuvent mener à des résultats discriminatoires ou inexacts, suscitant des inquiétudes quant à la transparence et à la responsabilité des systèmes. Sur le plan social, la confiance du public et l'acceptation de ces technologies sont des défis majeurs. La crainte de la censure et la stigmatisation des utilisateurs de médias sociaux peuvent freiner leur adoption, malgré leur potentiel à améliorer la vérification des faits. De plus, l'implémentation de ces outils dans divers contextes culturels peut rencontrer des résistances dues aux différences de perception et d'acceptation de l'intelligence artificielle. Ce chapitre explore les principales limites rencontrées par les modèles actuels et les solutions envisageables pour les surmonter.

5.1 Défis techniques

5.1.1 Fiabilité et robustesse des modèles

Les modèles de détection des *fake news*, bien qu'efficaces sur des jeux de données standardisés, sont souvent vulnérables aux contenus générés artificiellement par des modèles avancés comme GPT-4 [5]. De plus, les nouvelles formes de désinformation, combinant textes, images et vidéos, compliquent davantage leur détection. L'adaptation des modèles aux *fake news* multimodales reste un défi technique majeur [1, 3].

5.1.2 Biais et éthique des modèles

Les biais présents dans les jeux de données influencent les performances des modèles, menant à des erreurs de classification [13]. Par exemple, certains *Dataset* contiennent plus de *fake news* politiques que scientifiques, ce qui peut fausser les prédictions. Il est donc essentiel d'intégrer des techniques d'atténuation des biais, telles que l'augmentation des données et la réévaluation continue des modèles [14].

5.1.3 Problèmes de généralisation

Les modèles entraînés sur des *Dataset* spécifiques peinent parfois à s'adapter à de nouveaux contextes [9]. Un modèle performant sur des articles politiques peut échouer lorsqu'il est appliqué à des *fake news* scientifiques ou sanitaires [1]. La variabilité linguistique et culturelle représente un autre défi, rendant nécessaire l'entraînement des modèles sur des ensembles de données diversifiés [5].

5.2 Défis éthiques et sociaux

5.2.1 Impact sur la liberté d'expression

L'automatisation de la détection des *fake news* soulève des préoccupations quant à la censure et à la liberté d'expression [5]. Certains algorithmes peuvent erronément classer des opinions critiques ou des discours minoritaires comme de la désinformation [9]. Il est donc crucial de développer des méthodes permettant de distinguer les contenus frauduleux des débats légitimes [5].

5.2.2 Manipulation des modèles

Les adversaires peuvent exploiter les faiblesses des modèles pour contourner les systèmes de détection [5]. Les *fake news* évoluent rapidement et certaines techniques, comme la reformulation ou l'usage de synonymes, peuvent tromper les algorithmes [9]. Cela nécessite des systèmes de détection adaptatifs, capables d'apprendre en continu

et de détecter les tentatives de contournement [5].

5.3 Solutions potentielles

5.3.1 Approches hybrides

L'intégration de la vérification factuelle humaine avec des systèmes d'IA permettrait d'améliorer la précision des modèles [8]. Des plateformes comme Politifact pourraient être utilisées en complément des algorithmes automatisés pour renforcer la crédibilité des résultats [14].

5.3.2 Amélioration des jeux de données

L'expansion et l'amélioration des *Dataset* annotées sont essentielles pour accroître la robustesse et la généralisation des modèles. L'utilisation de *Dataset* multimodales intégrant texte, images et métadonnées pourrait renforcer l'efficacité des systèmes [9].

5.3.3 Développement de modèles explicables

Les futurs systèmes devront être plus transparents et fournir des explications sur leurs décisions afin d'accroître leur adoption et leur fiabilité. L'intégration de techniques d'apprentissage interprétable, comme les visualisations d'attention et les explications textuelles générées par IA, peut améliorer la confiance des utilisateurs.

Les défis liés à la détection automatique des *fake news* sont nombreux et nécessitent une approche combinant innovations technologiques et réflexion éthique. L'amélioration des *Dataset*, le développement de modèles plus robustes et la mise en place d'une régulation adaptée sont des pistes prometteuses pour renforcer la lutte contre la désinformation. Le chapitre suivant explore les perspectives et orientations futures dans ce domaine.

6 Perspectives et Travaux Futurs

L'évolution rapide des modèles d'intelligence artificielle et l'apparition constante de nouvelles menaces informationnelles nécessitent une adaptation continue des approches de détection des *fake news*. Ce chapitre explore les perspectives d'amélioration des modèles existants ainsi que les axes de recherche futurs, en mettant l'accent sur des stratégies innovantes et des collaborations interdisciplinaires.

6.1 Améliorations des modèles existants

6.1.1 Entraînement sur des données plus diversifiées

L'entraînement des modèles sur des ensembles de données plus vastes et variés, incluant des *fake news* issues de différents contextes culturels et thématiques, permettrait d'améliorer leur robustesse face aux désinformations émergentes [3]. L'intégration de *Dataset* multimodales contenant des articles politiques, scientifiques et sociaux contribuerait à renforcer la capacité de généralisation des modèles [3].

L'analyse exclusive du texte est parfois insuffisante pour détecter certaines formes de désinformation, notamment celles accompagnées d'images ou de vidéos manipulées. L'intégration d'informations provenant de plusieurs sources (texte, images, vidéos, métadonnées des publications) permettrait d'améliorer significativement la précision des modèles [3]. Des architectures basées sur des modèles multi-modaux, pourraient être explorées pour mieux appréhender ces défis.

6.2 Axes de recherche futurs

6.2.1 Modèles plus transparents et interprétables

Le développement de modèles d'intelligence artificielle explicables est un enjeu clé

pour accroître la confiance des utilisateurs et faciliter l'intégration de ces technologies dans les systèmes de fact-checking [6].

6.2.2 Détection proactive des *fake news*

Plutôt que de réagir après la diffusion de *fake news*, de nouvelles approches pourraient se concentrer sur la prédiction et la prévention de leur propagation [5]. L'analyse des dynamiques de diffusion sur les réseaux sociaux et la détection précoce des tendances virales permettraient d'identifier les contenus suspects avant qu'ils ne deviennent massivement partagés [1].

6.2.3 Adaptation continue des modèles

Les *fake news* évoluent rapidement, exploitant de nouvelles failles linguistiques et narratives. Ces failles incluent l'utilisation de reformulations, de synonymes, et de structures narratives crédibles pour contourner les systèmes de détection et manipuler l'opinion publique [5]. Les modèles statiques perdent en efficacité face à ces changements, comme le montre une chute de performance des CNNs, qui atteignent seulement 58,6 % de précision sur le dataset WELFake, comparé aux modèles plus avancés comme GPT-4o qui atteignent 98,6 % [9]. Le développement de modèles capables d'apprentissage continu (*continual learning*) permettrait une meilleure adaptation aux nouvelles formes de désinformation [1]. L'intégration de techniques de mise à jour dynamique des jeux de données et des pondérations des modèles améliorerait leur réactivité face aux nouvelles stratégies de désinformation [8].

6.2.4 Collaboration entre IA et vérificateurs humains

Les modèles d'IA, bien que performants, ne peuvent pas encore remplacer entièrement l'expertise humaine en *fact-checking*. Une synergie entre IA et experts permettrait d'optimiser la détection des *fake news*. L'IA pourrait pré-filtrer les contenus suspects pour permettre aux vérificateurs humains de se concentrer sur les cas les plus complexes. Des plateformes collaboratives, intégrant des algorithmes d'IA et des contributions d'experts, pourraient ainsi être développées pour renforcer l'efficacité de la vérification des faits.

6.3 Impact sur la société et les médias

L'amélioration des systèmes de détection des *fake news* pourrait renforcer la confiance dans les médias et améliorer la qualité du débat public [9]. Toutefois, des considérations éthiques et réglementaires devront être intégrées pour éviter les abus et garantir la liberté d'expression. La mise en place de cadres juridiques et de standards internationaux sur l'utilisation de l'IA dans la lutte contre la désinformation est essentielle pour assurer un équilibre entre contrôle et respect des droits fondamentaux [16]. Par ailleurs, récemment, des réglementations sur l'utilisation de l'IA au sein de l'Union Européenne visant à encadrer cela ont été mises en place, mais étant donné que ces régulations sont encore récentes, aucune de mes sources ne parle de ce fait.

Les recherches futures sur la détection des *fake news* devront concilier avancées technologiques, considérations éthiques et transparence des modèles. L'intégration d'approches hybrides combinant IA, vérification humaine et analyse multi-modale représente une voie prometteuse pour une lutte plus efficace contre la désinformation. La mise en œuvre de modèles évolutifs, capables d'apprentissage continu et dotés d'une forte explicabilité, constituera un levier majeur pour améliorer la fiabilité des systèmes de détection.

7 Conclusion

7.1 Synthèse des résultats

Cet état de l'art a exploré les avancées majeures en matière de détection automatique des *fake news* en utilisant les approches d'apprentissage automatique classiques, les architectures de deep learning et les grands modèles de langage. Nous avons mis en évidence les forces et les limites des différentes méthodes, allant des modèles SVM et Naïve Bayes jusqu'aux transformers tels que BERT et GPT-3.5.

L'analyse des modèles suggère que les méthodes classiques comme les SVM et Naïve Bayes restent efficaces sur des jeux de données simples, mais qu'elles sont dépassées par les architectures neuronales avancées, notamment les réseaux avec couches d'attention et les modèles transformers [4, 10]. Les travaux récents ont démontré que les modèles multi-modaux et les techniques de fine-tuning améliorent significativement la robustesse des systèmes [14, 1]. Cependant, les défis liés aux biais des modèles, à la capacité de généralisation et à l'explicabilité des décisions sont des défis majeurs [6, 5].

7.2 Implications pratiques et théoriques

D'un point de vue pratique, les avancées en IA permettent de développer des outils plus performants pour lutter contre la désinformation. Ces modèles peuvent être intégrés dans des plateformes de fact-checking, des outils de modération et des systèmes de surveillance médiatique. Toutefois, leur efficacité dépend fortement de la qualité des données d'entraînement et de leur application [8].

D'un point de vue théorique, ces résultats mettent en évidence la nécessité de développer des approches hybrides combinant l'analyse linguistique, l'apprentissage profond et la vérification contextuelle [3]. Par ailleurs, l'explicabilité des modèles est un enjeu clé pour garantir leur adoption et leur transparence, notamment dans les décisions automatisées [13].

7.3 Recommandations pour les recherches futures

Pour surmonter les limitations actuelles, plusieurs axes de recherche peuvent être envisagés :

- Développement de modèles plus robustes : Intégrer des mécanismes d'adaptation continue afin de détecter plus efficacement les *fake news* générées par IA tout en réduisant les biais algorithmiques [5].
- Amélioration des jeux de données : Construire des *Dataset* plus équilibrées et diversifiées en intégrant des sources variées pour assurer une meilleure généralisation des modèles [14].
- Approches multi-modales : Combiner texte, image et vidéo pour détecter des *fake news* sous différentes formes et renforcer la fiabilité des classifications [1].
- Collaboration entre IA et vérificateurs humains : Intégrer des systèmes hybrides où l'IA assiste les experts humains pour optimiser la détection et limiter les erreurs [8].
- Transparence et éthique : Renforcer l'explicabilité des décisions prises par les modèles et mettre en place des standards de validation ouverts [13].

7.4 Conclusion générale

La détection automatique des *fake news* est un domaine en constante évolution, nécessitant des efforts soutenus pour perfectionner la précision, la robustesse et l'éthique des modèles développés. Bien que les approches basées sur les transformers aient montré des résultats intéressants, elles ne constituent pas encore une solution exhaustive face à la complexité du phénomène de la désinformation. L'avenir de la lutte contre les *fake news* réside probablement

dans une approche multidimensionnelle, associant l'intelligence artificielle, le journalisme d'investigation, ainsi que des régulations adaptées et des politiques publiques cohérentes. En effet, la mise en place de modèles de détection performants ne pourra avoir un réel impact que si des cadres législatifs contraignants les accompagnent. À cet égard, les récents développements dans les plateformes sociales, telles que les déclarations de X (anciennement Twitter) et Facebook concernant la gestion des contenus politiques et la manipulation de l'information, soulignent l'absence de régulations claires et de mécanismes d'application. Il est donc crucial que les législations, en particulier au sein de l'Union européenne, soient renforcées pour encadrer l'usage de ces technologies et garantir leur mise en œuvre éthique et responsable. La recherche dans ce domaine pourra ainsi contribuer à la construction d'un environnement informationnel plus fiable, transparent et équitable.

Références

[1] Jawaher Alghamdi, Yuqing Lin et Suhuai Luo. "Towards COVID-19 fake news detection using transformer-based models". In : *Knowledge-Based Systems* 274 (15 août 2023), p. 110642. ISSN : 0950-7051. DOI : 10.1016/j.knosys.2023.110642. URL : https://www.sciencedirect.com/science/article/pii/S0950705123003921 (visité le 03/02/2025).

[2] Alim Al Ayub Ahmed Et Al. "Detecting Fake News using Machine Learning : A Systematic Literature Review". In : *Psychology and Education Journal* 58.1 (1er jan. 2021). Number : 1, p. 1932-1939. ISSN : 0033-3077. DOI : 10.17762/pae.v58i1.1046. URL : http://psychologyandeducation.net/pae/index.php/pae/article/view/1046 (visité le 02/02/2025).

[3] Majdi Beseiso et Saleh Al-Zahrani. "A Context-Enhanced Model for Fake News Detection". In : *Engineering, Technology & Applied Science Research* 15.1 (2 fév. 2025). Number : 1, p. 19128-19135. ISSN : 1792-8036. DOI : 10.48084/etasr.9192. URL : https://etasr.com/index.php/ETASR/article/view/9192 (visité le 02/02/2025).

[4] Nadia K. Conroy, Victoria L. Rubin et Yimin Chen. "Automatic deception detection : Methods for finding fake news". In : *Proceedings of the Association for Information Science and Technology* 52.1 (2015). Number : 1 _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.2015.145052010082, p. 1-4. ISSN : 2373-9231. DOI : 10.1002/pra2.2015.145052010082. URL : https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.2015.145052010082 (visité le 03/02/2025).

[5] Beizhe Hu et al. "Bad Actor, Good Advisor : Exploring the Role of Large Language Models in Fake News Detection". In : *Proceedings of the AAAI Conference on Artificial Intelligence* 38.20 (24 mars 2024), p. 22105-22113. ISSN : 2374-3468, 2159-5399. DOI : 10.1609/aaai.v38i20.30214. arXiv : 2309.12247[cs]. URL : http://arxiv.org/abs/2309.12247 (visité le 19/02/2025).

[6] Rohit Kumar Kaliyar, Anurag Goswami et Pratik Narang. "FakeBERT : Fake news detection in social media with a BERT-based deep learning approach". In : *Multimedia Tools and Applications* 80.8 (1er mars 2021). Number : 8, p. 11765-11788. ISSN : 1573-7721. DOI : 10.1007/s11042-020-10183-2. URL : https://doi.org/10.1007/s11042-020-10183-2 (visité le 03/02/2025).

[7] press.is. *Measuring the reach of "fake news" and online disinformation in Europe*. 2018. URL : https://www.press.is/static/files/frettamyndir/reuterfake.pdf.

[8] Nishant Rai et al. "Fake News Classification using transformer based enhanced LSTM and BERT". In : *International Journal of Cognitive Computing in Engineering* 3 (juin 2022), p. 98-105. ISSN : 26663074. DOI : 10.1016/j.ijcce.2022.03.003. URL : https://linkinghub.elsevier.com/retrieve/pii/S2666307422000092 (visité le 03/02/2025).

[9] Konstantinos I. Roumeliotis, Nikolaos D. Tselikas et Dimitrios K. Nasiopoulos. "Fake News Detection and Classification : A Comparative Study of Convolutional Neural Networks, Large Language Models, and Natural Language Processing Models". In : *Future Internet* 17.1 (jan. 2025). Number : 1 Publisher : Multidisciplinary Digital Publishing Institute, p. 28. ISSN : 1999-5903. DOI : 10.

3390/fi17010028. URL : https://www.mdpi.com/1999-5903/17/1/28 (visité le 02/02/2025).

[10] Sneha SINGHANIA, Nigel FERNANDEZ et Shrisha RAO. "3HAN : A Deep Neural Network for Fake News Detection". In : t. 10635. 2017, p. 572-581. DOI : 10.1007/978-3-319-70096-0_59. arXiv : 2306.12014[cs]. URL : http://arxiv.org/abs/2306.12014 (visité le 02/02/2025).

[11] STATISTA. *Fake news : opinion on most affected topics in France*. 2024. URL : https://www.statista.com/statistics/1198660/fake-news-most-affected-topics-young-people-opinion-france/.

[12] STATISTA. *Fake news : spread platforms on the Internet in France*. 2024. URL : https://www.statista.com/statistics/1198673/fake-news-spread-ways-internet-france/.

[13] Ciprian-Octavian TRUICĂ et Elena-Simona APOSTOL. "It's All in the Embedding ! Fake News Detection Using Document Embeddings". In : *Mathematics* 11.3 (jan. 2023). Number : 3 Publisher : Multidisciplinary Digital Publishing Institute, p. 508. ISSN : 2227-7390. DOI : 10.3390/math11030508. URL : https://www.mdpi.com/2227-7390/11/3/508 (visité le 03/02/2025).

[14] Pawan Kumar VERMA et al. "WELFake : Word Embedding Over Linguistic Features for Fake News Detection". In : *IEEE Transactions on Computational Social Systems* 8.4 (août 2021). Number : 4 Conference Name : IEEE Transactions on Computational Social Systems, p. 881-893. ISSN : 2329-924X. DOI : 10.1109/TCSS.2021.3068519. URL : https://ieeexplore.ieee.org/document/9395133 (visité le 03/02/2025).

[15] WILLIAM YANG WANG. *"Liar, Liar Pants on Fire" : A New Benchmark Dataset for Fake News Detection*. Issue : arXiv :1705.00648. 1er mai 2017. DOI : 10.48550/arXiv.1705.00648. arXiv : 1705.00648[cs]. URL : http://arxiv.org/abs/1705.00648 (visité le 03/02/2025).

[16] Shuo YANG et al. "Unsupervised Fake News Detection on Social Media : A Generative Approach". In : *Proceedings of the AAAI Conference on Artificial Intelligence* 33.1 (17 juill. 2019). Number : 01, p. 5644-5651. ISSN : 2374-3468. DOI : 10.1609/aaai.v33i01.33015644. URL : https://ojs.aaai.org/index.php/AAAI/article/view/4508 (visité le 03/02/2025).