République Algérienne Démocratique et Populaire
Ministère de l'enseignement supérieur et de la recherche scientifique.
Université Abderrehmane MIRA-BEJAIA

Faculté des sciences exactes
Département informatique

Mémoire de fin d'étude

Réalisé par :
DAHMANI MOHAMED
CHERKIT MALEK

En vue de l'obtention du diplôme de master 02 en informatique
Spécialité : Réseaux et Sécurité Informatique

Thème :

Proposition d'un système de détection des fausses informations

Soutenu le : 17/09/2025
Devant le jury composè de :
ALLEM Khaled (Président)
AIT HACEN Souhila
BOUDRIES Abdelmalek
NOUICER Amina

Encadré par :
Mme BATTAT Nadia

Année Universitaire 2024/2025

Remerciement

Avant tout, nous exprimons notre gratitude envers Dieu le tout-puissant, qui nous a donné la force et le courage nécessaires pour mener ce travail à terme.

Nous remercions particulièrement notre encadrant, Madame Nadia BATTAT, pour avoir accepté de nous encadrer et pour son aide précieuse ainsi que ses orientations.

Nos sincères remerciements vont également à nos familles qui nous ont grandement soutenus tout au long de la réalisation de ce mémoire.

Enfin, nous remercions toutes les personnes qui ont contribué de près ou de loin à la réalisation de ce travail.

Dédicace

Avec tout mon respect et en exprimant ma profonde gratitude, je souhaite dédier ma remise de diplôme et ma joie à ma **mère**, qui est mon paradis.
Je tiens également à exprimer ma reconnaissance envers mon **cher père** pour son soutien indéfectible, ainsi qu'envers tous les membres de ma petite famille et **Amina**, ma future femme, qui ont toujours été présents pour moi.
Je voudrais également remercier chaleureusement tous mes amis, mon encadrante et mon binôme.

DAHMANI MOHAMED

Dédicace

Avec tout mon respect et en exprimant ma profonde gratitude, je souhaite dédier ma remise de diplôme et ma joie à ma **maman**, qui est mon paradis.
Je tiens également à exprimer ma reconnaissance envers mon **cher papa** pour son soutien indéfectible, ainsi qu'envers tous les membres de ma petite famille qui ont toujours été présents pour moi.
Je voudrais également remercier chaleureusement tous mes amis, mon encadrante et mon binôme.

CHERKIT MALEK

Table des matières

Table des figures

TABLE 1 – Liste des abréviations

Abréviation	**Signification**
TTS	Text-To-Speech (Synthèse vocale)
GAN	Generative Adversarial Network (Réseau antagoniste génératif)
G	Generator (Générateur)
D	Discriminator (Discriminateur)
GPT	Generative Pre-trained Transformer (Transformeur génératif pré-entraîné)
MIT	Massachusetts Institute of Technology
TF-IDF	Term Frequency–Inverse Document Frequency
BoW	Bag of Words
NLP	Natural Language Processing (Traitement automatique du langage naturel)
GNN	Graph Neural Network (Réseau de neurones sur graphes)
GCN	Graph Convolutional Network (Réseau de graphes convolutif)
BERT	Bidirectional Encoder Representations from Transformers
RoBERTa	Robustly Optimized BERT Pretraining Approach
MLM	Masked Language Model
NSP	Next Sentence Prediction
CBOW	Continuous Bag of Words
NLI	Natural Language Inference
OOV	Out-Of-Vocabulary
AI	Artificial Intelligence (Intelligence artificielle)
ML	Machine Learning (Apprentissage automatique)
DL	Deep Learning (Apprentissage profond)
ZSL	Zero-Shot Learning (Apprentissage sans exemple préalable)
LM	Language Model (Modèle de langage)
API	Application Programming Interface (Interface de programmation d'applications)
RMI	Remote Method Invocation (Invocation de méthode à distance)
JMS	Java Message Service
CLI	Command Line Interface (Interface en ligne de commande)
UI	User Interface (Interface utilisateur)
GUI	Graphical User Interface (Interface graphique utilisateur)
LLM	Large Language Model (Grand modèle de langage)
GAT	Graph Attention Network (Réseau de graphes à attention)
LIAR	Dataset de classification de la véracité (Fake News dataset)
API REST	Representational State Transfer API
SSL	Semi-Supervised Learning (Apprentissage semi-supervisé)
TPR	True Positive Rate (Taux de vrais positifs)
FPR	False Positive Rate (Taux de faux positifs)
F1	F1-Score (Mesure harmonique de la précision et du rappel)
ROC	Receiver Operating Characteristic (Courbe de caractéristiques de réception)

1 Introduction générale

L'ère numérique a profondément bouleversé notre rapport à l'information. Grâce aux technologies de communication et à la généralisation des réseaux sociaux, la diffusion de contenus est devenue instantanée, interactive et accessible à une très large audience. Cependant, cette facilité de circulation s'accompagne d'un phénomène inquiétant : la prolifération des fausses informations, communément appelées fake news. Ces dernières, souvent créées dans un but de manipulation, d'influence ou de profit, représentent un danger considérable pour la société, car elles peuvent altérer la perception collective de la réalité, provoquer des tensions sociales, influencer des décisions politiques ou économiques, et fragiliser la sécurité nationale et internationale.

La détection et la lutte contre les fausses nouvelles constituent un défi majeur et multidimensionnel. En effet, les fake news ne se limitent pas uniquement au texte : elles se déclinent également sous forme d'images truquées, de vidéos manipulées (deepfakes) ou encore d'enregistrements audio falsifiés. De plus, les auteurs de désinformation exploitent les dynamiques propres aux réseaux sociaux – viralité, personnalisation algorithmique et effets de communauté – afin de maximiser leur portée et d'échapper aux systèmes classiques de vérification.

Dans ce contexte, l'intelligence artificielle offre des solutions prometteuses. Les approches basées sur le Traitement Automatique du Langage Naturel (NLP) permettent d'analyser le contenu textuel et d'identifier des incohérences sémantiques, tandis que les Réseaux de Neurones Graphiques (GNN) exploitent la structure des interactions sociales pour repérer des schémas de diffusion suspects. Combinées à des techniques multimodales (texte, image, audio, vidéo), ces approches hybrides ouvrent de nouvelles perspectives pour renforcer la robustesse des systèmes de détection.

Afin de mieux comprendre et proposer une solution adaptée, ce mémoire est structuré en quatre chapitres principaux :

Chapitre 1 : Les fausses informations Ce chapitre constitue la base théorique du travail. Il présente la définition des fausses nouvelles, leurs principales catégories (rumeurs, propagande, satire trompeuse, clickbait, etc.), ainsi que leurs caractéristiques spécifiques. Une attention particulière est accordée aux impacts des fake news sur les sociétés modernes, que ce soit sur le plan social, politique, économique ou sécuritaire. Ce chapitre met également en évidence les défis liés à leur identification et à leur classification.

Chapitre 2 : Les réseaux sociaux et la diffusion des fake news Le deuxième chapitre se concentre sur les réseaux sociaux, qui représentent aujourd'hui le canal privilégié de propagation des fausses informations. Il en propose une définition, une classification et une analyse de leurs usages. Les enjeux liés à leur rôle (sociaux, économiques, politiques) y sont discutés en détail. Ce chapitre examine également les différentes approches de détection déjà mises en place par les plateformes et les chercheurs, tout en soulignant leurs limites face à l'évolution des techniques de désinformation.

Chapitre 3 : Approches hybrides pour la détection des fausses informations Dans ce chapitre, nous présentons les fondements théoriques et méthodologiques des approches hybrides. Nous expliquons comment le NLP permet de traiter et analyser les contenus textuels, tandis que les GNN offrent la possibilité de modéliser et exploiter les graphes d'interactions et de propagation des messages. L'apport des méthodes multimodales est également discuté, notamment pour la prise en charge des contenus visuels et audio. Le chapitre met en évidence les avantages d'une telle approche intégrée, mais aussi les défis techniques qu'elle implique (complexité computationnelle, qualité des données, scalabilité).

Chapitre 4 : Proposition et conception d'un système hybride Le dernier chapitre est consacré à la proposition d'un modèle hybride pour la détection des fausses informations sur les réseaux sociaux. L'architecture du système est décrite en détail, en présentant ses principaux modules : collecte et prétraitement des données, analyse multimodale, détection via NLP et GNN, puis décision et réponse (alerte, blocage, notification). Une réflexion est menée sur la mise en œuvre pratique de ce système, en abordant les choix technologiques possibles, les contraintes de performance et les perspectives d'amélioration. Enfin, une simulation est esquissée afin de montrer la faisabilité de la solution proposée.

En somme, ce travail s'inscrit dans une dynamique de recherche visant à renforcer les outils de cybersécurité et de fiabilité de l'information. L'objectif est de contribuer, à travers une approche hybride et multimodale, à limiter la propagation de la désinformation, et ainsi à préserver l'intégrité de l'espace numérique et social.

Chapitre I :
Généralités sur les fausses informations

1.1 Introduction

Le transfert massif d'informations par le biais d'Internet et des plateformes sociales a rendu la propagation de contenus trompeurs, connus sous le nom de fausses informations ou "fake news", plus aisée. Qu'elles soient délibérées (désinformation) ou non (mésinformation), ces dernières peuvent engendrer des répercussions sérieuses dans plusieurs domaines : politique, santé, sécurité, etc.

Ce chapitre expose les bases de ce phénomène : définitions, traits distinctifs, catégories et formats des fausses informations. Il examine également les techniques contemporaines de création de contenu falsifié, comme les réseaux antagonistes génératifs (GAN) et les auto-encodeurs, ainsi que les effets de ces technologies.

1.2 Définition

Le terme Fake news en anglais, ou fausse information en français, semble désigner la même chose, mais en réalité, une distinction existe entre les deux. Le terme Fake news est souvent utilisé par certaines entités pour désigner une information sans source reconnue ou un opinion non vérifiée. Cette utilisation reste toutefois imprécise, car elle ne distingue pas une information douteuse (non confirmée) d'une information réellement fausse. En revanche, l'expression fausses informations se rapporte à des contenus identifiés comme erronés par une source fiable. Plus spécifiquement, certaines de ces fausses informations visent délibérément à induire le public en erreur, à capter l'attention grâce à des éléments prétendument authentiques, à créer le scandale ou à influencer l'opinion. Elles sont généralement produites par des individus ou des groupes agissant pour leurs propres intérêts, et leur diffusion répond le plus souvent à des motivations personnelles, politiques ou économiques [1].

1.3 Classification des fausses informations

— **Contenu fabriqué** : Ces informations sont créées de toutes pièces, sans aucun lien avec la réalité. Elles exploitent souvent l'actualité ou les préoccupations du public pour susciter des réactions émotionnelles. Par exemple, pendant la pandémie de COVID-19, de fausses informations sur des remèdes miracles ont circulé massivement, mettant parfois des vies en danger. Ces contenus sont particulièrement dangereux car ils s'appuient sur la crédibilité des plateformes qui les hébergent [2].

— **Contenu manipulé** : Il s'agit d'informations réelles qui ont été altérées de manière subtile mais significative. Un cas emblématique est celui des deepfakes - ces vidéos où le visage et la voix d'une personnalité sont reproduits artificiellement pour lui faire dire ce qu'elle n'a jamais dit ![3]. Plus courant, le simple recadrage d'une photo peut changer complètement sa signification, comme lorsqu'une image de manifestation pacifique est recadrée pour ne montrer que les quelques violences qui s'y sont produites ![4].

— **Contenu imposteur** : Cette catégorie regroupe les sites qui imitent délibérément des médias reconnus pour tromper les lecteurs. La technique va du nom de domaine similaire (comme "bbc-news.com.co" au lieu de "bbc.com") à la copie minutieuse de la charte graphique. Ces imitateurs profitent de la confiance que le public accorde aux vrais médias pour donner du crédit à leurs mensonges [5].

— **Contexte faux** : Ici, l'information de base est vraie, mais son cadre est falsifié. Une photo authentique d'un conflit ancien peut être présentée comme illustrant un événement récent. Les statistiques officielles sont souvent détournées de cette manière, en étant extraites de leur contexte temporel ou géographique pour servir un discours mensonger [2].

— **Satire exploitée** : Les sites parodiques comme Le Gorafi ou The Onion créent délibérément des informations humoristiques. Le problème survient quand ces contenus sont partagés hors de leur contexte comique, souvent avec des légendes les présentant comme vraies. Ce phénomène s'est accru avec les réseaux sociaux où le ton sarcastique n'est pas toujours perceptible [6].

1.4 Caractéristiques

1.4.1 Apparence credible

C'est l'une des caractéristiques majeures des fausses informations, défini pour reproduire les critères visuelles et textuelles des sources d'information crédibles. Pour gagner la confiance du lecteur et dissimuler leur nature trompeuse, ces contenus utilisent souvent des titres captivants, des mises en page soignées, des logos qui imite ceux de médias réputés, ainsi qu'un style d'écriture journalistique. Cette méthode a pour objectif de crée une illusion d'authenticité, en particulier auprès d'un public qui n'est pas familier avec les méthodes de contrôle de l'information [7].

1.4.2 Absence de sources fiables

Un autre critère clé pour détecter la désinformation est l'absence de source crédibles, identifiables et vérifiables. À l'inverse des articles rédigés par des journalistes compétents ou des organismes de compétence, les contenus trompeurs ont souvent tendance a ne pas mentionner leur source ou à se référer à des entités obscures, anormales ou peu fiables. Cette confusion rend la vérification des faits présentes presque impossibles on renforce leur caractère fallacieux [8].

1.4.3 Diffusion rapide sur les réseaux sociaux

C'est l'un des éléments essentiels de la propagation des fausses informations est leur circulation massive et rapide à travers les plateformes de médias sociaux. Ces plateformes occupent une position centrale dans la diffusion de contenus trompeurs, en raison de leur structure algorithmique élaborée pour optimiser l'implication des utilisateurs. Ces informations, généralement chargées d'émotion, sensationnelles ou polarisantes, profitent d'un taux de diffusion supérieur à celui des contenus factuels ou neutres [8].

1.4.4 Présence dans divers formats

Ces informations sont marquées par leur faculté à s'ajuster a plusieurs formats et supports de diffusion, ce qui amplifie leur portée et leur impact. Plutôt que de se diffuser uniquement sous forme d'écrits ou de blagues textuelles, elles se répandent aujourd'hui dans une multitude de formats tel que les vidéos falsifiées (deepfakes), images modifiées, messages courts sur les réseau sociaux, podcasts, ou encoure commentaires intègres dans discussions en ligne [9].

1.5 Types de fausses informations

Dans le discours actuel, des expressions telles que désinformation, mésinformation et mal information on évolue pour devenir des termes généraux utilises pour déterminer divers éléments comme les informations extrêmement partisanes, des rumeurs, des théories du complot et même encore des préjuges idéologiques pour mieux comprendre la terminologie, on propose les définitions suivantes :

1.5.1 Désinformation

Tout d'abord, la désinformation (fausse information sans intention malveillante) fait référence a une information clairement erronée qui a été créé et diffusée volontairement dans le but de produire une confusion, de manipuler ou de tromper. Elle peut comporter a la fois des vérités et des mensonges ou déformer intentionnellement le contexte, ce qui complique sa distinction par rapport à un contenu authentique [10].

1.5.2 Mésinformation

La mésinformation (fausse information délibérée) fait référence à des informations incorrectes qui ne sont pas forcément trompeuses et qui n'ont pas été diffusées dans l'intention de causer du tort. Un exemple fréquent est lors d'un évènement d'actualité largement médiatise et viral sur les réseaux sociaux ou on trouve des individus partagent des rumeurs, des vieilles images sans réaliser qui ne sont pas réellement associées à cet évènement [10].

1.5.3 Mal information

La mal information fait référence à une information authentique, mais exploitée de façon nuisible, habituellement elle est diffusée pour causer du tort à un individu, une organisation ou un collectif. Ce type d'information s'appuie sur des vérités, mais dont la diffusion ou la façon de les exposer vise clairement à nuire [11].

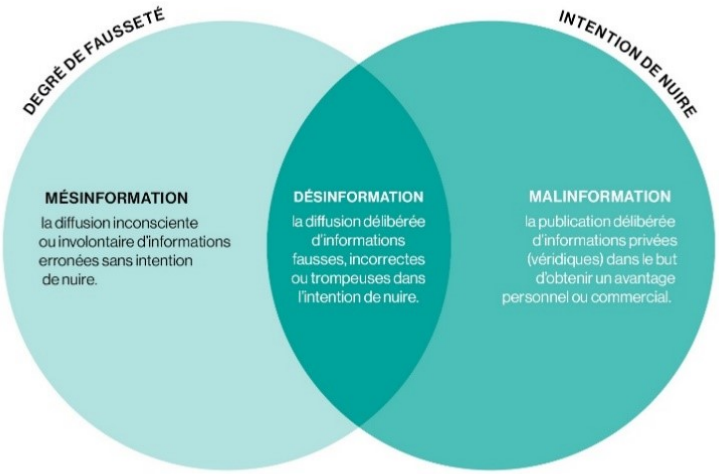

FIGURE 1 – Types des fausses informations[73]

1.6 Multimodalité des fausses informations (texte, image, audio, vidéo)

Les principales caractéristiques qui différencient les fausses informations sont leur source, leur format et leur objectif. Certaines informations erronées sont fabriquées sans intention de blesser, alors que d'autres sont intentionnellement élaborées pour induire en erreur et causer un effet défavorable. Il est fondamental de saisir ces diverses catégories pour étudier la complexité du phénomène de désinformation et ajuster les stratégies de détection et de prévention [12].

1.6.1 Audio Deepfake

C'est un enregistrement qui est modifier ou crée en utilisant une intelligence artificielle, qui vise à reproduire la voix d'une personne réelle afin de lui attribuer des paroles qu'elle n'a jamais prononcées, cette technologies base sur des modelés d'apprentissage automatique, dont les réseaux de neurones et les méthodes de text-to-speech(TTS) avancée, capable de reproduire le rythme et même les émotionnes d'une voix humaine en utilisant simplement un échantillon vocal [13].

1.6.2 Fausses images

Les fausses images sont des éléments visuels modifies et détacher de leur contexte ou bien totalement crée dans l'intention de manipuler la perception du public. Elles jeu un rôle important dans la diffusion des informations erronées, car l'image est un outil puissant. En peut considérer une image comme une preuve qui permet de valider une information fasse malgré l'absence du contenu textuel associé [14].

1.6.3 Fausses videos

Les vidéos falsifiées ou manipulées, représenter une forme persuasive et sophistiquée de la désinformation. Le but est de modifier ou de crée des séquences visuelles et sonores pour tromper les spectateurs,

FIGURE 2 – fausse image [74]

influencer l'opinion, nuire à une personne ou une institution. De nus joues grâce à l'intelligence artificielle cette méthode à évoluer pour modifier discrètement le contenu des vidéos réelles ou génère complètement du contenu artificiel [15] .

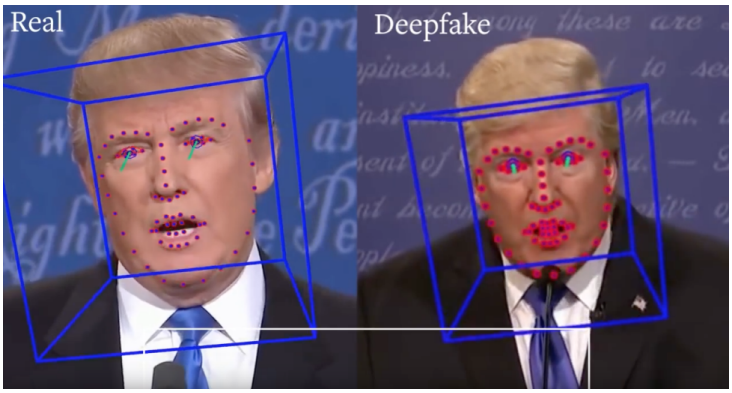

FIGURE 3 – fausse vidéo[75]

1.7 Méthodes utilisées pour la génération des fausses informations

Les algorithmes d'apprentissage automatique sont souvent utilisés par les générateurs automatiques de fausses informations pour produire des contenus qui ont l'apparence réels des articles journalistiques. Les réseaux de neurones générateurs adversaires (GAN) et les auto encodeurs sont les deux techniques les plus fréquemment employées pour produire ces textes.

1.7.1 Les Generatives Adversarials Networks (GANs)

Les Generative Adversarial Networks (GANs), présentés par Ian Goodfellow et ses collaborateurs en 2014, forment une catégorie de modèles d'intelligence artificielle générative capables à créer des contenus hautement réalistes à partir de données d'apprentissage. Elles sont souvent employées pour générer des images, vidéos ou sons totalement synthétiques, mais qui sont particulièrement complexes à différencier de la réalité, d'où leur association liée avec les deepfakes. Et ce réseau est composée de deux sous-réseaux formés en se opposant l'un à l'autre, dans un contexte d'apprentissage non supervisé.

- **Le Générateur (Generator, G) dans un GAN** Pour générer de fausses informations destinées à l'apprentissage du discriminateur, le générateur utilise un réseau de neurones. Il prend en entrée un vecteur de taille fixe (z), arbitraire et génère des données en sortie. L'objectif principal du générateur est de faire en sorte que le discriminateur classe les échantillons issus de la fausse base de données qu'il génère comme s'ils provenaient de la vraie base de données. Ainsi, lorsque D(G(z)) se rapproche de 1, cela signifie que le générateur a atteint

son objectif. L'objectif est donc d'optimiser D(G(z)), ce qui se traduit par la minimisation de $\log(1 - D(G(z)))$

$$\min_G \; E_{z\sim p_z(z)} \left[\log(1 - D(G(z)))\right]$$

[16]

- **Le Discriminateur (Discriminator, D) dans un GAN :** Le Discriminateur (Discriminator, D) constitue l'un des deux éléments cruciaux d'un réseau GAN. Sa tâche consiste à différencier les données authentiques (provenant du jeu d'entraînement) des données artificielles générées par le Générateur. En tant que classificateur binaire, il prend en entrée un échantillon, qu'il soit réel ou synthétique, pour produire une probabilité attestant de son authenticité. Le but du Discriminateur est d'optimiser la probabilité de classer les données avec précision, en assignant une valeur près de 1 aux données authentiques et proche de 0 aux données inauthentiques. Habituellement, cela s'appuie sur une architecture convolutive (CNN) pour les images, qui est capable d'extraire des caractéristiques visuelles distinctives [16].

$$\max_D \; E_{x\sim p_{\text{data}}(x)} \left[\log D(x)\right] + E_{z\sim p_z(z)} \left[\log(1 - D(G(z)))\right]$$

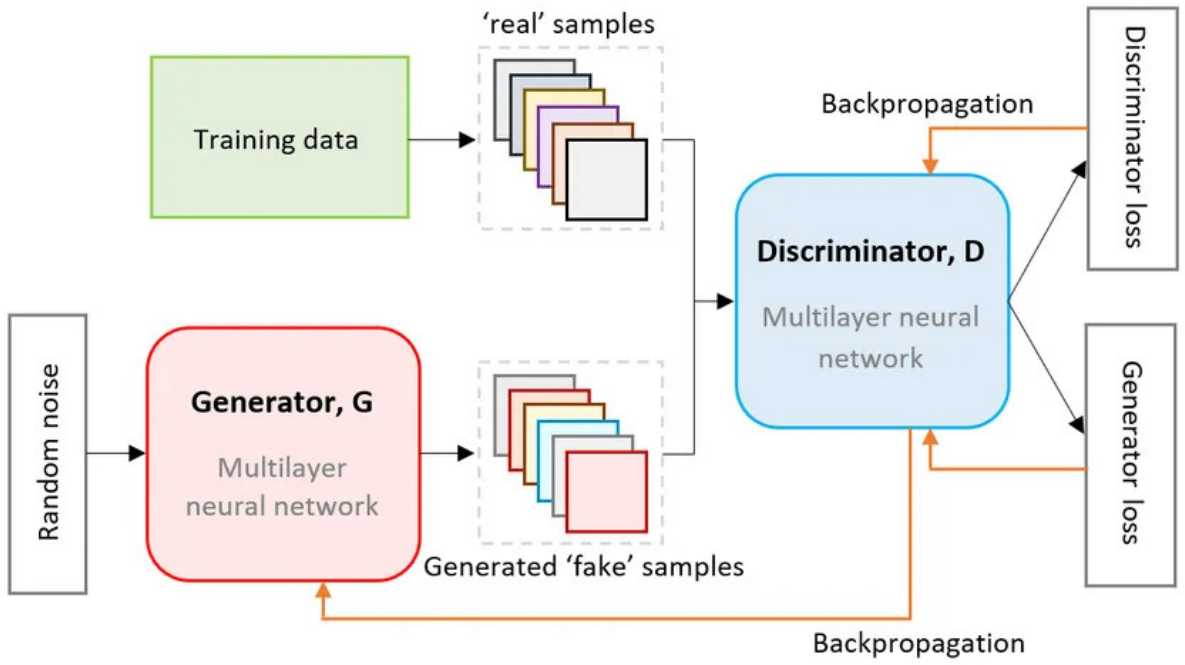

FIGURE 4 – Architecture du réseau GAN[76]

1.7.2 Les auto-encodeurs

: Les auto-encodeurs, qui sont des algorithmes d'apprentissage non supervisé basés sur des réseaux de neurones artificiels, rendent possible la création d'une information représentation d'un ensemble de données. En règle générale, ce dernier est de taille plus petite et renferme moins de descripteurs. Cela permet de diminuer la taille du jeu de données. La prochaine étape consiste à associer l'encodeur de la première image avec le décodeur de la seconde afin de générer une troisième image artificielle à partir des deux premières. Une auto-architecture est constituée de deux éléments : l'encodeur et le décodeur.

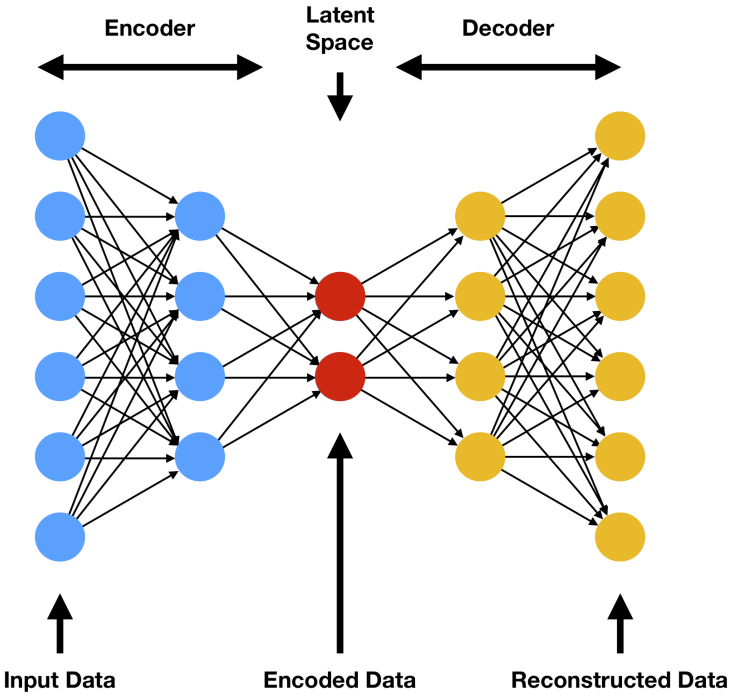

FIGURE 5 – Architecture de l'auto-encodeur[77]

- **Encodeur :** Le rôle de l'encodeur est de réduire la taille du jeu de données d'entrée en une représentation plus compacte. Dans cette optique, il puise les caractéristiques (features) les plus significatives à partir des données brutes. Cela entraîne une formation condensée appelée goulet d'étranglement, également connu sous le nom d'espace latent.
- **Décodeur : :** Contrairement à l'encodeur, le décodeur décompresse le goulot d'étranglement pour reconstruire les données. Son enjeu est d'exploiter les attributs présents dans le vecteur condensé pour tenter de reproduire avec la plus grande précision possible l'ensemble de données [17].

- **Modèles de langage GPT :** Des modèles de langage volumineux tels que GPT (Transformer Génératif Pré-entraîné) ont la capacité de produire de manière autonome des contenus textuels cohérents et fluides à partir d'un simple déclencheur. Grâce à des milliards de paramètres assimilés à partir d'immenses collections de textes, ils sont capables de générer des articles, des titres ou des messages convaincants et quasiment incomparables de ceux rédigés par des êtres humains. En matière de production de fausses informations, ces modèles servent à créer rapidement des contenus trompeurs, en reproduisant le style journalistique ou en ajustant le langage au public visé [17].

1.8 Impact global des fausses informations sur différents domaines

Les fausses informations ont un effet à la fois multidimensionnel et profond sur les sociétés contemporaines, influençant divers secteurs primordiaux :

- **Politique et processus démocratique :**

L'utilisation des fausses informations pour manipuler les opinions politiques est bien documentée. Pendant l'élection présidentielle de 2016 aux États-Unis, un grand nombre de contenus trompeurs soutenant certains candidats ont été largement diffusés, ce qui a pu influencer le résultat du vote [18].

- **Santé publique :**

La propagation des fausses informations dans le domaine de la santé peut entraîner un rejet des mesures

sanitaires, une baisse de la vaccination, ainsi qu'une augmentation des comportements à risque. Par exemple, les rumeurs liées au COVID-19 ont conduit à des comportements dangereux et à une méfiance accrue envers les vaccins [19].

- **Sécurité publique et gestion des crises :**

Lors de catastrophes naturelles, comme les tremblements de terre au Japon ou l'ouragan Sandy en 2012, des fake news ont été diffusées pour semer la panique. Cela a perturbé la coordination des secours et la diffusion d'informations fiables [20].

- **Économie et marchés financiers :**

En 2008, une fausse information sur la faillite de United Airlines a causé une chute instantanée de 76% du cours de l'action en quelques minutes. Malgré un démenti, l'action est restée affectée pendant plusieurs jours. De telles manipulations peuvent engendrer d'importantes pertes financières [21].

- **Cohésion sociale :**

Les fausses informations renforcent les divisions idéologiques et alimentent les discours polarisants. Le phénomène des chambres d'écho alimente les tensions sociales et les conflits communautaires c[22].

- **Crédibilité des médias et des institutions :**

Une exposition continue aux fausses informations peut affaiblir la confiance dans les médias traditionnels, les autorités publiques et scientifiques. Cela favorise l'émergence du scepticisme et du complotisme chez certains groupes de la population [23].

- **Radicalisation et violence physique :**

Dans des cas extrêmes, les fake news peuvent conduire à des actes violents. L'épisode du « Pizzagate » en 2016 montre comment une théorie du complot a poussé un homme armé à attaquer une pizzeria à Washington [24].

1.9 Origines psychologiques et sociales de la propagation des fausses informations

- **Biais cognitifs**

Les recherches montrent que les individus privilégient les informations conformes à leurs opinions préexistantes, un phénomène appelé biais de confirmation . Ce mécanisme s'explique par plusieurs facteurs psychologiques : la dissonance cognitive pousse au rejet des faits contradictoires, tandis que l'effet de simple exposition augmente la perception de véracité des informations fréquemment rencontrées . Des études expérimentales démontrent que ces biais affectent particulièrement le traitement de l'information politique [24].

- **Faible esprit critique**

L'analyse des capacités de détection des fake news révèle que seulement 4 % des individus identifient correctement les informations trompeuses dans des tests contrôlés . Cette difficulté varie selon le niveau d'éducation médiatique, l'âge et les compétences analytiques des individus . Les travaux de Pennycook et Rand (2018) montrent que les personnes dotées de fortes capacités analytiques sont 35 % moins vulnérables aux fake news [24].

- **Écho informationnel**

Les études sur les réseaux sociaux mettent en évidence la formation de chambres d'écho où 91,5 % des interactions conspirationnistes proviennent de cercles homogènes [22]. L'analyse de 1,1 million d'utilisateurs Twitter pendant les élections américaines de 2016 révèle que 76 % des conservateurs et 73 % des libéraux étaient principalement exposés à des contenus alignés avec leurs orientations politiques ![25]. Ces dynamiques sont amplifiées par les algorithmes de recommandation qui priorisent les contenus générant le plus d'engagement [22].

1.10 la détection des fausses nouvelles

1.10.1 définition

C'est le processus qui vise à identifier et à catégoriser automatiquement les informations fausses, fallacieuses ou intentionnellement incorrectes diffusées sous forme d'articles, de publications ou de titres. La détection des fausses informations peut être formulée comme un problème de classification binaire, où l'objectif est de développer une fonction prédictive permettant de classer un contenu d'actualité selon son

authenticité. Cette fonction pourrait retourner 1 si le contenu est entièrement faux et 0 s'il est totalement fiable [24].

1.10.2 buts

— **Contrôle de la véracité des contenus :** Un des buts principaux est de pouvoir détecter automatiquement les informations fallacieuses ou erronées circulant sur le web, qu'il s'agisse de textes, d'images ou de vidéos. Cela nécessite l'examen conjoint de divers types de données (multimodalité) et l'emploi d'algorithmes intelligents en mesure de repérer les signaux de désinformation dès leurs premières phases de propagation .

— **Précautions contre la diffusion de désinformation :** Il est crucial de minimiser la propagation des fausses informations avant qu'elles n'atteignent une ampleur considérable pour atténuer leur influence sociétale. Une identification précoce contribue à empêcher que des informations fallacieuses n'affectent de manière néfaste l'opinion publique ou ne déclenchent des répercussions significatives sur les plans politique, social ou économique .

— **Renforcer la confiance dans les médias numériques :** En privilégiant les sources dignes de foi et en écartant les contenus suspectés, les instruments de détection peuvent aider à rétablir la crédibilité des plateformes web. Cela offre aux utilisateurs la possibilité d'obtenir des informations plus fiables et de naviguer plus facilement dans un environnement numérique saturé .

— **Contribuer à des prises de décisions plus éclairées :** Que ce soit pour les citoyens, les journalistes, les dirigeants politiques ou les institutions, avoir accès à des informations précises est essentiel. Un système de détection performant offre à chacun la possibilité de fonder ses décisions sur des informations confirmées et authentiques, évitant ainsi les résolutions prises sur le fondement de croyances erronées ou d'informations manipulées [26].

1.11 défis dans la détection des fausses informations

- **Progression des méthodes de désinformation :** Les concepteurs des fausses informations modifient sans cesse leurs techniques pour esquiver les mécanismes de détection. L'emploi de deepfakes (vidéos truquées par intelligence artificielle), de textes produits par des modèles avancés de traitement du langage (tels que GPT-3) et d'altérations discrètes complique la tâche de détection [24].
- **Insuffisance de données labellisées de haute qualité** L'apprentissage des modèles d'intelligence artificielle nécessite de larges ensembles de données vérifiées. Cependant, le marquage manuel est couteux et subjectif, et les bases de données actuelles peuvent présenter des biais ou des incomplétudes [26].
- **Biais dans les algorithmes et leur interprétabilité**Les biais culturels ou politiques présents dans les données d'apprentissage peuvent être reproduits par les algorithmes de détection. Par ailleurs cela peut les amener a faire des erreurs ou prendre parti sans le vouloir ce qui réduit la confiance des utilisateurs [27].
- **Rapidité de propagation comparée à la réactivité des systèmes**Les fausses informations se propagent très rapidement alors que les processus de vérification prennent plus de temps à réagir. Selon une recherche menée par le MIT, les fausses iformations sont partagées 70% plus souvent que les informations véridiques [28].
- **Mise en contexte et vérification multilingue :**Certaines fausses informations utilisent des subtilités culturelles ou linguistiques (propres à un pays ou régions) que les systèmes automatisés ont du mal comprendre dû à leur complexité et leur spécificité, en particulier dans les langues qui sont moins présentes dans les données d'apprentissage [29].

1.12 Approche basée sur l'intervention

En plus de la détection, certaines approches cherchent à freiner de manière active la diffusion des fausses informations. Cela comprend la diminution algorithmique de leur présence sur les plateformes, l'intégration d'alertes contextuels signalant une information disputée ou incorrecte, ainsi que le traitement prioritaire des contenus les plus viraux ou risqués [39] [40]. On envisage également des démarches de prévention mentale visant à préparer les individus à identifier les tentatives de manipulation. Ces

méthodes se situent à la croisée de l'intelligence artificielle, de la psychologie sociale et de la régulation des plateformes. Même si elles sont performantes à grande échelle, des interrogations éthiques majeures émergent concernant la liberté d'expression et l'influence des entités privées sur l'information[23].

1.13 Conclusion

En conclusion, les informations erronées constituent un phénomène complexe aux répercussions importantes sur la société. Leur propagation rapide, amplifiée par les réseaux sociaux et les technologies avancées, pose un défi majeur dans de nombreux domaines. Les techniques modernes de génération de contenus falsifiés, telles que les GANs, rendent leur détection encore plus difficile. Néanmoins, plusieurs méthodes existent, chacune avec ses avantages et ses limites. La lutte contre ce phénomène nécessite donc une approche combinant différents outils, techniques et stratégies adaptées, tout en tenant compte des enjeux éthiques et pratiques qu'il soulève.

Chapitre II :

Détection de fausses informations dans les reseaux sociaux

2.1 Introduction

Aujourd'hui, les réseaux sociaux sont essentiels pour partager des infos partout dans le monde. On peut y accéder facilement, tout se passe en direct et on peut interagir, ce qui a complètement changé notre façon de communiquer, que ce soit entre nous ou en groupe. Mais ce changement a aussi des mauvais côtés, comme la multiplication des fausses informations, La détection automatique de fausses informations dans les réseaux sociaux est devenue un sujet suscitant l'intérêt de nombreuses équipes de recherche. Cet intérêt croissant s'explique, d'une part, par l'actualité, et d'autre part par les défis scientifiques que cela représente. Cette problématique concerne de nombreuses plateformes, chacune ayant des caractéristiques et des spécificités très diverses.

Dans ce contexte, la désinformation est un vrai problème, car elle influence nos idées et nos comportements. De plus, elle met en danger notre façon de vivre ensemble, notre confiance dans les institutions et la stabilité de la politique. Beaucoup de recherches ont été faites pour essayer de repérer et de ralentir la diffusion de ces fausses infos, surtout grâce à l'intelligence artificielle. Mais malgré les progrès, il reste des défis à relever : la variété des formes que prennent ces infos, leur capacité à devenir virales, la complexité des échanges entre les gens et la sophistication des méthodes de manipulation.

Dans ce chapitre, nous présenterons les particularités de la diffusion des fausses informations sur les réseaux sociaux, les méthodes existantes pour les détecter et une synthèse des approches récentes proposées dans la littérature [24].

2.2 Les réseaux sociaux

2.2.1 Définition et classification

Définition :
Un réseau social représente un ensemble organisé d'acteurs, des individus, des groupes ou des institutions — connectés par divers types de relations sociales, comme l'amitié, la coopération, le partage d'informations, le conseil, l'assistance ou l'influence. Ces liens, qu'ils soient formels ou non, symétriques ou asymétriques, constituent un réseau d'interactions connectées [58].

Classification :
La classification des réseaux sociaux est basée sur leurs objectifs, formats, structures de relations et types d'interaction. Cette variété facilite la compréhension de l'émergence de diverses modalités de communication et de communautés sur ces plateformes.

- **Selon la finalité ou le type d'usage :**
  On peut regrouper les réseaux sociaux en fonction de leurs objectifs. Les plateformes sociales (Facebook, Snapchat) cherchent à préserver des relations privées, alors que les plateformes professionnelles (LinkedIn) favorisent les interactions liées à la profession. Les plateformes de médias sociaux comme Twitter et Reddit facilitent le partage instantané d'informations, alors que les services de diffusion multimédia tels qu'Instagram, TikTok et YouTube se concentrent sur la distribution de contenus visuels. Les plateformes communautaires (telles que Discord et les groupes Facebook) rassemblent des utilisateurs partageant des centres d'intérêt similaires, et les applications de rencontres (Tinder, Bumble) encouragent les liens intimes ou amicaux. Cette catégorisation illustre la variété des applications sociales du numérique.
- **Selon le format du contenu :**
  Cette classe se concentre sur le type de contenu partagé par ces réseaux. Certains se focalisent sur le texte et les messages courts, d'autres préfèrent le contenu visuel (images ou vidéos), on a aussi des plateformes basées sur l'audio, telles que Clubhouse ou Spotify. Pour finir, quelques réseaux qualifiés de multiformats fusionnent divers types de contenu, tels que Facebook ou WhatsApp. Cette catégorisation souligne la capacité des plateformes à s'adapter aux diverses formes d'expression et de communication.
- **Selon le mode d'interaction :**
  On peut également classer les réseaux sociaux en fonction de la manière d'interaction entre les utilisateurs. Certains sont basés sur des relations bidirectionnelles, où l'ajout réciproque est indispensable pour communiquer, encourageant ainsi des échanges équilibrés. On trouve le modèle uni-

directionnel adopté par certaines plateformes, où un utilisateur peut suivre un autre à sens unique, sans réciprocité, ce qui facilite la diffusion de contenu à large échelle. Des plateformes comme Snapchat ou Instagram Stories introduisent des interactions éphémères, avec des contenus temporaires qui disparaissent après un temps donné, modifiant ainsi les dynamiques de visibilité et d'engagement. Finalement, quelques réseaux conservent une interaction constante, où les publications restent archivées et accessibles, ce qui a un impact sur la construction de l'identité numérique. Ces diverses modalités d'interaction organisent les pratiques sociales et les tactiques de communication sur le web.

— **Selon le degré d'ouverture :**
On peut aussi catégoriser les réseaux sociaux en fonction de leur ouverture, c'est-à-dire en basant sur la visibilité des contenus et la manière d'accéder aux interactions. Des plateformes telles que Twitter ou YouTube sont publiques, permettant à quiconque d'accéder à leurs contenus, ce qui facilite la propagation rapide et extensive de l'information. D'autres sont semi-ouvertes comme Facebook et Instagram, où les utilisateurs ont la possibilité de gérer la visibilité de leurs publications, cela réduit l'exposition à un public limité. Enfin, quelques réseaux sont privés ou fermés, comme Discord et Slack, qui limitent l'accès à des groupes déterminés, généralement sur invitation. Cette configuration entraîne un manque de transparence et d'accès aux données.

— **Selon la structure du réseau :**
Dans les réseaux sociaux, on peut distinguer deux structures techniques : centralisée ou décentralisée. Les réseaux centraux comme Facebook et Instagram sont centralisés par une entité unique qui gère les contenus, les algorithmes et la modération. Ce modèle simplifie la gestion mais soulève des questions liées au pouvoir, à la transparence et à la censure. En revanche, pour ceux qui sont décentralisés comme Mastodon et Diaspora, basés sur une architecture fédérée, les données sont réparties sur des serveurs indépendants. Cela permet d'avoir plus de liberté et de contrôle aux utilisateurs, mais complique la tâche de la modération. Finalement, la structure du réseau a un impact direct sur la diffusion de l'information et la lutte contre la désinformation [58].

2.2.2 Objectifs et usages

Dans la vie quotidienne, les réseaux sociaux sont devenus des outils incontournables, offrant les meilleurs moyens de communication. On les trouve dans divers domaines comme le partage de contenus et d'informations. Grâce à leurs objectifs et usages, ils jouent un rôle important dans la transformation des pratiques sociales et économiques. On trouve plusieurs objectifs qui se déclinent en usages variés :

— **Communication interpersonnelle instantanée :**
Faciliter une communication rapide et directe entre les utilisateurs par des messages, commentaires, appels et vidéo est l'un des rôles essentiels des réseaux sociaux. Cette interaction peut être synchrone (messages directs) ou asynchrone (messages différés), facilitant la coordination entre les individus en éliminant les distances géographiques. Cette rapidité de diffusion de l'information est devenue le pilier des réseaux sociaux modernes.

— **Partage de contenu multimédia :**
Le partage de contenu multimédia est très important dans les réseaux sociaux, permettant à l'utilisateur de publier des textes et images pour s'exprimer. Plusieurs plateformes offrent des environnements fortement visuels et interactifs, favorisant la participation des utilisateurs à travers des formats attractifs, donnant à ces derniers la chance de devenir des créateurs de contenu.

— **Réseautage professionnel et développement de carrière :**
Dans la gestion des relations de travail et la mise en valeur des parcours de carrière, les réseaux sociaux professionnels jouent un rôle essentiel. Ils permettent aux utilisateurs de construire un profil mettant en valeur leurs compétences, d'étendre leur réseau à travers des connexions ciblées comme collègues, recruteurs, associés, et de partager des contenus professionnels tels que des articles et des annonces. De plus, ces plateformes simplifient la recherche d'emploi et la surveillance sectorielle. Elles créent aussi des espaces semi-publiques combinant exposition personnelle et plan de carrière, contribuant à la construction d'un capital social numérique [63].

— **Veille et diffusion d'information en temps réel :**
Ces dernières années, les réseaux sociaux sont devenus des sources essentielles d'information, offrant aux utilisateurs la possibilité de rester informés en permanence, de repérer les nouvelles tendances

et de partager du contenu en temps réel. Certaines plateformes jouent un rôle important dans la surveillance de l'information, grâce à la rapidité des flux, aux notifications instantanées et à la réactivité des communautés. Les utilisateurs jouent à la fois le rôle de consommateurs et de diffuseurs d'information, en partageant des contenus et en créant leur propre histoire, fréquemment associée à des événements actuels (crises, manifestations, catastrophes). Cette pratique journalistique donne aux médias sociaux la capacité d'agir comme un système de surveillance environnementale [18].

— **Marketing d'influence et publicité ciblée :**
Les réseaux sociaux ont influencé le marketing numérique en devenant des outils stratégiques, grâce à leur efficacité dans la diffusion de l'information. Les marques font appel à des créateurs de contenu (influenceurs, micro-influenceurs) pour présenter leurs produits aux communautés spécifiques de manière plus authentique que la publicité traditionnelle. De même, des plateformes comme Instagram, Facebook, TikTok exploitent les données comportementales des utilisateurs afin de diffuser des annonces publicitaires personnalisées, optimisant ainsi leur impact et leur rentabilité. Ce modèle soulève des questions liées à la vie privée des consommateurs, car il repose sur une économie de l'attention visant à capter et monétiser le temps passé par les utilisateurs [63].

— **Mobilisation et participation citoyenne :**
L'importance des réseaux sociaux ne cesse d'augmenter dans l'organisation collective et la participation citoyenne. Ils permettent aux personnes et groupes d'avoir des moyens efficaces et économiques pour s'organiser, exprimer leurs demandes, attirer l'attention des autorités et soutenir des causes sociales ou politiques. Des mouvements tels que #MeToo et Black Lives Matter illustrent la capacité des plateformes telles que Facebook et Instagram à devenir des espaces de contestation, de visibilisation et de coordination militante. Ils permettent aussi aux citoyens d'être des acteurs de l'information, en créant et diffusant du contenu afin d'influencer le discours public. Cette relation dynamique entre citoyen et institutions donne naissance à une démocratie numérique. Toutefois, cette participation reste inégale et vulnérable aux manipulations, pouvant être censurée ou désinformée, ce qui pose des problèmes de régulation et d'authenticité des informations [48].

FIGURE 6 – Types des réseaux sociaux [78]

2.2.3 Enjeux sociaux, économiques et politiques

— **Enjeux sociaux :**
Les plateformes de médias sociaux modifient radicalement les formes de sociabilité, les méthodes de communication et les interactions sociales. Parmi les enjeux majeurs :

i. **Fragmentation sociale et polarisation :**
Même si les réseaux sociaux facilitent la communication à grande échelle, ils contribuent aussi à la fragmentation du discours public et à la polarisation des points de vue, en utilisant des contenus sur mesure qui renforcent les croyances des utilisateurs. Cela favorise à son tour la formation de bulles de filtre et chambres d'écho où les idées divergentes sont absentes ou rejetées, orientant les individus vers des positions plus radicales. Cette dynamique limite le dialogue contradictoire et favorise la désinformation, comme on a pu le constater lors d'événements tels que les élections américaines de 2016 ou la pandémie de Covid-19 [64].

ii. **Construction de l'identité numérique :**
Lorsqu'un individu crée, développe ou contrôle sa présence en ligne à travers ses profils et activités, ce processus désigne la construction de l'identité numérique. Celle-ci est généralement sélective et vise à répondre à des attentes sociales, professionnelles ou culturelles. Ce processus est façonné par la quête de reconnaissance et par des standards implicites de popularité, d'esthétisme ou de réussite sociale. Cette situation peut créer une forte pression sociale [47].

iii. **Solidarité et inclusion :**
Les réseaux sociaux peuvent également avoir un effet positif en favorisant la solidarité et l'inclusion sociale, notamment pour les groupes marginalisés, les minorités ou les communautés isolées. Ils permettent à des individus de se rassembler autour d'objectifs communs, d'échanger des expériences et de donner une visibilité à des situations fréquemment négligées par les médias traditionnels. On peut aussi construire à partir de ces plateformes des communautés de soutien, comme les luttes féministes ou antiracistes [49].

— **Enjeux économiques**
Les réseaux sociaux sont devenus des acteurs centraux de l'économie numérique, générant plusieurs enjeux majeurs, parmi lesquels :

i. **Capitalisme de surveillance et monétisation des données :** Le capitalisme de surveillance est l'un des enjeux économiques les plus critiques liés aux réseaux sociaux. Il repose sur la collecte massive des données personnelles, générées par les utilisateurs à travers leurs interactions, recherches, partages, voire même leurs déplacements. Ces données, une fois recueillies, sont analysées, classées puis vendues, notamment à des annonceurs, principalement à des fins publicitaires et commerciales. Ce mécanisme transforme l'attention humaine en une marchandise, chaque interaction se convertissant en un bénéfice économique pour les plateformes.

ii. **Émergence de nouveaux métiers et précarisation :** Avec l'essor des réseaux sociaux, de nouveaux métiers numériques ont émergé, tels que les influenceurs, gestionnaires de communauté, producteurs de contenu ou modérateurs de plateformes. Ces professions, issues de l'économie de plateforme, reposent souvent sur un mode de travail flexible, mais aussi fragile et incertain. Par exemple, les influenceurs dépendent de leur popularité et de leurs collaborations commerciales pour assurer leurs revenus, sans bénéficier de protection sociale ni de statut professionnel clairement défini. Ce phénomène tend à effacer la frontière entre la sphère personnelle de l'utilisateur et son activité professionnelle [63].

iii. **Impact sur les industries traditionnelles :** Ces dernières années, les réseaux sociaux ont profondément bouleversé les industries traditionnelles, notamment les médias, la publicité et la musique. Les contenus créés par les utilisateurs, comme les influenceurs, bénéficient d'un accès gratuit et d'une diffusion massive en temps réel, ce qui détruit peu à peu les modèles économiques classiques, en particulier celui de la presse écrite, fondé sur la publicité et la vente. Aujourd'hui, la majorité des revenus publicitaires est captée par les

plateformes numériques telles que Google et YouTube, créant un déséquilibre économique considérable qui menace la pérennité des médias traditionnels [18].

— **Enjeux politiques**
Les réseaux sociaux jouent un rôle central dans les dynamiques politiques modernes, avec quelques enjeux majeurs :

i. **Désinformation et ingérence électorale :** Les réseaux sociaux sont devenus aujourd'hui des canaux privilégiés de la désinformation politique, où des contenus faux ou manipulés se diffusent rapidement pour influencer l'opinion publique et les mouvements électoraux. Par exemple, lors de l'élection présidentielle américaine de 2016, plus de 115 sites pro-Trump ont diffusé de fausses informations, générant plus de 30 millions de partages sur Facebook. Par ailleurs, certaines campagnes ont créé des bots et des trolls pour amplifier ces messages, semer la division et polariser l'électorat. Ces dispositifs sont utilisés par des entités étatiques et non étatiques via des techniques automatisées et des algorithmes (propagande computationnelle), menant des opérations d'ingérence dont l'ampleur et l'efficacité sont préoccupantes. Ils illustrent comment la manipulation algorithmique et les réseaux de désinformation peuvent déformer les débats démocratiques et mettre en péril la fiabilité des processus électoraux [28].

ii. **Censure et surveillance d'État :** Dans plusieurs pays autoritaires, les réseaux sociaux ne sont pas seulement des espaces d'expression, mais aussi des outils de contrôle étatique. Les gouvernements utilisent ces plateformes pour surveiller, censurer ou réprimer les opposants politiques, journalistes, ou activités jugées suspectes. Par exemple, dans certains pays, les autorités combinent censure algorithmique, supervision humaine et régulations numériques sévères pour contrôler les contenus considérés comme sensibles. Sur des plateformes comme Weibo ou WeChat, il est fréquent que des hashtags ou conversations soient automatiquement supprimés ou rendus invisibles, et certains utilisateurs poursuivis pour avoir publié du contenu critique. Ces pratiques ne sont pas limitées aux régimes autoritaires : dans les démocraties, la surveillance des réseaux sociaux pour des objectifs de sécurité nationale ou d'ordre public soulève des questions relatives à la liberté d'expression, à la protection de la vie privée et au droit à l'anonymat. Ainsi, les réseaux sociaux peuvent à la fois mobiliser les citoyens et les contrôler, posant un dilemme fondamental entre liberté numérique et sécurité étatique.

iii. **Régulation, gouvernance et souveraineté numérique :** La régulation des réseaux sociaux est devenue un enjeu majeur dans les débats publics et politiques, face à la montée des abus tels que la désinformation, la manipulation électorale, les atteintes à la vie privée ou les discours de haine. Aujourd'hui, les grandes plateformes exercent une influence massive, souvent hors du contrôle direct des États, ce qui constitue un défi pour la gouvernance démocratique. L'Union européenne a tenté de répondre à cette situation en adoptant des cadres juridiques comme le RGPD en 2018 pour la protection des données, et en 2022 le Digital Services Act, imposant de nouvelles obligations en matière de transparence algorithmique, modération des contenus et lutte contre la désinformation. Ces mesures visent à reprendre le contrôle de la régulation des grandes entreprises du numérique et à instaurer une forme de souveraineté digitale. Par ailleurs, certains gouvernements cherchent à rapatrier leurs infrastructures digitales, développer des alternatives aux plateformes dominantes ou imposer des taxes aux entreprises technologiques. Ces démarches visent à redéfinir stratégiquement l'espace numérique, considéré non seulement comme un lieu de communication, mais aussi comme une zone d'influence géopolitique [27].

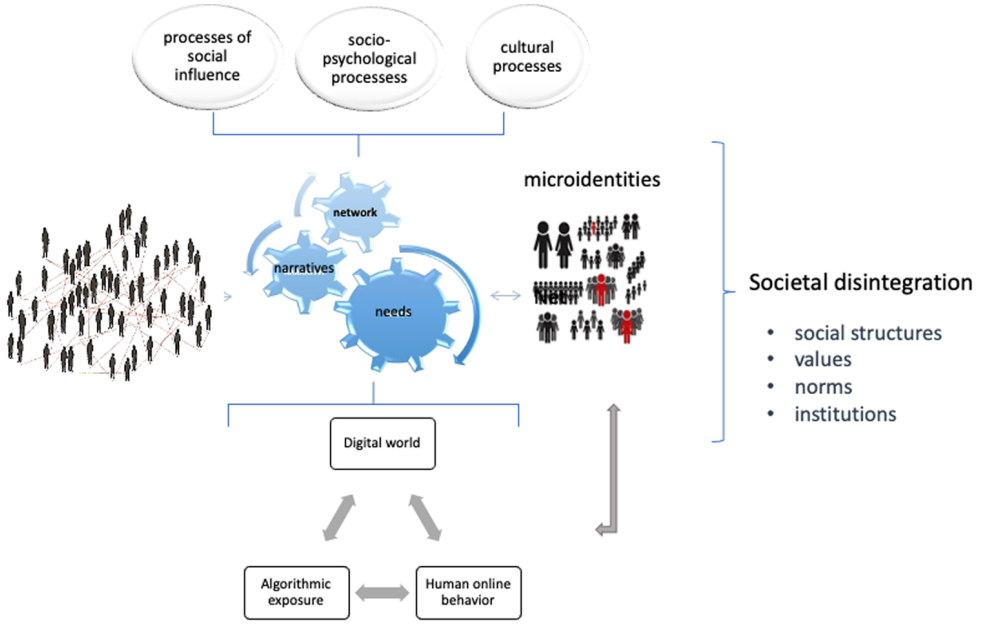

FIGURE 7 – Enjeux des réseaux sociaux [79]

2.2.4 L'intelligence artificielle dans les réseaux sociaux

2.2.4.1 Définition :

L'intelligence artificielle (IA) désigne un domaine scientifique et technique qui vise à développer des systèmes informatiques capables d'imiter un comportement jugé intelligent par un observateur humain. Sur le plan pratique, elle élabore des systèmes dont les actions semblent guidées par un raisonnement. Sur le plan théorique, elle cherche à modéliser la connaissance de manière opératoire, c'est-à-dire applicable pour l'action, la communication et le contrôle. L'IA est ainsi une science expérimentale de la connaissance, fondée sur une représentation explicite et structurée de ce qui peut être compris, traité et calculé. Cette approche nécessite une analyse épistémologique de la nature du savoir et de sa modélisation en informatique [51].

2.2.4.2 Usage de l'intelligence artificielle dans les réseaux sociaux :

L'IA est désormais au cœur du fonctionnement des réseaux sociaux, servant plusieurs finalités principales :

1. **Personnalisation des flux :**
   La personnalisation des flux est l'un des usages les plus visibles de l'IA. Grâce à des algorithmes de recommandation, des plateformes comme Facebook et Instagram sélectionnent et affichent des contenus adaptés aux utilisateurs selon leurs activités (likes, partages, commentaires). Le but est d'attirer l'attention pour maximiser le temps passé sur la plateforme. Ces fonctions reposent principalement sur des techniques d'apprentissage automatique, notamment l'apprentissage supervisé, les réseaux de neurones profonds (deep learning), et des modèles de filtrage collaboratif ou basés sur le contenu.
2. **Modération et détection de contenus :**
   Les réseaux sociaux utilisent largement l'IA pour modérer les contenus. Cela permet de filtrer, détecter et parfois bloquer les publications contenant des éléments problématiques tels que discours de haine, harcèlement, ou situations critiques. Des algorithmes de traitement automatique du langage naturel (NLP) analysent les textes, tandis que la vision par ordinateur traite les images et vidéos. Ces systèmes reposent sur des réseaux de neurones multicouches, avec des classificateurs CNN pour les images ou BERT pour les textes, souvent entraînés sur des exemples annotés.

3. **Chatbots et assistants virtuels :**
   Les agents conversationnels comme les chatbots et assistants virtuels améliorent l'expérience utilisateur sur les réseaux sociaux et applications de messagerie (WhatsApp, Telegram). Ils automatisent les interactions grâce au NLP et aux modèles de deep learning tels que GPT ou BERT, capables de répondre à des requêtes courantes, orienter les utilisateurs vers diverses procédures (réservations, commandes) et gérer plusieurs conversations simultanément. Cela augmente la réactivité et réduit la charge de travail humaine. Cependant, ces assistants deviennent inefficaces face à des demandes ambiguës ou émotionnellement chargées, nécessitant une supervision humaine pour les cas complexes [51].

2.2.5 Analyse des réseaux sociaux

2.2.5.1 Définition

L'étude de la structure des liens entre différents acteurs (individus, organisations, institutions) dans un réseau est réalisée grâce à l'Analyse des Réseaux Sociaux (SNA). Au lieu de se focaliser sur les caractéristiques individuelles de chaque acteur, l'analyse des réseaux sociaux se concentre sur la manière dont ces acteurs sont liés entre eux comme l'échange d'information, la collaboration, l'influence, le soutien. Cette approche est basée sur une modélisation par graphes, où les nœuds symbolisent les acteurs (personnes, entreprises, comptes...) et les arêtes représentent les liens (amitié, collaboration, suivi, échange d'information, etc.). Ces dernières peuvent être pondérées (force de la liaison) ou orientées (relation à sens unique comme « A suit B ») [22].

2.2.5.2 Étapes et concepts de la SNA

1. **Mesure de la centralité** :
   L'évaluation de la centralité est un point essentiel dans l'analyse des réseaux sociaux, utilisée pour évaluer l'importance ou l'influence d'un intervenant (nœud) au sein d'un réseau. On distingue diverses formes de centralité :
   - La centralité en degré reflète la popularité ou accessibilité d'un acteur par le nombre de connexions directes.
   - La centralité d'intermédiaire (*betweenness*) mesure la fréquence d'apparition d'un acteur sur les plus courts chemins entre d'autres.
   - La centralité de proximité mesure la distance moyenne d'un acteur à tous les autres.
   - Des méthodes avancées comme PageRank ou HITS prennent en compte la qualité des liens.
2. **Détection de communautés** :
   Cette détection identifie des sous-groupes fortement interconnectés, mais peu reliés au reste du réseau. Ces structures sociales sont explorées via des algorithmes comme Louvain ou Girvan-Newman. Elles permettent de comprendre les dynamiques sociales, les fractures informationnelles et les zones de polarisation.
3. **Analyse de la densité et de la portée** :
   Elle mesure la cohésion globale et l'accessibilité de l'information dans le réseau. Une forte densité indique une grande connectivité et une diffusion rapide. La portée est liée à la distance moyenne entre nœuds, au diamètre du réseau et aux composantes isolées [22].
4. **Modélisation de la diffusion** :
   Elle vise à modéliser la propagation d'une information ou d'une rumeur, via des modèles à la SIR ou SEIZ. Ces approches simulent les scénarios de propagation et identifient les nœuds influents. Des méthodes modernes incluent les Graph Neural Networks (GNNs) [20].
5. **Analyse des attributs et homophilie** :
   Cette analyse intègre les caractéristiques individuelles des nœuds (âge, sexe, opinion...) pour comprendre les dynamiques sociales. Le principe d'homophilie, « qui se ressemble s'assemble », joue un rôle crucial dans la formation de liens et la polarisation [25].
6. **Visualisation et cartographie** :
   Ces techniques permettent une représentation visuelle du réseau. Les outils comme Gephi, NodeXL, Pajek ou Cytoscape aident à identifier les communautés, les acteurs clés et à analyser la structure globale. Elles facilitent la compréhension et la communication des résultats [24, 35].

2.3 Méthodes de détection des fake news dans les réseaux sociaux :

La détection automatique de fausses informations sur les réseaux sociaux est un domaine de recherche multidimensionnel, qui a donné lieu à une grande variété d'approches exploitant différentes « modalités » des publications (texte, image, interactions sociales, propagation, contexte événementiel, etc.).

2.3.1 Travaux relatifs

2.3.1.1 Approches fondées sur les caractéristiques sociales

Les méthodes basées sur les attributs sociaux utilisent directement les données concernant les utilisateurs et leurs échanges sur les réseaux sociaux, au lieu de se concentrer uniquement sur le contenu écrit ou visuel. Le concept clé est que les attributs tirés du réseau social (profil de l'utilisateur, structure de son réseau, comportements d'engagement) peuvent offrir des indices distinctifs pour détecter la propagation de fausses informations. Nous identifions principalement trois types de descripteurs sociaux :

Descripteurs utilisateur :

— **Individus** : ils identifient l'auteur initial d'un message (ou tout participant impliqué dans sa diffusion). Cela comprend notamment des caractéristiques du profil (date d'enregistrement, âge, sexe, lieu de résidence déclaré, logiciel client utilisé) ainsi que des indicateurs quantitatifs comme le nombre d'abonnés, le nombre de personnes suivies (amis), le total des messages publiés, etc. Par exemple, Morris et al. (2012) se servent du descripteur "client "(type de programme ou d'application utilisée pour publier) et du qualificatif "emplacement "(précisant si le tweet ou le post est diffusé depuis le lieu de l'événement) afin de décrire le comportement de l'utilisateur. Castillo, Mendoza et Poblete (2011) ont démontré que le rapport entre les abonnements et les abonnés, l'ancienneté du compte et la régularité de ses publications peuvent déjà servir à identifier les comptes qui diffusent des rumeurs ou de fausses actualités sur Twitter.

— **Groupes d'utilisateurs** : ce sont des descripteurs consolidés à l'échelle d'un groupe d'utilisateurs contribuant à la diffusion d'une même information. Par exemple, Yang et al. (2012) regroupent des indicateurs tels que le taux d'utilisateurs certifiés ("verified "), le nombre moyen d'abonnés, la densité du sous-réseau constitué par ces utilisateurs, et ainsi de suite. Ces indicateurs de groupe visent à saisir des comportements collectifs (par exemple, la présence simultanée d'un grand nombre de comptes récents ou peu interconnectés diffusant une même information).

— **Stabilité dans le temps** : Kwon, Cha et Jung (2017) ont exploré l'évolution de certaines caractéristiques utilisateur au fil du temps. Ils ont constaté que pour la détection précoce des fausses informations, les indicateurs associés aux profils (nombre d'abonnés, ancienneté) et à l'engagement (taux de retweet, ratio likes/retweets) sont efficaces dès le début, tandis que les indicateurs structurels ou temporels (associés à la structure du réseau de diffusion) se révèlent plus discriminants sur une période prolongée.

Descripteurs de propagation :

Bien qu'ils soient fréquemment associés à une autre catégorie ("approches basées sur le parcours du message "), certains travaux incluent également dans les "informations sociales "les descripteurs qui illustrent comment l'information se propage à travers le réseau :

— Nous représentons le réseau de retweets/reposts sous la forme d'un arbre de diffusion, où chaque nœud correspond à un utilisateur qui a partagé le message. Chaque nœud présente des attributs (profil de l'utilisateur, nombre d'abonnés, taux d'engagement) qui sont intégrés dans une série chronologique. Kotteti et al. (2020) font appel à cette modélisation pour élaborer un classificateur RNN+CNN capable de déterminer en cinq minutes si une information est erronée, se basant uniquement sur la séquence d'attributs utilisateurs tout au long du processus de diffusion.

— Liu et Wu (2018) utilisent également des arbres de propagation pour saisir à la fois l'aspect global (largeur, profondeur) et local (rapidité du retweet, influence de chaque nœud) d'une rumeur sur Twitter et Sina Weibo, parvenant à plus de 85% de précision en 5 minutes.

Descripteurs temporels :

Ces indicateurs évaluent la dynamique dans le temps des interactions associées à une publication, c'est-à-dire le nombre de retweets et commentaires par période, le taux de création de nouveaux utilisateurs partageant l'information, etc. Par exemple, Ma et al. (2016) identifient des indices comme l'inclinaison du graphique des retweets, le nombre de réactions par heure, qui se révèlent bénéfiques pour repérer des pics inhabituels liés aux fausses informations.

Zhao, Resnick et Mei (2015) tirent des attributs linguistiques temporels des commentaires (l'apparition de questions telles que "Est-ce vrai ? ", manifestations d'incertitude), qui, lorsqu'ils sont analysés à la lumière des caractéristiques de l'utilisateur, perfectionnent la détection anticipée des rumeurs.

2.3.1.2 Modélisation de la propagation (cheminement du signal)

Les méthodes basées sur la diffusion se concentrent sur comment une information (qu'elle soit correcte ou incorrecte) se propage à travers le réseau social. L'hypothèse de base est que la structure et la dynamique de propagation d'une fausse information possèdent des traits distinctifs (rapidité, profondeur, amplitude, cohérence des sources) qui permettent de la différencier d'une information authentique. On identifie deux paradigmes principaux :

1. Propagation de la crédibilité :

Le modèle de diffusion de crédibilité se fonde sur l'établissement d'un réseau de confiance qui lie diverses sortes d'entités : utilisateurs, messages et événements. Chaque nœud du réseau reçoit une valeur initiale de crédibilité qui est ensuite progressivement diffusée en fonction de la structure du réseau et des poids assignés aux connexions, jusqu'à atteindre un état stable. Cette méthode permet de tirer parti des liens entre les entités pour améliorer la solidité de la détection, au lieu d'examiner chaque publication individuellement [49].

Construction du réseau de crédibilité

- Les nœuds peuvent représenter :
  - Les messages ou tweets (chaque publication est un nœud).
  - Les utilisateurs (chaque compte Twitter, Facebook, etc.).
  - Les événements (ensembles de messages liés au même sujet ou hashtag).
- Les liens sont pondérés selon des relations sémantiques ou d'interaction :
  - Un utilisateur est lié à chaque message qu'il a publié ou retweeté.
  - Deux messages sont liés s'ils apparaissent fréquemment ensemble sous le même hashtag ou s'ils partagent un contenu très similaire (mots-clés, URL partagée).
  - Un message est lié à l'événement qu'il illustre.
- La valeur initiale de crédibilité d'un message peut être estimée par un classificateur de type "texte pur "(par exemple, sur des descripteurs linguistiques ou lexicaux inspirés de Castillo et al. 2011). L'hypothèse est que les messages dont le contenu paraît "suspect "obtiennent une crédibilité initiale faible.

Propagation itérative

- Sous des hypothèses de cohérence (un utilisateur crédible est peu susceptible de relayer une fake news) et de régularité (deux publications crédibles se confirment mutuellement), on diffuse la crédibilité des nœuds selon un algorithme de type PageRank.
- Par exemple, Gupta, Zhao & Han (2012) proposent de pondérer les liens de façon à ce que :
  - La crédibilité d'un événement est soutenue par la somme des crédibilités de ses messages associés, pondérées par la réputation de leurs auteurs.
  - La crédibilité d'un message est renforcée s'il est relayé par des utilisateurs eux-mêmes jugés crédibles.

  - La crédibilité d'un utilisateur est ajustée en fonction de la crédibilité des messages qu'il a partagés.
- Ces valeurs sont itérées jusqu'à convergence, ce qui donne, au final, une estimation de la crédibilité de chaque message, événement et profil utilisateur du réseau.

Extensions multi-niveaux

Jin et al. (2016) améliorent ce schéma en introduisant un graphe de crédibilité à trois couches (messages ↔ sous-événements ↔ événements) :

- La couche message regroupe les publications individuelles.
- La couche sous-événement regroupe des "sous-thèmes "au sein d'un même événement.
- La couche événement regroupe tous les messages traitant du même sujet global (ex. "incendie dans la cathédrale ").
- En supposant que des publications sous un même sous-événement partagent une même vérité, la crédibilité se propage verticalement et horizontalement dans ce réseau tripartite, améliorant la cohérence globale de la détection.

2. Classification des chemins de propagation :

Cette deuxième méthode envisage la propagation d'un message sous la forme d'un arbre (ou graphe) de retweets/reposts : chaque nœud symbolise un utilisateur ayant diffusé l'information, et chaque arête incarne l'acte de retweet (ou de partage). On utilise donc la séquence chronologique des attributs des utilisateurs le long de chaque trajectoire de diffusion pour former un classificateur (généralement une combinaison hybride RNN+CNN). L'hypothèse fondamentale est que la structure temporelle et topologique d'une diffusion de fausses informations se distingue nettement de celle d'une information authentique.

Représentation des chemins comme série temporelle multivariée :

- Chaque chemin de propagation correspond à la chaîne des utilisateurs :

$$\text{User}_0 \to \text{User}_1 \to \text{User}_2 \to \cdots \to \text{User}_k,$$

  où User_0 est l'auteur initial et User_i est celui qui retweete User_{i-1} à l'instant t_i.
- À chaque position i on associe un vecteur de caractéristiques extraites de User_i au moment du retweet (nombre de followers, ancienneté du compte, ratio likes/retweets, type de client utilisé, etc.).
- Le chemin complet devient ainsi une séquence multivariée

$$\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_k,$$

  où $\mathbf{x}_i \in R^d$ capture les attributs du i-ème utilisateur[32].

Architecture RNN + CNN pour la détection précoce :

Kotteti et al. (2020) proposent d'alimenter un RNN (Recurrent Neural Network) par cette suite de vecteurs pour saisir la dynamique globale de la propagation (variations long-terme), puis de faire passer les mêmes vecteurs dans un CNN (Convolutional Neural Network) pour capturer les motifs locaux (p. ex. un pic soud'un d'utilisateurs peu fiables en début de propagation).

- La sortie du RNN reflète la tendance d'évolution de la crédibilité au fur et à mesure de la propagation.
- La couche CNN extrait des n-grammes temporels de vecteurs utilisateur, utiles pour détecter des "explosions "de retweets venant de comptes suspects en quelques instants.
- En combinant RNN et CNN, le modèle peut atteindre un bon compromis entre détection précoce (quelques minutes après le début de la diffusion) et robustesse (prise en compte de la topologie globale).
- Sur des jeux de données réels (Twitter16, Weibo), cette méthode atteint 85 % de précision en 5 minutes pour Twitter et 92 % pour Sina Weibo.

Variations et améliorations :

— Liu & Wu (2018) modélisent également l'arbre de propagation comme un graphe :
  — Ils calculent des descripteurs globaux (profondeur maximale, largeur moyenne des niveaux, vitesse moyenne de retweet) et des descripteurs locaux pour chaque nœud (influence individuelle, degré sortant), puis alimentent un classifieur RNN + CNN.

— Ma et al. (2015, 2016) exploitent une approche à noyau de graphes (propagation structure kernel) pour comparer directement la structure des arbres de diffusion, sans passer par une séquence temporelle explicite.

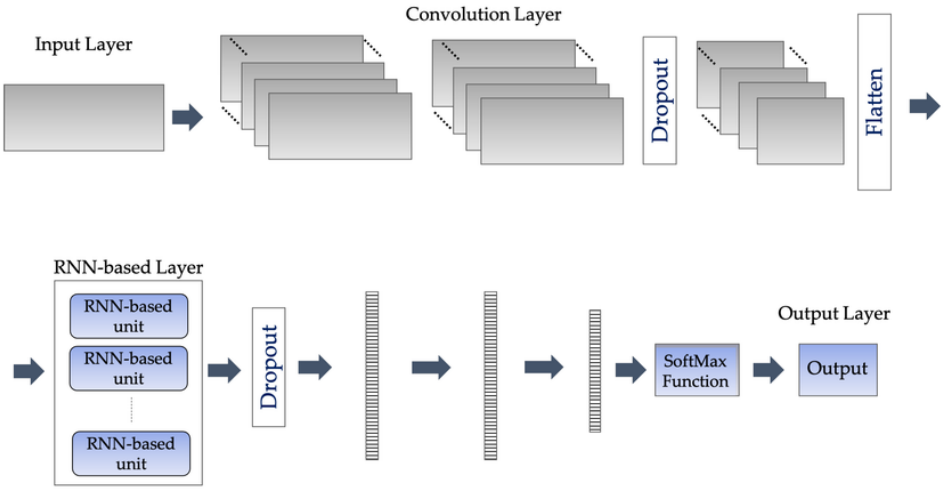

FIGURE 8 – Architecture RNN+CNN [80]

2.3.1.3 Approches centrées sur le contexte événementiel

Les méthodes basées sur le contexte (ou « event-based ») soutiennent que l'information ne doit pas être étudiée de manière indépendante, mais plutôt dans le cadre global de l'événement auquel elle est associée. L'hypothèse repose sur le fait que les schémas globaux (quantité de flux, signification commune, cohérence dans le temps) concernant un événement offrent des indices significatifs pour différencier une fausse information d'une information authentique.

1. Principe général

Au lieu d'étudier chaque publication de façon individuelle, ces techniques commencent par regrouper les messages associés à un même "événement "(comme un hashtag, une série de mots-clés ou une contrainte spatio-temporelle). Ensuite, elles déterminent des traits caractéristiques à ce niveau, tels que le nombre total de publications, l'aspect sémantique du propos ou les tendances en matière d'engagement. Elles s'efforcent ensuite d'établir la cohérence interne entre ces messages et leur contexte—en examinant, par exemple, la fiabilité des sources principales, l'harmonie entre le contenu textuel et les métadonnées ou le degré de compétence des intervenants majeurs—afin de nourrir un classificateur (réseau de neurones ou algorithme classique) qui se prononce sur l'événement dans son ensemble, pas uniquement sur des publications spécifiques [67].

2. Méthodologies typiques

Clustering & Filtrage par événement

— On constitue un ensemble d'événements à détecter en scrutant les flux sociaux (flux Twitter, groupes Facebook, forums). Chaque événement regroupe :
  — Un ensemble de publications (tweets, posts, commentaires) formant une conversation autour d'un lieu, d'un sujet ou d'un hashtag.

  - Des mots-clés / entités extraits via un algorithme de reconnaissance d'entités nommées (NER), de similarité sémantique ou de co-occurrence.
- Par exemple, dans Kotteti et al. (2020), pour chaque événement (ex. "attaque à Paris "), on recueille tous les tweets avec l'hashtag associé, puis on agrège métriques textuelles (fréquence d'apparition d'entités comme "terrorisme ", "gouvernement "), métriques temporelles (rythme de publication) et métriques sociales (répartition des auteurs selon leur ancienneté).

Extraction de descripteurs globaux d'événement

- **Volume de discussion :**
  - Nombre total de publications par unité de temps, vitesse de montée en charge (slope) et décroissance (heavy tail).
  - Un pic anormalement brusque, suivi d'une forte chute, peut indiquer une fake news (diffusée massivement par un noyau de comptes automatisés).
  - Ma et al. (2016) ont montré que de telles caractéristiques suffisent à séparer des rumeurs émergentes des flux d'actualité normaux.
- **Diversité lexicale & sémantique :**
  - Calcul de la richesse lexicale (nombre unique de n-grammes, diversité d'entités nommées), distance sémantique (modèles Word2Vec ou TF-IDF) entre messages du même événement.
  - Les fake news tendent à employer un vocabulaire plus restreint et à citer peu de sources externes fiables.
  - Zhu et al. (2015) ont extrait les topics latents via LDA pour mesurer la cohérence sémantique d'un événement : un faible nombre de topics dominants (< 3) avec une forte concentration peut signaler une manipulation.
- **Profil des sources principales :**
  - Répartition des auteurs selon
  - Ancienneté du compte (date d'inscription au réseau).
  - Crédibilité présumée (nombre moyen de followers, ratio followers/following, présence de "verified badge ").
  - Les fake news émergent souvent d'un sous-ensemble d'utilisateurs récents ou très actifs (< 3 mois) qui publient en rafale.
  - Par exemple, Castillo et al. (2011) rapportent qu'un ratio abonnements/abonnés atypique est un bon indicateur de rumeur.

Modèles de classification d'événements

- **Algorithmes traditionnels (SVM, Random Forest) :**
  - On alimente le classifieur de features agrégées (volume, diversité, profil des sources).
  - Ex. maillats d'expériences sur PHEME montrent qu'un SVM combinant variables temporelles (profil du pic) et variables sémantiques (diversité LDA) atteint $\sim 80\%$ de F1 en 30 minutes d'observation.
- **Réseaux de neurones (CNN/LSTM) appliqués aux séquences globales :**
  - On encode chaque événement comme une série temporelle multivariée :
  - Composantes : $[\text{volume}_t, \text{diversité}_t, \text{moyenne\_followers}_t, \dots]$ sur $t = \{t_1, \dots, t_k\}$.
  - Un CNN extrait des motifs locaux (ex. une explosion de volume couplée à une chute brutale de diversité lexicale).
  - Un LSTM saisit la tendance long-terme (augmentation graduelle de la part de comptes peu crédibles).
  - Ruchansky et al. (2017) proposent le modèle CSI (Capture, Score, Integrate) où :
    1. Capture des embeddings pour texte, utilisateur et propagation.
    2. Score un vecteur événementiel résumé via un CNN sur la structure de diffusion.
    3. Integrate ces vecteurs pour décider fake vs non-fake.
  - CSI dépasse un SVM basique de 7 points de F1 sur des données Twitter annotées [33].

2.3.1.4 Approches multimodales

Les méthodes multimodales intègrent diverses modalités (texte, image/vidéo, audio, métadonnées sociales, diffusion) afin de tirer parti de la complémentarité des informations variées. Dans un univers où les fausses informations peuvent se manifester à travers d'images manipulées ou de vidéos modifiées, une simple étude de texte ne suffit pas ; il est nécessaire d'intégrer l'examen du contenu visuel et la concordance entre l'image et le texte.

1. Principe général

Extraction de modalités indépendantes :

- **Texte** (contenu textuel du message) : embeddings Word2Vec/GloVe, TF-IDF, analyse syntaxique et stylistique.
- **Image/vidéo** (photo jointe, capture d'écran) : CNN (ex. ResNet, VGG) pour extraire des embeddings visuels, détection de manipulations (incohérences JPEG, artefacts).
- **Métadonnées sociales** : caractéristiques utilisateurs (followers, réputation), données spatio-temporelles (géolocalisation, horodatage).
- **Propagation** : séquences temporelles d'engagement (nombre de retweets, likes, commentaires).

Fusion des représentations

- **Early fusion** : concaténation des embeddings de chaque modalité avant le classifieur.
- **Late fusion** : classifieurs séparés (texte, image, social) dont les scores sont agrégés (moyenne pondérée, vote majoritaire).
- **Hierarchical fusion** : fusion progressive : on combine d'abord texte + image pour chaque message, puis on agrège au niveau événementiel avec propagation et métadonnées sociales.

Classification

- **Réseaux profonds hybrides** : architectures "multi-branch "(une branche CNN texte, une branche CNN image, une branche LSTM propagation) qui fusionnent leurs sorties dans une couche fully connected, puis Softmax pour la décision finale.
- **Modèles graphiques** : inclusion des embeddings multimodaux dans un "Graph Neural Network "où chaque nœud représente un utilisateur ou un message, lié par des arêtes (retweet, mention), et où les attributs node-level incluent texte, image et scores de propagation [37].

2. Exemples concrets

CSI (Capture, Score, Integrate) – Ruchansky et al. (2017)

- **Capture :**
  - Texte : embeddings via RNN sur le contenu du post.
  - Image : features extraites par CNN (Inception).
  - Social : vecteurs d'attributs utilisateur (nombre de followers, ratio followers/following).
- **Score :**
  - Chaque dimension (texte, image, social) produit un score partiel via un classifieur binaire.
  - Un module Attention pèse l'importance de chaque modalité selon l'événement courant.
- **Integrate :**
  - Les scores partiels sont concaténés, puis un réseau fully connected entraîne la décision finale.
- **Résultats** : sur un corpus Twitter annoté, la F1 passe de 0,82 (seulement texte + social) à 0,89 en ajoutant l'image et la propagation.

MVNN (Multi-View Neural Network) – Xu et al. (2018)

- **Branches** : texte, propagation, utilisateurs et visuel.
- **Texte** : CNN 1D sur le message pour capturer n-grammes discriminants.
- **Propagation** : LSTM appliqué à la séquence temporelle du nombre de retweets par minute.
- **Utilisateur** : vecteur dense construit à partir de l'ancienneté, du nombre de followers, fréquence de tweet.
- **Visuel** : CNN (VGG16) appliqué à l'image jointe, réduit en embedding 256-D.
- **Fusion** : chaque embedding passe par une couche Dense (128 unités), puis concaténation globale, Dropout 0,5, et Softmax.
- **Perf** : 0,91 de précision sur un dataset Weibo, dépassant les méthodes unimodales (max. 0,82).

DeepTrust – Nguyen et al. (2019)

- **Objectif** : détecter image + légende falsifiée.
- **Technique :**
  1. OCR sur l'image pour extraire tout texte incrusté (ex. fausses manchettes).
  2. Fusion texte-image :
     - Texte extrait (OCR + légende) passe dans un BERT finetuné pour embeddings.
     - Image passe dans EfficientNet pour vecteur visuel.
  3. Ressemblance inter-modale : on calcule un score de cohérence sémantique entre l'embedding texte et l'embedding image (cosine similarity).
  4. Classifieur final :
     - Score "cosine "
     - Score de crédibilité de la source (compte utilisateur)
     - Score d'engagement (vitesse de propagation)
     - Tous agrégés dans un MLP qui prédit fake vs non-fake.
- **Résultats** : 0,94 d'AUC (Area Under the ROC Curve) sur un corpus de tweets avec images fausses (doctoring).

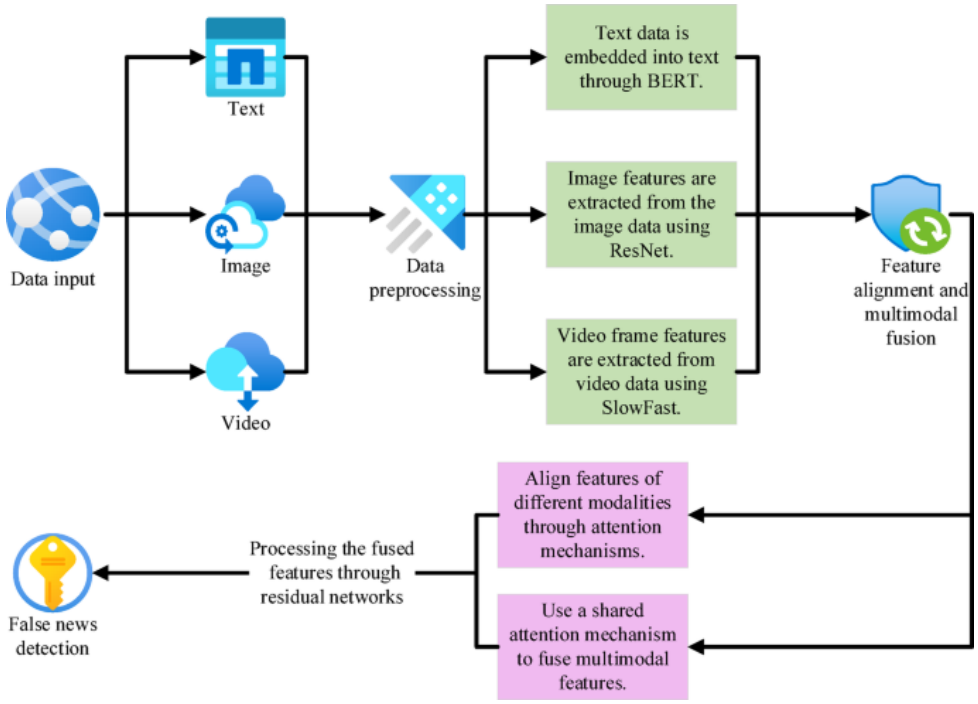

FIGURE 9 – Approche multi-modale[81]

2.3.2 Synthèse des approches existantes

2.3.2.1 Analyse des interactions sociales et caractéristiques utilisateurs

Objectif

Ces méthodes visent principalement à identifier des comportements ou des profils inhabituels susceptibles de propager ou de transmettre de fausses informations. On cherche particulièrement à identifier certains types de comportements, tels que les comptes automatisés (bots) ou semi-automatisés, les comptes nouvellement créés qui n'ont pas d'historique fiable, les utilisateurs dont le réseau est peu connecté (faible regroupement ou nombre limité de liens réciproques), ainsi qu'une activité excessive par rapport à la notoriété du profil (par exemple, un volume extrêmement important de publications sans précédent établi). On suppose généralement que, dans la plupart des situations, les acteurs malintentionnés se basent sur des comptes "bottom-up "(créés expressément pour la propagation) ou sur des bots afin de diffuser rapidement une information fallacieuse. En se concentrant sur ces signaux sociaux, nous pouvons identifier la provenance avant même d'examiner le contenu textuel ou visuel du message.

Techniques

1. Métriques statistiques d'influence et de réputation

— L'indicateur du nombre de followers par rapport au nombre d'abonnements (amis) : un déséquilibre notable entre les deux (peu d'abonnés en comparaison aux abonnements) ou l'opposé peut indiquer un compte factice ou trop récent. Selon Castillo, Mendoza et Poblete (2011), ces indicateurs sont déjà distinctifs pour identifier les comptes qui déclenchent des rumeurs sur Twitter.

— Durée du compte : La date d'enregistrement peut être utilisée pour apprécier la maturité du profil. De nombreux comptes propagateurs de fausses informations sont créés juste avant leur première diffusion trompeuse.

— Indice de réputation : combinaison des indicateurs précédents, éventuellement ajustée par un signet "vérifié "(pour les plateformes qui le proposent), ou par une note de confiance calculée selon le degré d'interaction authentique reçu par le compte (mentions, retweets provenant de profils bien établis).

2. Historique de contenu

— Volume et fréquence de publication : un compte qui publie de façon excessive, sans phase d'inactivité ou variation graduelle, pourrait signaler un comportement automatisé. Nous calculons la moyenne quotidienne de posts et leurs écarts types, ainsi que la régularité (étude de la série temporelle des timestamps).

— Thématique variée : l'examen des thèmes traités (à travers le TF-IDF ou le regroupement thématique) permet d'identifier les comptes fortement focalisés sur la propagation d'un sujet particulier (généralement la désinformation), sans autre trace de centres d'intérêt.

3. Mesures d'engagement

— Ces indicateurs, à savoir le nombre de retweets, likes et commentaires par publication, évaluent la portée et le niveau de viralité d'un post. Un compte affichant un engagement extrêmement élevé sur des publications dénuées de sens ou alimentant des rumeurs suscite des soupçons.

— Proportion retweets/contenu original : un compte qui se contente de repartager ou de retweeter, sans produire de contenu novateur, est généralement un bot de diffusion conçu pour propager un message.

4. Critères de confiance du profil

— Réseau de relations : en élaborant le réseau personnel (le sous-graphe d'amitié ou d'abonnements associés au compte), nous déterminons des indicateurs tels que le coefficient d'agglomération, le degré moyen des interlocuteurs ou encore la centralité de proximité. Il est possible qu'un compte faiblement intégré (avec un faible regroupement) ou isolé (degré d'isolement très faible) soit "créé "dans le but de diffuser du contenu.

— Hétérogénéité des interactions : si un compte interagit majoritairement avec un cercle restreint d'autres comptes récents ou douteux, cela suggère une diffusion orchestrée. Selon Maigrot et al. (2017), l'étude de la structure du réseau des retweets (qui retweete qui) sur des périodes temporelles restreintes permet d'identifier des accélérations inhabituelles attribuées à des bots.

Avantages

— **Disponibilité rapide des données :** Dès la publication initiale d'un message, toutes les métriques sont à votre disposition (profil public, métadonnées sociales, historique visible). À la différence de l'analyse de contenu qui requiert parfois une manipulation complexe du sens ou de l'analyse de diffusion qui demande un minimum de partages, l'extraction des caractéristiques utilisateur peut être initiée sans délai pour générer une alerte anticipée.

— **Robustesse face aux manipulations de contenu :** Ces indicateurs ne sont pas dupés par un texte sophistiqué ou une image modifiée, car ils se fondent sur le comportement du compte plutôt que sur le contenu sémantique ou visuel. Même si un opposant parvient à rédiger un message très convaincant, si le compte est fraîchement créé, avec un taux d'abonnés en déséquilibre et un modèle de retweet suspect, il sera signalé comme potentiellement nuisible.

— **Facilité d'implémentation** : l'évaluation de métriques telles que le nombre d'abonnés, l'ancienneté, le volume de publications ou le taux de retweets exige peu de ressources informatiques. Cela permet l'intégration de ce module dans un flux de traitement en temps réel avec une complexité algorithmique proportionnelle au nombre de comptes analysés.

Inconvénients

— **Faux positifs générés par des comptes légitimes très actifs** : les influenceurs, les journalistes ou les figures publiques peuvent afficher des indicateurs semblables à ceux de comptes douteux (nombre important de publications, engagement fort, large réseau). Si une contextualisation plus précise n'est pas fournie, ces profils pourraient être erronément identifiés comme diffuseurs de fausses informations, ce qui diminue la crédibilité du système.

— **Limites lors de la diffusion par des comptes "inoffensifs "** : si une fausse information est d'abord propagée par un compte respecté (par exemple, un compte d'actualité reconnu détourné ou piraté), puis partagée par des utilisateurs ayant une bonne réputation, l'approche qui repose uniquement sur les interactions sociales perd de sa pertinence. Ainsi, aucun profil utilisateur inhabituel n'est présent dans cette situation.

— **Éventuelles stratégies d'évasion** : Il est possible pour des adversaires d'investir afin de "légitimer "des comptes avant de les exploiter : ils gonflent délibérément le nombre de leurs abonnés, engagent des interactions avec des comptes dignes de confiance, ou mettent en place un historique varié de publications pour dissimuler une intention nuisible. Ces actions rendent graduellement ces indicateurs moins fiables, puisque le comportement des robots se rapproche de plus en plus de celui des véritables utilisateurs humains [33].

2.3.2.2 Modélisation de la propagation (cheminement du signal)

Objectif

L'objectif de la modélisation de la propagation est d'utiliser la structure et le mouvement temporel d'un graphe de diffusion pour différencier le comportement viral d'une fausse information de celui d'une information authentique. En pratique, on envisage chaque message (tweet, post, publication) comme l'origine d'un arbre de retweets ou de reposts, où chaque point représente un utilisateur qui a propagé l'information à un moment précis, et chaque lien symbolise l'action de retweeter (ou de partager). Ainsi, l'idée est de détecter au fil du temps des signaux épidémiques spécifiques aux fake news comme des hausses rapides d'engagement, une profondeur (nombre de retweets) remarquablement basse ou hautement élevée, et une amplitude (nombre d'utilisateurs à chaque étape de l'arbre) inhabituelle pour identifier une propagation suspecte lors des toutes premières minutes.

Techniques utilisées

1. Arbres de propagation

On débute en créant un arbre orienté où chaque nœud symbolise une instance de diffusion du message : l'auteur initial est la racine, et chaque descendant est un usager qui a retweeté ou reposté le message de son parent. Chaque nœud comporte des attributs tels que les vecteurs de caractéristiques utilisateurs obtenus lors du moment du retweet (nombre de followers, ancienneté du compte, ratio likes/retweets, type de client utilisé, etc.).

Par exemple, Liu & Wu (2018) élaborent cet arbre de propagation pour chaque rumeur sur Twitter et Sina Weibo : ils recueillent des données pour chaque nœud concernant le degré d'entrée/sortie, la fréquence locale de retweet ainsi que la réputation de l'utilisateur dans le but d'appréhender aussi bien la structure topologique (profondeur, largeur) que la dynamique locale (vitesse de retweet, influence individuelle).

2. Séries temporelles de vecteurs de caractéristiques

— Une fois l'arbre construit, on représente chaque chemin de propagation (depuis la racine jusqu'à un nœud feuille) comme une séquence temporelle $[x_0, x_1, \ldots, x_k]$, $\in R^d$ où x_i est le vecteur de caractéristiques de l'utilisateur i au moment t_i du retweet.

— Kotteti et al. (2020) mettent en œuvre cette représentation : ils nourrissent un Réseau de Neurones Récurrents (RNN) avec la séquence complète afin de comprendre l'évolution globale de la propagation, tout en passant ces vecteurs identiques par un Réseau de Neurones Convolutionnels (CNN) consacré à la dimension temporelle pour repérer des modèles locaux (par exemple, une succession de retweets effectués par des comptes récents).

3. Graph Neural Networks (GNN)

— Au lieu d'examiner chaque voie séparément, nous pouvons visualiser l'intégralité de l'arbre (ou même le sous-graphe de propagation intégrale) comme un graphe où chaque nœud est doté d'attributs multimodaux (profil utilisateur, indicateurs d'engagement temporel, note initiale de crédibilité) et chaque lien représente une action de retweet ou de mention.

— Un Graph Neural Network permet alors de transmettre l'information entre les nœuds voisins, en apprenant à doser l'impact des utilisateurs adjacents (par exemple, un nœud central ayant un grand nombre de voisins peu fiables contribue à réduire la crédibilité de ces derniers) ; cette méthode saisit donc simultanément la topologie globale (comment l'information se répand dans le réseau) et la dynamique locale (l'influence comparative de chaque utilisateur sur ses proches).

4. RNN/CNN hybrids

— En pratique, les architectures hybrides combinant RNN et CNN se révèlent particulièrement efficaces pour la détection anticipée. Le RNN (généralement un LSTM ou GRU) encode la tendance à long terme de la chaîne de retweets (synthétisation de l'expansion graduelle de la diffusion), tandis que le CNN appliqué aux vecteurs temporels saisit des motifs localisés (comme une montée soudaine de retweets provenant de comptes récents).

— Selon Kotteti et Liu & Wu, l'union de ces deux méthodes leur a permis d'obtenir environ 85 % de précision sur des données Twitter en à peine 5 minutes, et 92 % sur Sina Weibo. Ceci met en lumière l'aptitude de ces modèles à réagir promptement face aux premiers indices d'une propagation virale suspecte.

Un exemple représentatif est le modèle **GRACE (Graph-based Attention for Coherent Explanation)**. GRACE exploite une architecture de graphe avec mécanisme d'attention pour capter les relations complexes entre utilisateurs et messages. Ce modèle ne se contente pas de classer une information comme vraie ou fausse : il fournit également une *explication cohérente* en mettant en évidence les parties du graphe de propagation les plus déterminantes dans la décision. Ainsi, GRACE combine performance de détection et interprétabilité, ce qui le rend particulièrement adapté à l'analyse de la diffusion de fausses informations sur les réseaux sociaux.[65]

Avantages

— **Capture de la dimension "épidémiologique "** : en représentant la diffusion comme un processus de contagion, ces approches saisissent la cadence (intervalle de temps entre les retweets), la profondeur (quantité de générations dans l'arbre) et l'ampleur (nombre d'utilisateurs à chaque niveau), ce qui s'avère fréquemment plus discriminant que l'étude du contenu exclusivement.

— **Détection anticipée** : dès l'apparition de quelques retweets, un modèle RNN+CNN est en mesure de déclencher une alerte, offrant ainsi la possibilité d'intervenir promptement avant que la fausse information ne se répande à grande échelle.

— **Résilience face aux contenus élaborés** : malgré un texte ou une image très convaincante, la diffusion d'une fausse nouvelle révèle souvent des profils d'utilisateurs (comptes récents, fortement connectés à d'autres comptes douteux) qui sont compliqués à répliquer en grande quantité sans laisser de traces dans l'arbre de propagation.

Inconvénients

— **Dépendance à un faible historique de retweets** : si une information (véridique ou erronée) n'est que peu diffusée, par exemple dans un cercle privé ou un réseau limité, le nombre de retweets accessible est trop faible pour élaborer une structure fiable, la détection s'en trouve alors retardée, voire impossible.

— **Sensibilité au "silence "des utilisateurs** : dans certains scénarios (par exemple, durant le week-end ou à des heures de faible affluence), une information erronée peut rester immobile avant de prendre de l'ampleur, faisant en sorte que la période d'observation initiale ne reflète pas son potentiel de propagation.

— **Coût de la collecte et du calcul** : assurer en continu la récolte des retweets, l'actualisation des vecteurs de caractéristiques utilisateurs et l'inférence d'un RNN+CNN ou d'un GNN sur des réseaux extrêmement actifs (millions d'utilisateurs) nécessite une infrastructure de calcul décentralisée et un système de streaming à haute fréquence, ce qui pourrait restreindre son déploiement à grande échelle.

— **Biais associés à certains événements** : lors de crises ou d'événements d'actualité majeurs, la diffusion de contenus véridiques peut être aussi fulgurante et étendue qu'une fausse information. En l'absence d'un module supplémentaire d'analyse de contenu, le modèle risque de confondre un flux légitime très viral avec une rumeur malintentionnée [42, 43].

2.3.2.3 Approche s'appuyant sur le contexte événementiel

Objectif

Cette méthode cherche à examiner les informations erronées au niveau dún événement mondial plutôt que de se focaliser sur chaque communication individuelle. Le concept consiste à rassembler toutes les publications (tweets, posts, commentaires) liées à un même sujet ou événement marquant, tel quúne catastrophe naturelle, une attaque, une élection ou un événement sportif et déxtraire des caractéristiques globales. Ces critères portent généralement sur la quantité de publications, l´volution sémantique des échanges au cours du temps, la cohérence temporelle des messages partagés, et le profil des sources concernées. Le but est dídentifier des anomalies à grande échelle ou des schémas récurrents associés à la désinformation liée à un événement spécifique, plutôt que de juger la véracité dún message pris individuellement. Cette méthode sávère particulièrement utile pour repérer des opérations organisées de propagation de fausses informations ou pour examiner la structure narrative partagé dún événement suspect.

Techniques utilisées

Pour mettre en œuvre cette approche, plusieurs étapes sont nécessaires :

1. Clustering des messages par similarité lexicale et visuelle

Nous utilisons initialement des algorithmes de regroupement (comme le k-means, DBSCAN ou des modèles récents basés sur des embeddings) sur le contenu textuel des publications, généralement vectorisé par TF-IDF ou grâce à des modèles pré-entraînés (Word2Vec/GloVe/BERT). Cela permet une classification automatique des messages traitant du même sujet, même si les termes ou hashtags utilisés ne sont pas identiques. Quand il ságit dímages ou de vidéos associées aux publications, on fait appel à des détecteurs visuels (CNN, tel que ResNet ou VGG) pour obtenir des embeddings qui capturent la similarité sémantique visuelle. Une recherche dímages similaires peut également révéler lútilisation répétée dúne même image modifiée pour plusieurs messages suspects (par exemple, une photo de catastrophe retravaillée ou extraite de son contexte).

Dans Kotteti et al. (2020), pour chaque incident tel que “attaque à Paris ”, on regroupe tous les tweets liés à lhashtag, puis on compile des indicateurs textuels (taux dápparition déntités telles que “terrorisme ”ou “gouvernement ”), visuels (images similaires), temporels (cadence de publication) et sociaux (distribution des contributeurs en fonction de leur expérience) dans le but de créer un ensemble homogène relatif à un événement spécifique.

2. Extraction de topics par LDA et mesure de la cohérence sémantique

Après avoir effectué le regroupement, nous appliquons un modèle de Latent Dirichlet Allocation (LDA) afin dídentifier les sujets latents prédominants dans le flux d´vénements. Lhypothèse est quún événement authentique présente souvent une grande diversité lexicale (plusieurs sujets distincts, référence à différentes sources), alors quúne fausse nouvelle tend à se caractériser par un petit nombre de sujets dominants et une forte concentration sur des arguments répétitifs. Par exemple, Zhu et al. (2015) démontrent quún événement où moins de trois sujets latents concentrent plus de 70 % des messages pourrait signaler une manipulation organisée.

Nous complétons cette étude par le calcul dúne distance sémantique entre les messages (en utilisant Word2Vec ou TF-IDF) afin de mesurer la richesse du vocabulaire et le niveau de redondance : Un nombre restreint dúniques n-grammes et une diversité limitée déntités nommées (personnes, lieux) indiquent fréquemment une campagne de désinformation.

3. Comparaisons croisées entre sources

Par la suite, nous classons les messages en fonction de leur source (comptes vérifiés, médias réputés, comptes récents ou anonymes) et nous analysons la contribution de chaque catégorie au discours général. L´cart notable entre un contenu majoritairement issu de comptes récents et anonymes et celui provenant de comptes bien établis (médias traditionnels, experts reconnus) renforce la suspicion de fausses informations.

Lánalyse peut comporter la vérification des citations, en associant chaque citation à des sources externes de confiance (agences de presse, revues universitaires) afin d´valuer la correspondance des faits. Lábsence complète de citations de sources vérifiables ou, à línverse, la récurrence dádresses URL douteuses (diffusées par un groupe restreint de comptes) constitue un indicateur fort dálerte.

4. Détection d’anomalies temporelles

En analysant la ligne temporelle de publication, on détermine des mesures comme le rythme dáugmentation (slope) et la durée de vie du pic (tail heavy) du volume des messages. Un pic soudain suivi dúne dégradation rapide, soulignant quún événement a provoqué une soudaine explosion de messages presque identiques, puis a disparu, est typique dúne fausse information propagée par un groupe de comptes automatisés. Ma et al. (2016) démontrent que ces caractéristiques temporelles sont suffisantes pour distinguer les rumeurs naissantes des courants dínformation habituels.

Il est aussi possible dísoler des caractéristiques linguistiques temporelles (comme les variations dans lútilisation de mots interrogatifs ou déxpressions díncertitude), qui, combinées à des descripteurs statistiques, renforcent la détection des anomalies.

Un exemple récent est l’approche **Veracity-Oriented Context-Aware Prompting** fondée sur les grands modèles de langage (LLM). Cette méthode exploite des stratégies de *prompting optimisé* pour

évaluer la véracité d'une information en tenant compte du contexte événementiel associé (articles liés, discussions en ligne, métadonnées temporelles). Contrairement aux méthodes purement textuelles, cette approche ne se limite pas au contenu d'une phrase mais intègre des signaux contextuels plus larges, ce qui permet d'améliorer la précision et de réduire la vulnérabilité face aux manipulations subtiles.

Avantages

— **Contexte sémantique** : En regroupant tous les tweets associés à un même événement, nous obtenons une perspective globale qui empêche de se concentrer sur un seul tweet détaché. Cela aide à repérer des motifs récurrents (par exemple, références fréquentes à un même compte, hashtags modifiés) qui, pris individuellement, pourraient paraître inoffensifs, mais qui, en groupe, dévoilent une tentative concertée de désinformation.

— **Résilience face aux manipulations superficielles** : Si un diffuseur de fausse nouvelle tente de dissimuler son texte en utilisant un vocabulaire diversifié ou en modifiant sans cesse límage associée, lánalyse général de l´vénement révélera le manque de variété sémantique ou la rare présence de sources crédibles, signes dínformation suspecte.

— **Validation croisée** : Les résultats obtenus grâce à lápproche événementielle sont fréquemment utilisés comme un outil de vérification pour les alertes générées par dáutres techniques (analyse du profil utilisateur, diffusion de crédibilité). En combinant les signaux, on diminue les occurrences de faux positifs et on augmente la fiabilité.

Inconvénients

— **Dépendance à l'identification intégrale de l'événement** : Si le clustering initial néglige certains messages par exemple, en raison de variations de hashtags, d'erreurs d'orthographe ou de néologismes, l'évaluation globale peut être faussée ou partielle. Cette vulnérabilité du regroupement implique que la qualité des résultats est largement déterminée par le seuil de similitude ou l'algorithme de NER utilisé.

— **Insuffisant pour les fausses informations isolées** : Une fausse information qui ne circule que dans un cadre limité, comme par exemple un forum privé ou un groupe confidentiel ne produit pas une quantité d'informations similaire à celle d'un événement public. Dans cette situation, l'approche basée sur les événements n'a pas suffisamment de données consolidées pour identifier des tendances significatives. Par conséquent, elle n'est pas recommandée pour repérer de petites campagnes de désinformation très spécifiques.

— **Coûts de mise en œuvre et de flexibilité** : L'assemblage en temps réel des messages provenant de milliers d'événements potentiels, la réalisation de clustering lexical/visuels et l'application de LDA sur chaque ensemble nécessitent une infrastructure de calcul décentralisée. En l'absence de ressources appropriées (cluster Big Data, stockage en temps réel, puissance de calcul GPU), la solution risque d'être trop lente pour une exploitation en direct de grandes quantités de données sociales.

— **Bruit lié à des événements d'envergure** : Quand un événement légitime (attentat, élection, catastrophe naturelle) génère un flot considérable de messages, la diversité du vocabulaire peut naturellement varier (par exemple, plusieurs sources citent les mêmes mots-clés) et occulter les signaux de désinformation, ce qui pourrait mener à des faux positifs si l'on ne parvient pas à différencier le "véritable buzz "du "buzz trompeur " [44, 45].

2.3.2.4 Approches multimodales (fusion de modalités hétérogènes)

Objectif

Ces méthodes visent principalement à fusionner au moins deux des dimensions précédemment citées (texte, image, social, propagation, contexte) dans le but de diminuer la vulnérabilité des systèmes de détection face aux contournements. Par exemple, un contenu écrit qui paraît fiable peut être démasqué s'il est associé à une image de type "deepfake "ou si son mode de diffusion sur les réseaux sociaux présente des anomalies.

Techniques et Architectures

Les méthodes multimodales se basent sur des architectures sophistiquées en mesure de gérer et combiner des données de différentes natures :

1.Architectures multi-branches

Chaque type de modalité (texte, image, etc.) est géré par une branche séparée de l'architecture. Par exemple, une branche peut représenter un réseau de neurones convolutifs (CNN) dédié aux images, tandis qu'une autre pourrait être un réseau de neurones récurrents (RNN) ou un transformeur pour le traitement du texte.

2.Fusion des informations

Suite à la première étape de traitement pour chaque modalité, les informations obtenues sont combinées. Il est possible de réaliser cette fusion à divers niveaux, on a :

- **Fusion précoce (early fusion)** : Les caractéristiques élémentaires ou de bas niveau de chaque modalité sont assemblées avant d'être soumises à un modèle unique.
- **Fusion tardive (late fusion)** : Chaque modalité est analysée de manière autonome jusqu'à l'obtention de prédictions ou de scores, qui sont par la suite assemblés (par exemple, via un vote majoritaire ou une moyenne pondérée).
- **Fusion Intermédiaire** : Les représentations de niveau supérieur obtenues par chaque branche sont fusionnées à une étape intermédiaire de l'architecture, souvent via des couches de concaténation, des réseaux neuronaux spécifiquement conçus pour la fusion (tels que des réseaux neuronaux complètement connectés), ou des mécanismes d'attention qui apprennent à évaluer l'importance de chaque modalité.

3.Méthodes de fusion

Parmi les méthodes précises, on retrouve le stacking (où les résultats d'un modèle sont employés comme entrées pour un autre), le vote majoritaire (pour les missions de classification), ou encore l'emploi de réseaux de neurones sophistiqués qui apprennent à fusionner efficacement les informations provenant des diverses modalités.

Un exemple représentatif de cette approche est proposé par l'étude *User Comment-Guided Cross-Modal Attention for Interpretable Multimodal Fake News Detection*. Dans ce travail, les auteurs combinent le contenu textuel des articles, les images associées ainsi que les commentaires des utilisateurs grâce à un mécanisme d'attention croisée guidée par les interactions sociales. Cette architecture permet non seulement d'améliorer les performances de détection, mais aussi d'assurer une meilleure interprétabilité : le modèle peut mettre en évidence quels commentaires ou quelles parties du contenu multimodal influencent la décision finale.

Avantages

- **Robustesse accrue face aux tentatives de manipulation** : C'est l'avantage principal, si un attaquant essaie de rendre une modalité (comme le texte) crédible, les incohérences ou irrégularités dans les autres modalités (telles que les métadonnées sociales, la diffusion graphique ou l'image) peuvent trahir la tromperie. Par exemple, si le texte reproduit un style de rédaction journalistique, mais que les indications sociales signalent une provenance douteuse ou que le modèle de propagation ressemble à celui d'une opération de désinformation, le système multimodal sera plus susceptible d'identifier la fausseté.
- **Méthode plus complète et globale** : En tenant compte de divers aspects du contenu et de sa diffusion, ces systèmes proposent une perspective plus complète et subtile, ce qui entraîne une augmentation de la précision dans la détection.
- **Meilleure capacité à gérer l'ambiguïté** : Une information ambiguë dans une modalité peut être clarifiée par les informations provenant d'une autre modalité.

Inconvénients

— **Complexité de calcul** : Le traitement en parallèle de divers types de données et l'emploi d'architectures de modèles plus sophistiquées exigent des ressources informatiques considérables (processeur, carte graphique, mémoire), ce qui peut entraîner des coûts significatifs lors du développement et du déploiement de ces systèmes.

— **Exigence de disposer simultanément de plusieurs types de données** : Pour que l'approche multimodale soit efficace, il est essentiel d'avoir accès à tous les modes pertinents pour un contenu spécifique. Par exemple, pour étudier la propagation d'une image, il est nécessaire de suivre son trajet à travers les différentes plateformes, une tâche qui n'est pas toujours simple ou réalisable.

— **Problème d'intégration en temps réel** : L'intégration et la nécessité que toutes les modalités soient disponibles peuvent compliquer le placement de ces systèmes dans des flux de vérification en temps réel, où la rapidité de la prise de décision est primordiale. Le délai requis pour rassembler, traiter et combiner toutes les informations peut constituer un point de débordement.

— **Problèmes de synchronisation et d'alignement** : Veiller à ce que les données issues de diverses modalités soient convenablement alignées et synchronisées (soit sur le plan temporel, soit sur le plan sémantique) peut représenter un défi d'ordre technique.

— **Nécessité de jeux de données annotés multimodaux** : L'entraînement de ces modèles exige des ensembles de données sophistiqués, où toutes les modalités d'information sont accessibles et correctement annotées, ce qui s'avère généralement plus difficile à réaliser que des ensembles de données unimodaux [37].

2.3.3 Étude comparative des approches

Afin de comparer concrètement ces méthodes, on peut s'appuyer sur plusieurs critères :

Critère	Approches sociales (profil utilisateur)	Propagation (cheminement du signal)	Contexte événementiel	Approches multimodales
Signal exploité	Métadonnées utilisateur (ancienneté, activité, influence)	Structure et dynamique de diffusion (graphes, arbres de retweet)	Agrégation des messages liés à un événement (volume, diversité, cohérence)	Combinaison de plusieurs modalités : texte, image, social, propagation
Temps de détection	Très rapide (dès la publication)	Semi-rapide (besoin de données de diffusion initiale)	Moyen (après agrégation de messages)	Variable (selon la disponibilité des modalités)
Robustesse aux manipulations	Moyenne (contournable avec des comptes bien construits)	Bonne (analyse du comportement de diffusion)	Très bonne (analyse de patterns collectifs)	Excellente (corrèle plusieurs sources d'information hétérogènes)
Coût computationnel	Faible	Élevé (construction de graphes + séquences)	Moyen à élevé (clustering, topic modeling)	Élevé à très élevé (plusieurs réseaux de neurones + fusion)
Qualité requise des données	Profils utilisateurs accessibles	Historique de diffusion suffisant	Volume minimum d'informations pour former un cluster	Données complètes sur chaque modalité (texte, image, etc.)
Complexité d'implémentation	Faible à moyenne	Élevée (synchronisation temporelle + agrégée, LDA, etc.)	Moyenne (analyse agrégée, LDA, etc.)	Très élevée (fusion de branches, traitement parallèle)
Limites principales	Faux positifs pour comptes légitimes très actifs	Inefficace en cas de faible propagation ou d'événements dormants	Inefficace pour fake news isolées ou de petite ampleur	Difficulté d'implémentation en temps réel, dépendance à la complétude des données
Outils / Logiciels utilisés	Python (Pandas, Scikit-learn), NetworkX, Gephi, Neo4j	NetworkX, SNAP, PyTorch Geometric, TensorFlow/Keras pour RNN/CNN	Gensim (LDA), Scikit-learn (clustering), SpaCy, NLTK, Elasticsearch, Neo4j	TensorFlow, PyTorch, Hugging Face Transformers, OpenCV, DGL, multimodal fusion pipelines

— Les diverses méthodes de détection de fausses informations font appel à une gamme d'instruments selon les genres de signaux utilisés. L'approche sociale repose généralement sur des indicateurs simples collectés grâce à des APIs tels que Twitter API ou Facebook Graph, puis traités à l'aide de librairies comme Scikit-learn ou Pandas, avec parfois l'usage de Gephi pour la représentation graphique des réseaux. L'approche de propagation nécessite une modélisation explicite des graphes temporels, généralement élaborés à l'aide d'outils comme NetworkX, SNAP ou PyTorch Geometric, alors que l'analyse des séquences d'engagement se fait grâce à des réseaux LSTM ou CNN.

— En revanche, l'approche basée sur le contexte événementiel se concentre sur l'identification de regroupements thématiques en employant des méthodes de vectorisation telles que TF-IDF, Word2Vec ou BERT, associées à Gensim (pour l'analyse de sujets via LDA) et Scikit-learn pour le regroupement. Pour finir, les méthodes multimodales font appel à des frameworks complets de deep learning comme PyTorch, TensorFlow ou Keras. Elles intègrent également des modèles déjà formés provenant de Hugging Face, ainsi que des outils pour le traitement d'images tels qu'OpenCV. Parfois, elles utilisent aussi des bibliothèques de graphes comme DGL ou GraphGym afin d'exploiter des structures de diffusion complexes.

2.3.4 Défis et limites :

Repérer les fausses informations sur les réseaux sociaux est une problématique complexe qui génère plusieurs défis et contraintes pour les méthodes actuelles. Ces enjeux sont variés, englobant la nature du contenu, les mécanismes de diffusion, la disponibilité et la qualité des informations, sans oublier les contraintes inhérentes aux techniques actuelles.

1. Défis liés à la nature du contenu et à son interprétation

Un des défis majeurs est la subtilité et la complexité de l'analyse des informations actuelles. Les informations erronées sont généralement élaborées pour être trompeuses, en reproduisant le style et la structure des nouvelles authentiques. La compréhension de ceux-ci nécessite souvent une connaissance détaillée du contexte politique, social, culturel ou simplement du « bon sens » humain. Actuellement, les algorithmes de traitement du langage naturel (TLN) ont encore du mal à saisir ces nuances, ce qui complique l'identification entre une information authentique et une désinformation délibérément trompeuse. Le niveau croissant de sophistication des contenus produits par les entités malintentionnées rend cette mission d'autant plus complexe, même pour des spécialistes humains qualifiés.

Un autre problème de taille concerne la brièveté et la multimodalité du contenu sur les plateformes des réseaux sociaux. Les publications sur des réseaux sociaux tels que Twitter sont fréquemment très brèves, ce qui restreint le volume d'éléments linguistiques utilisables pour une analyse précise. En outre, un nombre grandissant de fausses informations ne se restreint pas à la simple textualité, mais englobe également des images, des vidéos et d'autres contenus multimédias. Ces éléments visuels peuvent être modifiés (comme par des modifications, des collages ou des vidéos falsifiées) et constituent un défi majeur pour les méthodes exclusivement basées sur l'examen du texte. Pour identifier ces modifications multimédias, des méthodes précises et sophistiquées, comme l'analyse forensique d'images ou la détection de deepfakes, sont indispensables, qui sont actifs et évoluent constamment.

Finalement, l'approche axée sur le contenu souffre d'une limitation significative due à sa dépendance linguistique. La majorité des caractéristiques linguistiques et des modèles de traitement du langage naturel sont spécifiques à une langue distincte. Cela implique que les modèles conçus pour une langue ne peuvent pas être transférés directement à une autre sans un travail de réajustement et de nouvel entraînement conséquent, ce qui restreint la capacité de généralisation et d'évolution des solutions de détection à l'échelle internationale.

2. Défis liés à la temporalité et à la dynamique de propagation

L'identification anticipée des fausses informations est un but essentiel mais complexe à réaliser. Lors de leur première diffusion, les fausses informations sont souvent dépourvues des indicateurs comportementaux et structurels requis pour leur détection. Les utilisateurs ont la possibilité de retweeter ou de diffuser des informations sans apporter de commentaires pertinents, et les tendances de diffusion n'ont pas encore eu assez de temps pour dévoiler des irrégularités. Les caractéristiques linguistiques et temporelles sont plutôt superficielles, ce qui entraîne une précision de détection limitée à cette étape préliminaire. Cependant, la vitesse à laquelle les fausses informations se propagent sur les plateformes sociales rend indispensable une identification précoce pour en réduire l'effet.

Manipuler les commentaires et les interactions initiaux constitue aussi un défi. Les individus malintentionnés peuvent essayer d'orienter les premiers commentaires et réactions pour créer une sensation de légitimité autour de leurs informations trompeuses. Cette opération peut complexifier la différenciation entre un engagement authentique et un engagement faux, altérant ainsi les signaux qui pourraient par ailleurs faciliter l'identification. De plus, l'évolution constante des tactiques utilisées par les créateurs de fake news pose un défi permanent. Les techniques de désinformation évoluent et se perfectionnent pour contrer les progrès des méthodes de détection. Ce jeu impose que les modèles de détection soient sans cesse actualisés et ajustés pour maintenir leur efficacité, ce qui demande des efforts permanents en matière de recherche et développement.

3. Défis liés aux données et à leur qualité

La collecte de données représente un défi crucial. Constituer des jeux de données qui soient à la fois vastes, complets et annotés pour former et évaluer des modèles de détection solides s'avère complexe. Le terme « fake news » est fréquemment confronté à interprétations et subtilités, sans qu'il y ait un accord général sur ce qui représente une « news » ni sur la façon de l'identifier comme étant « vraie » ou « fausse

». Cette incertitude complique l'étiquetage manuel des données et le rend susceptible d'erreurs humaines.

Le biais des sources est une autre préoccupation. Plusieurs recherches se basent sur le concept de sources sûres ou non pour évaluer la véracité des informations. Toutefois, cette méthode est souvent remise en question pour son caractère trop approximatif, car même des sources habituellement dignes de confiance peuvent ponctuellement partager des données trompeuses, tandis que des sources moins fiables peuvent parfois fournir des informations véridiques. Par ailleurs, la polarisation politique et idéologique peut affecter la manière dont on perçoit la fiabilité des sources, introduisant ainsi des biais dans les ensembles de données.

4. Limites inhérentes aux approches existantes

Les méthodes basées sur le contenu sont bloquées par leur incapacité à traiter les fausses informations sophistiquées qui sont conçues pour ne pas sembler immédiatement erronées. Elles peuvent aussi s'avérer inefficaces lorsqu'il s'agit de contenus concis ou exclusivement multimédias. Les méthodes basées sur les utilisateurs sources peuvent négliger des signaux distinctifs clés en omettant de considérer les caractéristiques des utilisateurs qui retweetent ou diffusent l'information, et pas simplement la source originale. Les caractéristiques « faites à la main » (handcrafted features) employées dans certaines méthodes graphes, comme la centralité ou les cliques, peuvent être aléatoires et ne pas nécessairement avoir une pertinence pour la tâche spécifique de détection des fausses informations. Elles peuvent ne pas avoir la capacité de saisir les modèles complexes et changeants de la désinformation.
Pour finir, la globalité des modèles représente une préoccupation de premier ordre. Les modèles formés sur des données propres à une époque ou à un contexte culturel spécifique peuvent ne pas être performants sur des données plus anciennes ou plus récentes, ou dans des contextes linguistiques ou culturels divergents. Cela met l'accent sur l'importance de concevoir des modèles plus adaptatifs et solides, capables de se modifier en fonction des changements dans le paysage de la désinformation.

Ces obstacles et contraintes mettent en évidence le besoin d'élaborer des méthodes plus solides, adaptatives et aptes à combiner des données variées pour une identification plus performante et généralisable des fausses informations sur les médias sociaux. Les recherches à venir devraient se focaliser sur l'amélioration de la compréhension du contexte, l'intégration multimodale, la détection préventive et la capacité de résistance face aux tactiques adverses [24, 26].

2.4 Conclusion

Il est devenu absolument essentiel de combattre les fausses informations qui circulent sur les réseaux sociaux. Leur diffusion rapide et leur aptitude à contourner les mécanismes de contrôle traditionnels mettent directement en péril la crédibilité de l'information, la cohésion sociale et la stabilité démocratique. Les moyens de détection actuels, bien qu'utiles dans certains contextes, présentent encore des lacunes notables concernant leur passage à grande échelle, leur fiabilité et leur rapidité d'exécution.

Dans ce chapitre, nous avons examiné les spécificités des réseaux sociaux, les caractéristiques propres aux fausses informations ainsi que leurs conséquences, et nous avons étudié les principales méthodes actuellement employées pour les repérer. L'évaluation de ces approches a révélé des limites structurelles, confirmant la nécessité d'une stratégie plus globale et diversifiée, capable de prendre en compte la complexité des contenus, les comportements des utilisateurs et la dynamique de propagation.

Face à ces constats, nous suggérons une approche hybride fondée sur l'analyse sémantique des contenus (NLP), la modélisation de la diffusion par les réseaux de neurones graphiques (GNN), la détection de l'intention à l'aide de modèles Transformer en mode zero-shot, ainsi que l'intégration multimodale (texte, image, audio, vidéo). Cette combinaison permet un repérage plus précis, contextuellement adapté et robuste face à l'évolution constante des tactiques de manipulation présentes sur les plateformes sociales..

Chapitre III :

Approches Hybrides pour la Détection des Fausses Informations dans les Réseaux Sociaux

3.1 Introduction

La prolifération des fausses informations, communément appelées « fake news », est devenue un défi majeur à l'ère numérique, menaçant la cohésion sociale, la stabilité politique et la confiance dans les institutions. Les réseaux sociaux, par leur nature même de diffusion rapide et virale de l'information, sont devenus des vecteurs privilégiés pour la propagation de ces contenus trompeurs. La détection automatique et efficace des fausses informations est donc une priorité de recherche cruciale. Ce chapitre se propose d'explorer en profondeur une approche de pointe pour relever ce défi : l'approche hybride combinant le Traitement du Langage Naturel (NLP) et les Réseaux de Neurones Graphiques (GNN).

3.1.1 L'Évolution du Paysage de la Désinformation en Ligne

: Historiquement, la désinformation a toujours existé, mais l'avènement d'Internet et des plateformes de médias sociaux a radicalement transformé son échelle, sa vitesse et sa complexité. Les fausses informations ne se limitent plus à de simples rumeurs ; elles sont souvent produites de manière sophistiquée, imitant le style et le format des nouvelles légitimes, et exploitant les biais cognitifs des utilisateurs pour maximiser leur diffusion [1]. Cette évolution a conduit à l'émergence de phénomènes tels que les « deepfakes » (vidéos ou audios manipulés de manière réaliste) et les campagnes d'influence coordonnées, rendant la détection manuelle par les fact-checkers de plus en plus difficile et insuffisante.

3.1.2 Limites des Approches Traditionnelles de Détection

Les premières tentatives de détection des fausses informations se sont principalement concentrées sur deux axes : l'analyse du contenu textuel et l'analyse des caractéristiques de propagation. Les approches basées sur le contenu utilisaient des techniques de NLP pour identifier des patterns linguistiques suspects, des incohérences sémantiques ou des tonalités émotionnelles extrêmes. Cependant, ces méthodes peinent à distinguer les fausses infrmationss bien écrites des vraies, et sont vulnérables aux techniques d'obfuscation. D'autre part, les approches basées sur la propagation analysaient la manière dont l'information se diffuse sur le réseau social (vitesse, nombre de partages, structure du réseau des diffuseurs) pour identifier des schémas anormaux. Bien que prometteuses, ces méthodes peuvent être trompeuses, car certaines vraies nouvelles peuvent également se propager rapidement, et les acteurs malveillants peuvent manipuler les schémas de diffusion pour échapper à la détection. De plus, ces approches ne prennent pas en compte la sémantique intrinsèque du contenu.

3.1.3 Justification de l'Approche Hybride NLP+GNN

Face aux limites des méthodes unimodales, l'approche hybride NLP+GNN se présente comme une solution robuste et complète. Elle capitalise sur les forces complémentaires du NLP et des GNN. Le NLP permet une analyse approfondie du contenu textuel, capturant la sémantique, le style et les éventuels signaux de désinformation intégrés dans le message lui-même. Simultanément, les GNN excellent dans la modélisation des relations complexes et des interdépendances au sein des réseaux sociaux, permettant d'analyser la structure de propagation, les interactions entre utilisateurs et la détection de communautés malveillantes. En fusionnant ces deux perspectives, l'approche hybride vise à fournir une capacité de détection plus précise, plus robuste et plus précoce, capable de s'adapter à la nature évolutive de la désinformation en ligne. Ce chapitre détaillera les fondements théoriques, l'architecture, les avantages, les défis et les perspectives de cette approche prometteuse.

3.2 Fondements Théoriques et Conceptuels

La détection des fausses informations est intrinsèquement un problème multidisciplinaire, nécessitant une compréhension approfondie des mécanismes linguistiques de la désinformation et des dynamiques de propagation au sein des réseaux sociaux. L'approche hybride NLP+GNN s'appuie sur des fondements théoriques solides issus de l'intelligence artificielle, de l'apprentissage automatique et de la théorie des graphes. Cette section détaillera les concepts clés du NLP et des GNN qui sont essentiels à la compréhension de cette approche.

3.2.1 Rappel sur le Traitement du Langage Naturel (NLP)

Le Traitement du Langage Naturel (NLP) est un domaine de l'intelligence artificielle qui vise à permettre aux ordinateurs de comprendre, d'interpréter et de générer le langage humain. Dans le contexte de la détection des fausses informations, le NLP est utilisé pour analyser le contenu textuel des messages et en extraire des caractéristiques pertinentes qui peuvent indiquer leur véracité ou leur fausseté.

3.2.1.1 Techniques d'Extraction de Caractéristiques Textuelles (TF-IDF, Word Embeddings)

Pour qu'un modèle d'apprentissage automatique puisse traiter le texte, celui-ci doit d'abord être converti en une représentation numérique. Historiquement, des techniques comme le *Bag-of-Words* (BoW) et le *Term Frequency-Inverse Document Frequency* (TF-IDF) ont été largement utilisées.

- **Bag-of-Words (BoW)** : Cette approche simple représente un document comme un ensemble de mots, ignorant l'ordre des mots et la structure grammaticale. La fréquence d'apparition de chaque mot dans le document est utilisée comme caractéristique. Bien que simple, BoW peut être efficace pour des tâches de classification de texte de base.
- **TF-IDF (Term Frequency-Inverse Document Frequency)** : Le TF-IDF est une amélioration du BoW qui pondère l'importance d'un mot non seulement par sa fréquence dans un document (TF), mais aussi par sa rareté dans l'ensemble du corpus (IDF). Un mot fréquent dans un document mais rare dans le corpus aura un score TF-IDF élevé, indiquant sa pertinence pour ce document.

Cependant, ces méthodes traditionnelles souffrent d'une limitation majeure : elles ne capturent pas les relations sémantiques entre les mots. Pour y remédier, les *Word Embeddings* ont été développés.

- **Word Embeddings (Word2Vec, GloVe, FastText)** : Les embeddings de mots sont des représentations vectorielles denses de mots qui capturent leur signification sémantique et leurs relations contextuelles. Les mots ayant des significations similaires sont situés à proximité dans l'espace vectoriel. Des modèles comme Word2Vec, GloVe et FastTextont révolutionné le NLP en permettant aux modèles de comprendre le sens des mots au-delà de leur simple présence.
  - **Word2Vec** : Développé par Google, Word2Vec est un groupe de modèles qui utilisent un réseau neuronal à deux couches pour apprendre les embeddings de mots. Il existe deux architectures principales : *Skip-gram* (qui prédit les mots contextuels à partir d'un mot cible) et *CBOW* (qui prédit un mot cible à partir de son contexte).
  - **GloVe (Global Vectors for Word Representation)** : GloVe est un modèle d'apprentissage non supervisé pour obtenir des représentations vectorielles de mots. Contrairement à Word2Vec qui se concentre sur les fenêtres de contexte locales, GloVe intègre des statistiques de co-occurrence globales du corpus pour construire ses embeddings.
  - **FastText** : Développé par Facebook AI, FastText est une extension de Word2Vec qui prend en compte les sous-mots (n-grammes de caractères). Cela permet à FastText de gérer efficacement les mots rares et les mots hors vocabulaire (OOV), et de capturer des informations morphologiques. [61, 62]

3.2.1.2 Modèles de Langage Avancés (Transformers : BERT, RoBERTa, GPT)

L'avènement des architectures basées sur les *Transformers* a marqué un tournant majeur dans le traitement automatique du langage naturel (NLP), permettant aux modèles de capturer des dépendances à longue portée dans le texte et de générer des représentations contextuelles dynamiques des mots.

- **Transformers** : Introduits par Vaswani et al, les Transformers sont des architectures de réseaux neuronaux qui s'appuient sur des mécanismes d'attention (*self-attention*) pour pondérer l'importance des différentes parties d'une séquence d'entrée lors de la génération d'une représentation. Cela leur permet de traiter les mots en parallèle et de capturer des relations complexes, contrairement aux RNN et LSTM qui traitent les séquences de manière séquentielle.

— **BERT (Bidirectional Encoder Representations from Transformers)** : Développé par Google, BERT est un modèle de langage pré-entraîné bidirectionnel qui a révolutionné de nombreuses tâches en NLP. Il est entraîné sur deux tâches principales :
  — le *Masked Language Model (MLM)*, où le modèle prédit des mots masqués dans une phrase ;
  — le *Next Sentence Prediction (NSP)*, où il détermine si deux phrases se suivent logiquement.

  La bidirectionnalité de BERT lui permet de comprendre le contexte d'un mot en fonction de tous les autres mots de la phrase, ce qui est crucial pour la détection des nuances dans les fausses informations.

— **RoBERTa (A Robustly Optimized BERT Pretraining Approach)** : RoBERTa est une version optimisée de BERT, développée par Facebook AI. Les améliorations incluent :
  — un entraînement sur un corpus de données plus large,
  — la suppression de la tâche NSP,
  — et l'utilisation de masquages dynamiques.

  Ces modifications ont permis à RoBERTa de surpasser BERT sur de nombreux benchmarks NLP.

— **GPT (Generative Pre-trained Transformer)** : Les modèles GPT (GPT-1, GPT-2, GPT-3, etc.), développés par OpenAI, sont des modèles de langage génératifs basés sur l'architecture Transformer. Contrairement à BERT, qui est un encodeur bidirectionnel, GPT est un décodeur unidirectionnel, principalement conçu pour la génération de texte. Bien que leur application directe dans la détection de fausses informations soit différente (par exemple, pour générer des exemples de fausses informations ou analyser la cohérence), leur capacité à comprendre et générer du langage naturel reste fondamentale pour ce domaine.

3.2.1.3 Analyse Stylistique et Sémantique pour la Détection de Fausses informations

L'analyse stylistique et sémantique est au cœur de l'utilisation du NLP pour la détection des fausses informations. Elle vise à identifier les caractéristiques intrinsèques du texte qui peuvent révéler sa nature trompeuse.

— **Analyse stylistique** : Les fausses informations présentent souvent des caractéristiques stylistiques distinctes. Cela peut inclure l'utilisation excessive de superlatifs, de langage émotionnel ou sensationnaliste, de fautes d'orthographe ou de grammaire (bien que les générateurs modernes soient de plus en plus sophistiqués), ou un style d'écriture qui diffère des sources d'information légitimes. Le NLP peut quantifier ces caractéristiques en analysant la richesse lexicale, la complexité syntaxique, la densité des mots-clés, et la présence de marqueurs de subjectivité ou de partialité.

— **Analyse sémantique** : L'analyse sémantique se concentre sur le sens du texte. Elle implique la détection d'incohérences factuelles, de contradictions logiques, ou de la manipulation du sens. Les *word embeddings* et les modèles Transformers sont essentiels ici, car ils permettent de capturer les relations sémantiques entre les mots et les phrases. Par exemple, un modèle peut identifier si une affirmation est en contradiction avec des faits établis ou si elle utilise un langage ambigu pour tromper le lecteur. Les techniques d'inférence textuelle (*Natural Language Inference* - NLI) peuvent être utilisées pour déterminer si une hypothèse est vraie, fausse ou indéterminée par rapport à un texte donné, ce qui est directement applicable à la vérification des faits.

— **Détection de *clickbait* et de titres trompeurs** : Les fausses informations utilisent souvent des titres accrocheurs (*clickbait*) pour attirer l'attention. Le NLP peut identifier ces titres en analysant leur structure, l'utilisation de points d'exclamation, de questions rhétoriques, ou de phrases qui créent un sentiment d'urgence ou de curiosité excessive.

— **Analyse de la subjectivité et de la polarité** : Le NLP peut évaluer le degré de subjectivité et la polarité (positive, négative, neutre) du langage utilisé. Les fausses informations tendent à être plus subjectives et à exprimer des opinions polarisées, contrairement aux reportages factuels qui visent l'objectivité [59, 60].

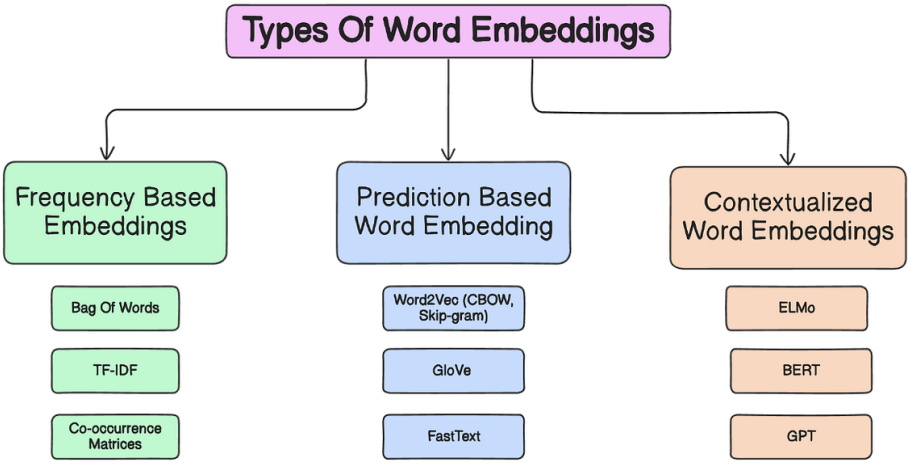

FIGURE 10 – Schéma comparatif des techniques d'embedding lexical[82]

En combinant ces différentes techniques d'analyse stylistique et sémantique, le module NLP de l'approche hybride peut générer des représentations riches et discriminantes du contenu textuel, qui serviront de base pour l'étape suivante de l'analyse graphique.

3.2.2 Introduction aux Réseaux de Neurones Graphiques (GNN)

Les Réseaux de Neurones Graphiques (GNN) sont une classe de réseaux de neurones profonds spécialement conçus pour opérer sur des données structurées sous forme de graphes. Contrairement aux données euclidiennes (images, texte séquentiel) où les CNN et RNN excellent, les données de graphes sont non-euclidiennes et complexes, avec des relations arbitraires entre les nœuds. Les GNN permettent d'exploiter la structure relationnelle des données, ce qui est particulièrement pertinent pour les réseaux sociaux où les informations se propagent à travers des connexions complexes entre utilisateurs et messages.

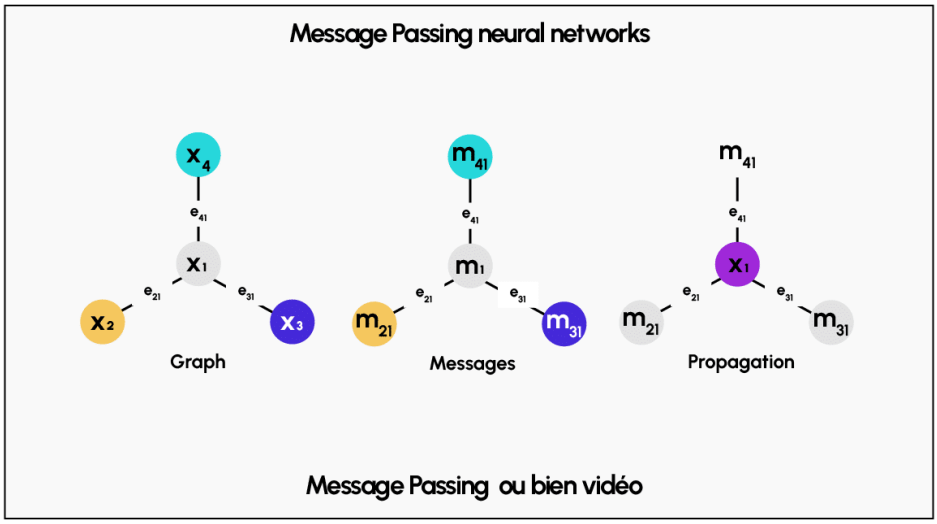

FIGURE 11 – Réseaux de Neurones Graphiques (GNN) [83]

3.2.2.1 Principes Fondamentaux des Graphes et des GNN

Un graphe est défini par un ensemble de nœuds (ou sommets) V et un ensemble d'arêtes (ou liens) E qui connectent ces nœuds. Dans le contexte des réseaux sociaux, les nœuds peuvent représenter des utilisateurs, des messages, des articles, des mots-clés, etc., et les arêtes peuvent représenter des relations d'amitié, des interactions (likes, partages, commentaires), des citations ou des similarités sémantiques. Chaque nœud peut également avoir des caractéristiques (features) associées, telles que l'embedding textuel d'un message ou les informations démographiques d'un utilisateur.

Le principe fondamental des *Graph Neural Networks* (GNN) est d'apprendre des représentations (embeddings) pour chaque nœud en agrégeant les informations de ses voisins dans le graphe. Ce processus d'agrégation est itératif : à chaque couche du GNN, la représentation d'un nœud est mise à jour en combinant sa propre représentation avec les représentations agrégées de ses voisins. Cela permet aux GNN de capturer à la fois les caractéristiques locales (des nœuds eux-mêmes) et les caractéristiques structurelles (des relations et du voisinage) du graphe.

Le processus de propagation de l'information dans un GNN peut être généralisé comme suit :

$$h_v^{(k)} = \text{AGGREGATE}^{(k)} \left(\left\{ h_u^{(k-1)} \mid u \in \mathcal{N}(v) \right\} \right), \tag{1}$$

où $\mathcal{N}(v)$ est l'ensemble des voisins du nœud v à la couche k.

$$h_v^{(k)} = \text{COMBINE}^{(k)} \left(h_v^{(k-1)}, h_v^{(k)} \right), \tag{2}$$

où $h_v^{(k)}$ est la représentation du nœud v à la k-ième couche.

Après plusieurs couches d'agrégation et de combinaison, chaque nœud possède une représentation riche qui encode des informations sur son propre contenu ainsi que sur sa position et ses relations dans le graphe. Ces représentations peuvent ensuite être utilisées pour des tâches en aval telles que la classification de nœuds (comme la détection de fausses informations), la classification de graphes ou la prédiction de liens.

3.2.2.2 Types de GNN Couramment Utilisés (GCN, GAT, GraphSAGE)

Plusieurs architectures de *Graph Neural Networks* (GNN) ont été proposées, chacune avec des mécanismes d'agrégation et de combinaison différents. Les plus courantes incluent :

— **Graph Convolutional Networks (GCN)** : les GCN sont l'une des architectures GNN les plus influentes. Ils généralisent l'opération de convolution des CNN aux graphes. L'idée est de propager les caractéristiques des nœuds voisins vers le nœud central, en les pondérant par la structure du graphe. La mise à jour des nœuds dans un GCN peut être formulée comme :

$$H^{(k+1)} = \sigma \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(k)} W^{(k)} \right), \tag{3}$$

où $\tilde{A} = A + I$ est la matrice d'adjacence avec boucles ajoutées, $\tilde{D}$ est la matrice de degré correspondante, $H^{(k)}$ est la matrice des caractéristiques des nœuds à la k-ième couche, $W^{(k)}$ est la matrice de poids entraînable, et σ est une fonction d'activation. Les GCN sont efficaces pour capturer les dépendances locales et globales dans le graphe [53].

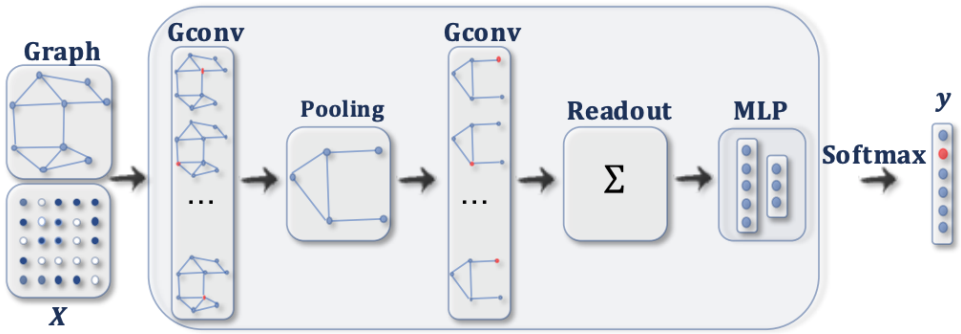

FIGURE 12 – Graph Convolutional Networks (GCN)[84]

— **Graph Attention Networks (GAT)** : les GAT introduisent un mécanisme d'attention dans le processus d'agrégation. Au lieu de pondérer uniformément les voisins comme dans les GCN, les GAT apprennent des poids d'attention pour chaque voisin, permettant au modèle de se concentrer sur les nœuds les plus pertinents pour une tâche donnée. Cela rend les GAT plus flexibles et robustes aux variations de la structure du graphe. Le mécanisme d'attention améliore aussi l'interprétabilité, car les poids révèlent l'importance relative des différentes connexions.

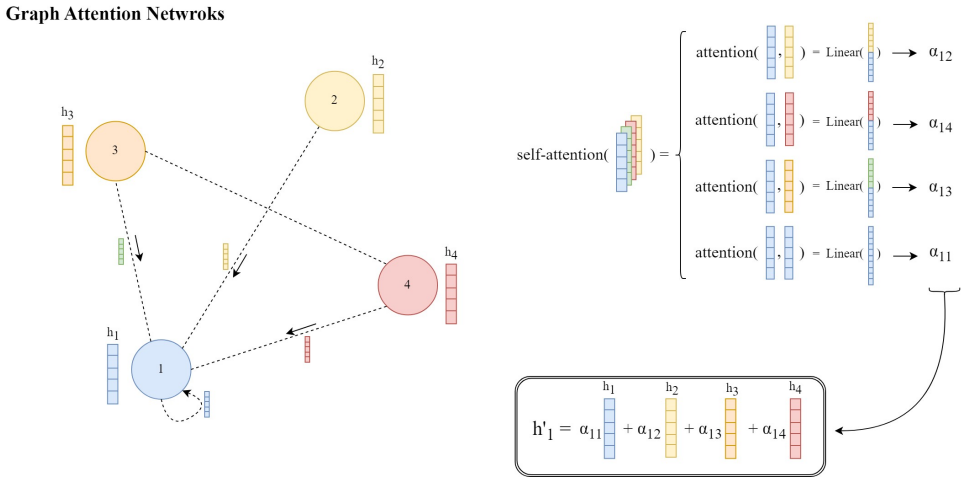

FIGURE 13 – Graph Attention Networks (GAT) [85]

— **GraphSAGE (SAmple and aggreGatE)** : GraphSAGE est un cadre d'apprentissage inductif permettant de générer des embeddings de nœuds. Contrairement aux GCN qui apprennent des transformations pour un graphe fixe, GraphSAGE apprend une fonction d'agrégation applicable à de nouveaux nœuds ou graphes non vus pendant l'entraînement. Cela le rend particulièrement adapté aux grands graphes dynamiques comme les réseaux sociaux. GraphSAGE permet l'utilisation de différentes fonctions d'agrégation (moyenne, LSTM, pooling) pour combiner les informations des voisins [54].

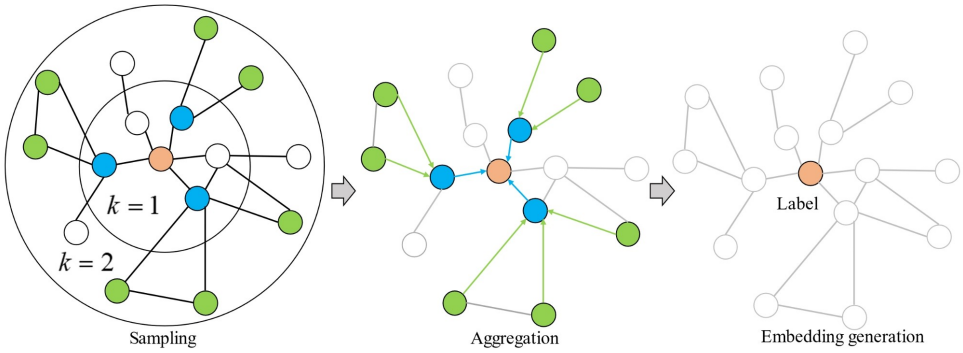

FIGURE 14 – GraphSAGE (SAmple and aggreGatE) [86]

3.2.2.3 Modélisation des Réseaux Sociaux comme Graphes pour la Détection de Fausses informations

La modélisation des réseaux sociaux sous forme de graphes est une approche naturelle pour la détection des fausses informations, car elle permet de capturer les interactions complexes et les dynamiques de propagation. Dans ce contexte, différents types de graphes peuvent être construits :

- **Graphes de propagation** : Ces graphes représentent la manière dont une information se diffuse. Les nœuds peuvent être les messages (posts, tweets), et les arêtes représentent les relations de propagation (retweets, partages, citations). L'analyse de ces graphes peut révéler des schémas de diffusion anormaux, tels que des propagations rapides et explosives souvent associées aux fausses iformations, ou des structures de diffusion par des bots.
- **Graphes d'interaction utilisateur** : Ces graphes se concentrent sur les relations entre les utilisateurs. Les nœuds représentent les utilisateurs, et les arêtes les interactions (mentions, réponses, abonnements mutuels, etc.). L'analyse de ces graphes peut aider à identifier des communautés d'utilisateurs qui partagent et amplifient les fausses informations, ou des comptes suspects (bots, trolls) qui sont au centre de ces réseaux de désinformation.
- **Graphes hétérogènes** : Les réseaux sociaux sont intrinsèquement hétérogènes, contenant différents types d'entités (utilisateurs, messages, hashtags, images) et de relations. Les graphes hétérogènes permettent de modéliser cette complexité en intégrant des nœuds et des arêtes de différents types. Par exemple, un graphe hétérogène peut inclure des nœuds pour les utilisateurs et les messages, avec des arêtes reliant un utilisateur à ses messages, ou des messages citant d'autres messages. Les GNN hétérogènes sont particulièrement adaptés pour traiter ces structures complexes et intégrer des informations multimodales.

L'utilisation des GNN dans ce contexte permet de capturer non seulement les caractéristiques intrinsèques des nœuds (par exemple, le contenu textuel d'un message), mais aussi les informations contextuelles dérivées de la structure du graphe. Par exemple, un message publié par un utilisateur ayant de nombreuses connexions avec d'autres diffuseurs de fausses informations pourrait être considéré comme plus suspect, même si son contenu textuel n'est pas explicitement trompeur. Cette capacité à intégrer le contenu et le contexte relationnel est la pierre angulaire de l'approche hybride NLP+GNN.

3.3 Architecture du Modèle Hybride NLP+GNN

L'architecture d'un modèle hybride NLP+GNN pour la détection des fausses informations est conçue pour intégrer de manière synergique les informations sémantiques extraites du contenu textuel et les

informations structurelles dérivées du graphe de propagation social. Bien que des variations existent en fonction des spécificités de la tâche et des données, une architecture générique peut être décomposée en plusieurs modules interconnectés, chacun jouant un rôle crucial dans le processus de détection.

3.3.1 Module d'Encodage du Contenu Textuel (NLP)

Ce module est responsable de la transformation du contenu textuel brut des messages (articles, posts, tweets, commentaires) en représentations numériques riches et informatives. Ces représentations serviront de caractéristiques initiales pour les nœuds du graphe.

3.3.1.1 Prétraitement du Texte et Normalisation

Avant l'encodage, le texte brut doit subir une série d'étapes de prétraitement pour le nettoyer et le normaliser, réduisant ainsi le bruit et améliorant la qualité des représentations. Ces étapes incluent généralement :

— **Tokenisation** : Division du texte en unités plus petites (mots, sous-mots, caractères) appelées *tokens*. Par exemple, la phrase « La détection des fausses informations est cruciale. » pourrait être tokenisée en `[« La », « détection », « des », « fausses », « nouvelles », « est », « cruciale », « . »]`.

— **Nettoyage** : Suppression des caractères spéciaux, des chiffres, de la ponctuation (sauf si pertinente), des balises HTML, des URLs, et des mentions d'utilisateurs ou de hashtags (si non pertinents pour l'analyse sémantique). La conversion en minuscules est également une étape courante pour réduire la variabilité.

— **Suppression des mots vides (Stop Words)** : Élimination des mots très fréquents mais peu informatifs (ex : « le », « la », « et », « est ») qui n'apportent pas beaucoup de sens sémantique.

— **Lemmatisation ou Stemming** : Réduction des mots à leur forme de base (lemme ou racine). Par exemple, « courir », « courait », « couru » seraient réduits à « courir » (lemmatisation) ou « cour » (stemming). Cela permet de regrouper les différentes formes flexionnelles d'un mot. [61]

3.3.1.2 Génération d'Embeddings Contextuels (Ex : BERT-based Embeddings)

Après le prétraitement, le texte est transformé en représentations vectorielles. Les *embeddings* contextuels, notamment ceux générés par des modèles basés sur les *Transformers*, sont privilégiés pour leur capacité à capturer les nuances sémantiques et le contexte des mots.

— **Utilisation de modèles Transformers pré-entraînés** : Des modèles comme *BERT*, *RoBERTa*, ou *XLM-RoBERTa* (pour le multilingue) sont utilisés pour générer des *embeddings* pour chaque token ou pour l'ensemble du message. Ces modèles sont pré-entraînés sur de vastes corpus de texte et ont appris des représentations linguistiques riches. Pour un message donné, le modèle Transformer produit un vecteur de haute dimension qui encode son contenu sémantique et syntaxique.

— **Fine-tuning (réglage fin)** : Dans certains cas, le modèle Transformer peut être *fine-tuné* sur un corpus spécifique à la détection des fausses informations. Cela permet au modèle d'adapter ses représentations aux particularités linguistiques et stylistiques des fausses informations dans le domaine cible.

— **Agrégation des embeddings** : Si le modèle génère des embeddings pour chaque token, une stratégie d'agrégation (par exemple, moyenne, *max-pooling*, ou utilisation du vecteur `[CLS]` pour BERT) est appliquée pour obtenir une représentation unique de l'ensemble du message. Ce vecteur agrégé devient la caractéristique initiale du nœud de message dans le graphe. [60]

3.3.1.3 Intégration des Caractéristiques Stylistiques et Linguistiques

En plus des *embeddings* sémantiques, des caractéristiques stylistiques et linguistiques peuvent être extraites et intégrées pour enrichir la représentation du contenu. Ces caractéristiques peuvent inclure :

— **Mesures de lisibilité** : Indices de lisibilité (par exemple, indice de Flesch-Kincaid) qui évaluent la complexité du texte.

— **Caractéristiques lexicales** : Richesse du vocabulaire (nombre de mots uniques), longueur moyenne des mots, fréquence des mots rares.

— **Caractéristiques syntaxiques** : Complexité des phrases (nombre de propositions, profondeur de l'arbre syntaxique), utilisation de la voix passive ou active.

— **Caractéristiques émotionnelles / sentimentales** : Score de sentiment (positif, négatif, neutre), détection d'émotions spécifiques (colère, peur, surprise) pour identifier un langage sensationnaliste ou manipulateur.

— **Caractéristiques de crédibilité** : Présence de sources citées, utilisation de pronoms personnels (ce qui peut indiquer une subjectivité), utilisation de mots d'incertitude ou d'hyperbole.

Ces caractéristiques peuvent être concaténées avec les *embeddings* contextuels pour former un vecteur de caractéristiques de contenu complet pour chaque message. Ce vecteur sera ensuite utilisé comme attribut initial pour les nœuds de message dans le graphe. [4]

3.3.2 Module de Construction et de Représentation du Graph

Ce module est fondamental pour l'approche hybride, car il transforme les données brutes des réseaux sociaux en une structure de graphe exploitable par les GNN. La qualité et la pertinence de la construction du graphe influencent directement la capacité du modèle à capturer les dynamiques de propagation et les relations sociales.

3.3.2.1 Définition des Nœuds (Messages, Utilisateurs, Entités)

Le choix des entités à représenter comme nœuds dans le graphe est crucial et dépend des aspects de la désinformation que l'on souhaite modéliser.

Types de nœuds dans le graphe de réseau social

— **Nœuds de message (Post, Tweet, Article)** : Chaque publication individuelle sur le réseau social est représentée comme un nœud. Ces nœuds portent les caractéristiques de contenu textuel générées par le module NLP (voir Section 3.3.1). C'est le type de nœud le plus direct pour la classification des fausses informations.

— **Nœuds d'utilisateur** : Chaque utilisateur participant à la diffusion ou à l'interaction avec les messages est représenté comme un nœud. Les caractéristiques associées à ces nœuds peuvent inclure des informations de profil (nombre d'abonnés, ancienneté du compte), des métriques d'activité (fréquence de publication, d'interaction), ou des caractéristiques comportementales (modèles de retweet, de mention). L'intégration des utilisateurs permet de modéliser la crédibilité de la source et les schémas de comportement des diffuseurs.

— **Nœuds d'entité** : Pour une analyse plus fine, des entités nommées (personnes, organisations, lieux) ou des hashtags mentionnés dans les messages peuvent également être représentés comme des nœuds. Cela permet de construire des graphes sémantiques où les relations entre les messages peuvent être inférées par les entités qu'ils partagent. Par exemple, deux messages mentionnant la même entité peuvent être liés, même s'ils ne sont pas directement connectés par une relation de propagation.

— **Nœuds hétérogènes** : Dans des architectures plus complexes, différents types de nœuds peuvent coexister dans le même graphe (par exemple, messages et utilisateurs). Ces graphes hétérogènes nécessitent des GNN spécifiques (comme les HGNN) capables de gérer la diversité des types de nœuds et d'arêtes [19].

3.3.2.2 Définition des Arêtes (Relations de Propagation, Interactions, Similarité)

Les arêtes du graphe représentent les relations entre les nœuds. Le choix des types d'arêtes est tout aussi important que le choix des nœuds, car elles encodent les dynamiques de propagation et les interactions sociales.

Types de relations dans le graphe

- **Relations de propagation** : Ce sont les liens les plus directs pour modéliser la diffusion de l'information. Elles incluent :
  - **Retweets / Partages** : Une arête est créée entre un message original et un retweet/partage, ou entre l'utilisateur qui a publié le message original et l'utilisateur qui l'a partagé. Ces arêtes sont souvent dirigées, reflétant le flux de l'information.
  - **Citations / Réponses** : Des arêtes peuvent relier un message à un autre message auquel il répond ou qu'il cite. Cela permet de suivre les conversations et les chaînes de discussion.
- **Relations d'interaction** : Ces arêtes représentent les engagements des utilisateurs avec le contenu ou entre eux :
  - **Mentions** : Une arête est ajoutée entre un utilisateur et un message s'il a été mentionné, ou entre deux utilisateurs s'ils se sont mutuellement mentionnés.
  - **Likes / Réactions** : Des arêtes peuvent indiquer qu'un utilisateur a aimé ou réagi à un message. Bien que moins informatives sur la propagation directe, elles reflètent l'engagement et l'approbation.
- **Relations de similarité** : Ces arêtes sont construites sur la base de la similarité entre les nœuds :
  - **Similarité de contenu** : Des arêtes peuvent relier des messages ayant un contenu sémantique similaire (calculé par exemple à partir de la similarité cosinus de leurs embeddings NLP), même s'ils ne sont pas directement liés par une action de propagation. Cela peut aider à identifier des groupes de messages traitant du même sujet ou des variantes d'une même fausse nouvelle.
  - **Similarité d'utilisateur** : Des arêtes peuvent relier des utilisateurs ayant des comportements similaires (par exemple, partageant les mêmes sources, interagissant avec les mêmes types de contenu) ou des profils similaires. Cela peut révéler des groupes d'utilisateurs coordonnés ou des communautés de désinformation [28, 22].

3.3.2.3 Représentation des Caractéristiques des Nœuds et des Arêtes

Une fois les nœuds et les arêtes définis, leurs caractéristiques doivent être représentées numériquement pour être traitées par le GNN. Ces caractéristiques peuvent être de différentes natures :

- **Caractéristiques des nœuds** :
  - Pour les **nœuds de message**, il s'agit des *embeddings* de contenu textuel générés par le module NLP (voir Section 3.3.1.2), potentiellement enrichis par des caractéristiques stylistiques (voir Section 3.3.1.3).
  - Pour les **nœuds d'utilisateur**, les caractéristiques peuvent inclure :
    - des métadonnées de profil (ancienneté du compte, nombre de followers/followings),
    - des métriques d'activité (nombre de publications, de retweets),
    - des caractéristiques comportementales (ratio retweets/posts, diversité des sources).
  - Ces caractéristiques sont généralement représentées sous forme de vecteurs numériques.
- **Caractéristiques des arêtes** :
  - Les arêtes peuvent également être porteuses de caractéristiques associées.
  - Pour les **arêtes de propagation**, il peut s'agir :
    - du temps écoulé depuis la publication du message original,
    - du nombre de partages du message,
    - de la force de la relation (par exemple, relation d'amitié ou d'abonnement simple).
  - Pour les **arêtes de similarité**, la caractéristique principale peut être le *score de similarité* entre les nœuds.
  - Ces caractéristiques peuvent être exploitées par le GNN pour pondérer l'importance des informations reçues depuis les nœuds voisins [54, 53].

La combinaison de ces informations de contenu et de structure dans une représentation de graphe riche est la première étape vers une détection plus efficace des fausses informations. Le module suivant, le GNN, exploitera cette structure pour apprendre des représentations contextuelles pour chaque nœud.

3.4 Module de Traitement et d'Apprentissage Graphique (GNN)

Le cœur de l'approche hybride réside dans le module GNN, qui prend en entrée le graphe construit avec les caractéristiques de nœuds (incluant les embeddings NLP) et apprend des représentations enrichies pour chaque nœud en exploitant la structure du graphe. Ce processus permet au modèle de comprendre comment le contenu est influencé par sa propagation et par les entités avec lesquelles il interagit.

3.4.1.1 Agrégation d'Informations Locales et Globales sur le Graphe

Le processus d'apprentissage dans un GNN est itératif et repose sur l'agrégation d'informations provenant des voisins d'un nœud. À chaque couche du GNN, la représentation (embedding) d'un nœud est mise à jour en combinant sa propre information avec les informations agrégées de ses voisins. Cette agrégation peut se faire de différentes manières, selon l'architecture GNN utilisée (GCN, GAT, GraphSAGE, etc.).

- **Agrégation locale** : Les GNN commencent par agréger les informations des voisins directs d'un nœud. Par exemple, pour un nœud de message, cela pourrait inclure les utilisateurs qui l'ont partagé, les messages auxquels il répond, ou les entités qu'il mentionne. Cette agrégation locale permet de capturer le contexte immédiat du nœud dans le graphe.
- **Propagation multi-couches** : En empilant plusieurs couches GNN, le modèle peut agréger des informations provenant de voisins de plus en plus éloignés dans le graphe. Une GNN à k couches peut capturer des informations sur le voisinage à k sauts. Cela permet au modèle de comprendre les schémas de propagation à plus grande échelle et les influences indirectes. Par exemple, un message peut être influencé non seulement par l'utilisateur qui l'a partagé, mais aussi par les amis de cet utilisateur, et ainsi de suite.
- **Pondération des voisins** : Certaines architectures GNN, comme les GAT, utilisent des mécanismes d'attention pour pondérer l'importance des informations provenant de différents voisins. Cela signifie que le modèle peut accorder plus d'importance aux voisins jugés plus pertinents pour la tâche de détection des fausses informations (par exemple, les utilisateurs particulièrement influents ou fiables).

3.4.1.2 Fusion des Embeddings NLP et GNN

La fusion des embeddings NLP (qui représentent le contenu) et des informations apprises par le GNN (qui représentent la structure et le contexte) est une étape cruciale pour l'approche hybride. Il existe plusieurs stratégies pour réaliser cette fusion :

— **Concaténation** : La méthode la plus simple consiste à concaténer les embeddings NLP initiaux des nœuds avec les embeddings appris par le GNN. Cela crée un vecteur de représentation final qui combine explicitement les informations de contenu et de structure. Par exemple, si un message a un embedding NLP de dimension d_{NLP} et un embedding GNN de dimension d_{GNN}, l'embedding fusionné aura une dimension $d_{\text{NLP}} + d_{\text{GNN}}$.

— **Fusion Précoce (Early Fusion)** : Dans cette approche, les embeddings NLP sont utilisés comme caractéristiques initiales des nœuds du graphe, et le GNN apprend directement sur ces caractéristiques. La fusion se produit donc dès le début du processus d'apprentissage du GNN, permettant au modèle de considérer le contenu et la structure de manière intégrée tout au long de la propagation de l'information.

— **Fusion Tardive (Late Fusion)** : Ici, les embeddings NLP et les embeddings GNN sont appris séparément, puis combinés à un stade ultérieur, souvent juste avant la couche de classification. Cela peut impliquer des couches de fusion spécifiques (par exemple, des réseaux neuronaux fully connected) qui apprennent à combiner les deux types d'informations de manière optimale.

— **Mécanismes d'Attention Croisée** : Des mécanismes d'attention peuvent être utilisés pour permettre au modèle de pondérer l'importance relative des informations de contenu et de structure. Par exemple, un mécanisme d'attention pourrait décider si, pour un nœud donné, l'information de contenu est plus pertinente que l'information structurelle, ou vice-versa, pour la tâche de détection des fausses informations.

3.4.1.3 Apprentissage des Représentations Latentes du Graph

L'objectif final du module GNN est d'apprendre des représentations latentes (embeddings) pour chaque nœud du graphe. Ces embeddings sont des vecteurs de haute dimension qui encodent de manière compacte toutes les informations pertinentes sur le nœud, y compris son contenu sémantique, sa position dans le graphe, et ses relations avec les autres nœuds. Ces représentations latentes sont optimisées pour la tâche de détection des fausses informations.

- **Optimisation par Descente de Gradient** : Comme pour les autres réseaux de neurones, les paramètres du GNN (poids des couches d'agrégation, matrices de transformation, etc.) sont appris via la descente de gradient, en minimisant une fonction de perte (par exemple, l'entropie croisée pour la classification) qui mesure l'écart entre les prédictions du modèle et les étiquettes de vérité terrain (vrai/faux).
- **Apprentissage End-to-End** : Dans de nombreux modèles hybrides, le module NLP et le module GNN sont entraînés de manière *end-to-end*. Cela signifie que les gradients sont propagés à travers les deux modules, permettant aux représentations NLP d'être optimisées spécifiquement pour la tâche de détection des fausses informations dans le contexte du graphe, et vice-versa. Cette co-optimisation est un avantage clé de l'approche hybride.

Les représentations latentes ainsi apprises par le GNN sont ensuite transmises au module de classification, où la décision finale concernant la véracité de l'information sera prise

3.4.2 Module de Classification et de Décision

Le module de classification est la dernière étape du pipeline de détection. Il prend en entrée les représentations latentes enrichies générées par le module GNN et les utilise pour prédire la probabilité qu'un message soit une fausse information ou non. Ce module est généralement composé de couches de réseaux de neurones traditionnelles

3.4.2.1 Couches de Classification (Fully Connected, Softmax)

- **Couches Fully Connected (Dense)** : Les embeddings finaux des nœuds (messages) sont passés à une ou plusieurs couches fully connected (ou denses). Chaque neurone de ces couches est connecté à tous les neurones de la couche précédente.

  Ces couches apprennent des transformations non linéaires sur les représentations latentes, permettant au modèle de capturer des motifs complexes et de prendre des décisions de classification. Des fonctions d'activation non linéaires (comme ReLU) sont généralement appliquées après chaque couche fully connected pour introduire de la non-linéarité.
- **Couche de Sortie Softmax (pour la classification binaire ou multi-classes)** : Pour la tâche de classification finale, une couche de sortie est utilisée. Pour la détection de fausses informations, qui est souvent un problème de classification binaire (vrai/faux), une couche avec un seul neurone et une fonction d'activation sigmoïde peut être utilisée pour prédire la probabilité d'être une fausse nouvelle.

  Pour une classification multi-classes (par exemple, vrai, faux, satire, opinion), une couche avec autant de neurones que de classes et une fonction d'activation softmax est employée. La fonction softmax convertit les sorties brutes du réseau en probabilités, où la somme des probabilités pour toutes les classes est égale à 1.
- **Fonction Sigmoïde (pour classification binaire)** : $\sigma(z) = \frac{1}{1+e^{-z}}$. La sortie est une probabilité entre 0 et 1.
- **Fonction Softmax (pour classification multi-classes)** :

$$P(y = j \mid x) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$$

  où z_j est la sortie du neurone j et K est le nombre de classes.

3.4.2.2 Stratégies de Classification (Binaire, Multi-classes)

La détection des fausses informations peut être formulée comme un problème de classification binaire ou multi-classes, en fonction de la granularité souhaitée pour les étiquettes de vérité terrain.

- **Classification Binaire** : C'est l'approche la plus courante, où un message est classé comme « vrai » ou « faux ». Cela simplifie le problème mais peut masquer des nuances, comme la satire ou l'opinion.
- **Classification Multi-classes** : Certains jeux de données et certaines applications nécessitent une classification plus fine, distinguant par exemple les vraies nouvelles, les fausses informations, la satire, la désinformation, la mésinformation, etc. Cela fournit une compréhension plus nuancée de la nature de l'information.

Le module de classification prend la décision finale en se basant sur les représentations contextuelles et structurelles apprises par le GNN, qui ont été enrichies par les informations sémantiques du NLP. Cette intégration permet au modèle de prendre des décisions éclairées, en considérant à la fois ce qui est dit et comment cela est diffusé.

3.5 Concept de la détection d'intention

La détection de l'intention consiste à identifier la finalité communicative d'un message, c'est-à-dire ce que l'émetteur cherche réellement à accomplir au-delà du simple contenu textuel. Dans le cadre de la désinformation, il ne s'agit pas seulement de vérifier la véracité d'un énoncé, mais de comprendre l'objectif sous-jacent, comme manipuler l'opinion publique, renforcer des croyances erronées, mobiliser un groupe social ou simplement divertir par la satire.

Cette tâche est particulièrement complexe sur les réseaux sociaux car un même contenu peut avoir des intentions plurielles ou ambiguës. Par exemple :

— « La Terre est plate » → intention **désinformation**.

— « Certains pensent encore que la Terre est plate (absurde, non ?) » → intention **divertissement**.

La détection de l'intention complète donc la classification de vérité/fausseté en fournissant une dimension pragmatique essentielle à l'analyse [46].

3.5.1 Zero-Shot Learning appliqué à la classification d'intentions

Le *Zero-Shot Learning* (ZSL) est une approche d'apprentissage qui permet à un modèle de classer un texte dans des catégories jamais vues lors de la phase d'entraînement. Cela représente un changement de paradigme par rapport aux méthodes classiques de *supervised learning*, qui nécessitent de disposer d'un dataset annoté exhaustif pour chaque nouvelle tâche ou intention cible.

Principe général

Le ZSL exploite les capacités des modèles de langage pré-entraînés (PLMs), tels que BERT, RoBERTa, GPT, ou encore T5, qui ont appris des représentations linguistiques universelles sur d'immenses corpus. Ces représentations permettent de projeter simultanément :

— le texte du message,

— les labels d'intention (formulés en langage naturel),

dans un même espace vectoriel latent.

La classification repose alors sur un calcul de similarité entre l'encodage du texte et celui des labels. Plus le message est proche d'un label dans l'espace sémantique, plus la probabilité qu'il lui corresponde est élevée.

Mécanisme technique

Le processus Zero-Shot pour la détection d'intention suit généralement trois étapes :

1. **Encodage du texte** : le message est transformé en un vecteur dense via un Transformer (ex. embedding [CLS] de BERT).
2. **Encodage des intentions** : chaque intention est reformulée en phrase naturelle ("Ce texte vise à désinformer", "Ce texte vise à informer") et encodée de la même manière.
3. **Mesure de compatibilité** : le modèle calcule une similarité (souvent cosinus ou entailment probability) entre le vecteur du texte et ceux des intentions, attribuant le label le plus proche.

Exemple pratique

— Texte : « Le vaccin modifie l'ADN humain ».
— Labels candidats : { « désinformer », « informer », « divertir » }.
— Encodage → projection dans l'espace latent → calcul de similarité.
— Résultat : le texte est classé comme **« désinformer »**, avec une forte probabilité.

Stratégies avancées en Zero-Shot

Deux grandes variantes existent pour la classification Zero-Shot :

— **Approche basée sur l'entailment (NLI)** : formulation de la tâche comme une inférence logique. Exemple : "Ce texte désinforme." → le modèle de NLI évalue si le texte implique ou contredit cette hypothèse. Des modèles comme BART-NLI ou RoBERTa-NLI sont utilisés dans ce cadre.
— **Approche basée sur l'encodage sémantique** : représentation conjointe du texte et des labels dans le même espace vectoriel, suivie d'un calcul de similarité. Les modèles de type *Sentence-BERT* (SBERT) sont souvent employés ici.

Atouts de l'approche Zero-Shot pour l'intention

— **Pas besoin de données annotées spécifiques** : déploiement rapide sur de nouvelles plateformes ou langues.
— **Évolutivité** : de nouvelles intentions peuvent être intégrées simplement en ajoutant des labels formulés en langage naturel.
— **Adaptabilité multilingue** : les modèles comme XLM-R permettent d'appliquer la méthode dans plusieurs langues sans entraînement supplémentaire.
— **Utilité en détection de manipulation** : le ZSL peut identifier des intentions implicites (influencer, manipuler, ironiser), même sans dataset spécialisé [72].

3.5.2 Étapes du processus de détection

La détection d'intention Zero-Shot via Transformers repose sur un pipeline rigoureux, structuré en plusieurs phases successives. L'objectif est de mapper simultanément le message et les intentions candidates dans un espace vectoriel latent, puis de mesurer leur proximité sémantique pour décider de l'étiquette la plus probable.

3.5.2.1 Prétraitement linguistique Avant tout encodage, le texte brut subit une **normalisation linguistique** afin de réduire le bruit et standardiser les séquences :

— **Minuscule** : transformation de tous les caractères en minuscules pour uniformiser le vocabulaire.
— **Nettoyage** : suppression de la ponctuation non informative, des caractères spéciaux, ainsi que des artefacts sociaux (URLs, hashtags, emojis).
— **Tokenisation** : application d'une segmentation adaptée au modèle choisi, par exemple *WordPiece* pour BERT ou *Byte-Pair Encoding (BPE)* pour RoBERTa.

Formellement, une séquence textuelle brute $S = \{w_1, w_2, \ldots, w_n\}$ est transformée en une séquence de tokens $T = \{t_1, t_2, \ldots, t_m\}$, où $m \geq n$ dépend du vocabulaire du modèle.

3.5.2.2 Encodage du message et des intentions L'encodage repose sur les couches **Transformer Encoder** qui génèrent des représentations contextuelles :

$$h_i = \text{Transformer}(t_i), \quad i \in [1, m]$$

où $h_i \in R^d$ est le vecteur d'embedding associé au token t_i. Le vecteur de la séquence complète est extrait via le token spécial `[CLS]` :

$$z_S = h_{\texttt{[CLS]}} \in R^d$$

En parallèle, chaque intention candidate est reformulée en phrase naturelle (par ex. : "Ce message vise à informer", "Ce message vise à désinformer"). Ces phrases sont encodées dans le même espace vectoriel :

$$z_{L_j} = h^{(L_j)}_{\texttt{[CLS]}}, \quad j \in [1, k]$$

où k est le nombre d'intentions définies.

3.5.2.3 Projection dans un espace latent commun Les représentations z_S (message) et z_{L_j} (intentions) sont intégrées dans un espace sémantique commun grâce aux **Transformers pré-entraînés** (ex. BERT, RoBERTa). L'objectif est de maximiser la cohérence contextuelle entre le texte et l'intention.

On obtient ainsi :

$$Z = \{z_S, z_{L_1}, z_{L_2}, \ldots, z_{L_k}\} \subset R^d$$

3.5.2.4 Calcul de la similarité Le degré de correspondance entre un message et une intention est mesuré via la **similarité cosinus** :

$$\text{sim}(z_S, z_{L_j}) = \frac{z_S \cdot z_{L_j}}{\|z_S\| \cdot \|z_{L_j}\|}$$

Cette mesure varie entre $[-1, 1]$, avec une valeur proche de 1 indiquant une forte proximité sémantique.

3.5.2.5 Décision finale La décision est prise en sélectionnant l'intention dont la similarité est la plus élevée :

$$\hat{y} = \arg\max_{j \in [1,k]} \text{sim}(z_S, z_{L_j})$$

Afin d'éviter les erreurs sur des cas ambigus, un **seuil de confiance** τ peut être introduit :

$$\hat{y} = \begin{cases} \arg\max_j \text{sim}(z_S, z_{L_j}), & \max_j \text{sim}(z_S, z_{L_j}) \geq \tau \\ \text{rejet}, & \text{sinon} \end{cases}$$

3.5.2.6 Exemple illustratif Un tweet tel que :

"Partagez vite, le vaccin est dangereux et tue des enfants"

1. Prétraitement → tokenisation → encodage.
2. Projection dans l'espace latent.
3. Comparaison avec les intentions "Informer", "Désinformer", "Manipuler".
4. Calcul des scores → plus forte similarité avec l'intention "désinformation/manipulation".

Ainsi, le système étiquette automatiquement le message comme **désinformation**. [71]

3.6 Intégration de Zero-Shot et Transformers dans une architecture hybride de détection

La détection d'intention via Zero-Shot Learning (ZSL) et Transformers s'intègre dans une architecture hybride multimodale, combinant l'analyse de contenu, la modélisation des interactions sociales et, le cas échéant, les signaux multimédias. Cette approche améliore à la fois la robustesse et l'interprétabilité du système.

Analyse NLP du contenu

Le module NLP exploite les Transformers pré-entraînés (BERT, RoBERTa, DeBERTa, T5) pour encoder le texte du message dans un espace vectoriel dense $\mathbf{h}_t \in R^d$, où d est la dimension de l'embedding contextuel.

Prétraitement linguistique :

— Normalisation : conversion en minuscules, suppression des ponctuations inutiles.

— Tokenisation : WordPiece pour BERT, Byte-Pair Encoding (BPE) pour RoBERTa, ou SentencePiece pour T5.

— Nettoyage social : suppression des hashtags, mentions, URLs, emojis si nécessaire.

Encodage Transformers :

$$\mathbf{h}_t = \text{TransformerEncoder}(\text{Tokens}(x_t))$$

où x_t est le texte du message. Cette représentation capture la sémantique fine, les relations syntaxiques et les patterns rhétoriques (hyperboles, exagérations, manipulation de sentiment). Les embeddings peuvent ensuite être utilisés pour calculer un score de véracité ou de risque de désinformation via un classifieur linéaire ou MLP.

Outils pratiques : HuggingFace Transformers (BertModel, RobertaModel), PyTorch, TensorFlow.

Modélisation des interactions sociales via GNN

Les Graph Neural Networks (GNN) représentent les utilisateurs et messages comme un graphe $G = (V, E)$:

— $V = \{v_i\}$: nœuds représentant messages et utilisateurs.

— $E = \{e_{ij}\}$: arêtes représentant interactions (partage, réponse, mention, retweet).

Propagation de l'information :

$$\mathbf{h}_v^{(l+1)} = \sigma\Big(\sum_{u \in N(v)} W^{(l)} \mathbf{h}_u^{(l)} + b^{(l)} \Big)$$

où $\mathbf{h}_v^{(l)}$ est l'embedding du nœud v à la couche l, $N(v)$ est l'ensemble des voisins, $W^{(l)}$ les poids appris, $b^{(l)}$ le biais et σ une fonction d'activation (ReLU, GELU).

Cette étape permet de capturer :

— Les motifs de propagation rapide ou ciblée.

— Les clusters d'utilisateurs partageant des contenus similaires.

— Les relations structurées entre utilisateurs et messages.

Outils pratiques : PyTorch Geometric (GCNConv, GraphSAGE), DGL (dgl.nn), NetworkX pour l'analyse préliminaire des graphes.

Zero-Shot Learning pour la détection d'intention

Le Zero-Shot Learning repose sur la projection conjointe du texte et des labels sémantiques dans un espace vectoriel partagé $Z \subset R^d$.

Texte du message : $x_t \to \mathbf{h}_t = f_\theta(x_t)$

Label intention : $y_j \to \mathbf{h}_{y_j} = f_\theta(y_j)$

Similarité cosinus :

$$s(x_t, y_j) = \frac{\mathbf{h}_t \cdot \mathbf{h}_{y_j}}{\|\mathbf{h}_t\| \|\mathbf{h}_{y_j}\|}$$

Prédiction :

$$\hat{y} = \arg\max_j s(x_t, y_j)$$

Les labels sont formulés comme phrases naturelles : "Ce message vise à informer" ou "Ce message vise à désinformer". Cette approche permet de classifier un message dans une catégorie d'intention non vue pendant l'entraînement, éliminant le besoin de datasets annotés exhaustifs.

Outils pratiques : HuggingFace (pipeline("zero-shot-classification")), modèles `bart-large-mnli`, `roberta-large-mnli`.

Fusion multimodale

Si le message contient des images, vidéos ou audio, ces signaux sont encodés via des modèles adaptés :

— Images : CLIP (clip-ViT-B/32) $\rightarrow$ embedding $\mathbf{h}_i \in R^d$

— Audio : Wav2Vec2 $\rightarrow$ embedding $\mathbf{h}_a \in R^d$

La fusion multimodale s'effectue par concaténation ou attention pondérée :

$$\mathbf{h}_{fusion} = \text{Attention}([\mathbf{h}_t, \mathbf{h}_i, \mathbf{h}_a])$$

Cette représentation combinée capture les indices de désinformation qui ne sont pas visibles dans le texte seul.

Pipeline de décision et interprétabilité

Après encodage et fusion :

$$\text{Score global} = \alpha \cdot s_{\text{NLP}} + \beta \cdot s_{\text{ZSL}} + \gamma \cdot s_{\text{GNN}} + \delta \cdot s_{\text{MM}}$$

où $\alpha, \beta, \gamma, \delta$ sont des poids ajustables selon l'importance relative des modules. Un seuil de confiance τ permet de rejeter les prédictions ambiguës.

Cette combinaison offre :

— Une robustesse accrue face aux messages manipulatoires.

— Une explicabilité, car chaque module fournit un score séparé justifiant la décision finale.

Exemple concret :

Message : "Le vaccin modifie l'ADN humain et tue des enfants"

— ZSL : intention dominante = "désinformer" ($s_{\text{ZSL}} = 0.92$)

— GNN : propagation rapide dans un cluster homogène ($s_{\text{GNN}} = 0.87$)

— NLP : contenu linguistiquement manipulatoire ($s_{\text{NLP}} = 0.78$)

Score global combiné : $0.85 > \tau \Rightarrow$ message classé comme fake news.

3.7 Avantages et Contributions de l'Approche Hybride

L'intégration synergique du Traitement du Langage Naturel (NLP) et des Réseaux de Neurones Graphiques (GNN) dans la détection des fausses informations offre des avantages substantiels par rapport aux approches unimodales. Cette section explore les contributions majeures et les bénéfices de cette méthodologie hybride.

3.7.1 Analyse Multidimensionnelle et Complète

L'un des principaux avantages de l'approche hybride est sa capacité à effectuer une analyse multidimensionnelle et complète de l'information. Les fausses informations sont des phénomènes complexes qui ne peuvent être pleinement caractérisés par le seul contenu textuel ou la seule structure de propagation. Elles sont souvent conçues pour manipuler à la fois le message et la manière dont il est diffusé.

- **Intégration du Contenu et du Contexte** : Le module NLP excelle dans l'extraction de caractéristiques sémantiques, stylistiques et linguistiques du texte, permettant de comprendre ce qui est dit. Le module GNN, quant à lui, modélise les relations complexes entre les entités (utilisateurs, messages) et la dynamique de propagation, révélant comment l'information est diffusée et par qui. En combinant ces deux perspectives, le modèle hybride peut identifier des signaux de désinformation qui seraient invisibles pour une approche isolée. Par exemple, un message au contenu apparemment inoffensif pourrait être identifié comme faux s'il est diffusé par un réseau de bots ou par des utilisateurs ayant un historique de propagation de fausses informations.

- **Vue Holistique du Phénomène** : Cette approche permet une vue plus holistique du phénomène de la désinformation. Elle reconnaît que la véracité d'une information ne dépend pas uniquement de son contenu intrinsèque, mais aussi du contexte dans lequel elle est partagée et des acteurs impliqués dans sa diffusion. Cette compréhension globale est essentielle pour une détection précise et robuste.

3.7.2 Robustesse face à la Sophistication des Fausses informations

Les créateurs de fausses informations adaptent constamment leurs tactiques pour contourner les systèmes de détection. Les approches hybrides sont intrinsèquement plus robustes face à cette sophistication.

- **Résilience aux Manipulations** : Si un acteur malveillant tente de rendre le contenu textuel d'une fausse nouvelle indétectable par le NLP (par exemple, en utilisant un langage neutre et factuel), le GNN peut toujours identifier des schémas de propagation suspects (par exemple, une diffusion rapide et coordonnée par des comptes inauthentiques). Inversement, si la structure de propagation est masquée (par exemple, en utilisant des comptes humains légitimes), le NLP peut détecter des anomalies dans le contenu. Cette complémentarité rend le système plus difficile à tromper.
- **Détection des Nuances** : Les fausses informations peuvent prendre de nombreuses formes, allant de la désinformation flagrante à la mésinformation involontaire, en passant par la satire ou les opinions biaisées. L'approche hybride, en combinant des signaux de contenu et de structure, est mieux équipée pour distinguer ces nuances et éviter les faux positifs ou les faux négatifs. Par exemple, un article satirique pourrait avoir un contenu qui semble trompeur pour un NLP pur, mais le GNN pourrait identifier qu'il est partagé au sein d'une communauté connue pour la satire, réduisant ainsi le risque de classification erronée.

3.7.3 Potentiel de Détection Précoce

La rapidité de propagation est une caractéristique clé des fausses informations, rendant la détection précoce essentielle pour limiter leur impact. L'approche hybride offre un avantage significatif à cet égard.

- **Identification des Signaux Faibles** : Les GNN sont particulièrement efficaces pour identifier des schémas de propagation anormaux dès les premières étapes de la diffusion d'une information. Même avec un nombre limité d'interactions, la structure du graphe (par exemple, des clusters d'utilisateurs suspects, des schémas de retweet non organiques) peut révéler des signaux faibles indiquant une fausse nouvelle. Le NLP peut alors analyser le contenu de ces messages émergents pour confirmer les soupçons.
- **Réduction du Temps de Réponse** : En identifiant rapidement les fausses informations, les plateformes peuvent prendre des mesures correctives plus tôt, telles que la suppression du contenu, l'ajout d'avertissements, ou la réduction de sa visibilité. Cela permet de limiter l'exposition des utilisateurs à la désinformation et de freiner sa propagation virale avant qu'elle n'atteigne une audience massive.

3.7.4 Amélioration de l'Interprétabilité des Résultat

L'interprétabilité des modèles d'IA est cruciale, surtout dans des domaines sensibles comme la détection des fausses informations, où les décisions peuvent avoir des implications importantes. L'approche hybride peut offrir une meilleure interprétabilité.

- **Attribution des Causes** : En ayant des modules distincts pour le contenu (NLP) et la structure (GNN), il est possible de mieux comprendre pourquoi un message est classé comme faux. Le modèle peut indiquer si la décision est principalement basée sur des anomalies linguistiques dans le texte, sur des schémas de propagation suspects, ou sur une combinaison des deux. Par exemple, un message pourrait être signalé parce que son contenu est incohérent avec des faits connus (signal NLP) et parce qu'il est partagé par un réseau de bots (signal GNN).
- **Visualisation des Preuves** : Les représentations graphiques des réseaux de propagation peuvent être visualisées, permettant aux analystes humains de comprendre les chemins de diffusion et d'identifier les acteurs clés. De même, les techniques d'explicabilité du NLP (par exemple, l'attention des Transformers) peuvent mettre en évidence les parties du texte qui ont le plus contribué à la classification. Cette transparence aide les fact-checkers et les modérateurs à valider les décisions du modèle et à affiner leurs propres stratégies.

3.7.5 Apport du Zero-Shot Learning à la Détection d'Intention

Le Zero-Shot Learning (ZSL) représente une contribution majeure à la détection d'intentions dans le cadre de la lutte contre les fausses informations. Contrairement aux méthodes supervisées classiques qui nécessitent des jeux de données massifs et annotés, le ZSL permet de classifier des messages dans des catégories jamais vues lors de l'entraînement. Cette capacité est rendue possible grâce à l'utilisation de modèles de langage pré-entraînés, tels que BERT, RoBERTa ou GPT, qui capturent des représentations sémantiques universelles.

Le principe repose sur la projection conjointe des messages et des étiquettes d'intentions formulées en langage naturel dans un espace vectoriel partagé. La similarité cosinus entre ces représentations permet d'attribuer une intention probable au message. Par exemple, pour un tweet affirmant que « le vaccin modifie l'ADN humain », le système peut comparer sa représentation avec celles de labels tels que « informer », « désinformer » ou « manipuler », et déterminer qu'il correspond le mieux à l'intention « désinformer ».

Ainsi, le ZSL permet de déployer rapidement des systèmes de détection sans dépendre de processus longs et coûteux d'annotation, tout en maintenant une capacité d'adaptation à de nouvelles formes de fausses informations [72].

3.7.6 Intégration du Zero-Shot avec les Modules NLP et GNN

L'apport du ZSL prend toute sa valeur lorsqu'il est intégré dans une architecture hybride combinant NLP et GNN. Chaque module joue un rôle complémentaire :

- Le **NLP** extrait les caractéristiques linguistiques, stylistiques et sémantiques du message.
- Le **GNN** modélise la structure sociale de diffusion et met en évidence les schémas relationnels suspects.
- Le **Zero-Shot Learning** identifie l'intention sous-jacente au message, qu'il s'agisse d'informer, de désinformer, de manipuler ou de divertir.

Le processus d'intégration consiste à aligner les sorties des trois modules dans un espace décisionnel commun. Par exemple, un message peut sembler informatif selon son contenu (NLP), mais être diffusé de manière anormale par un cluster de comptes automatisés (GNN), tout en portant l'intention de manipuler (ZSL). La combinaison de ces signaux hétérogènes renforce la robustesse du modèle hybride, en réduisant la vulnérabilité aux manipulations isolées sur une seule dimension d'analyse.

Cette complémentarité permet également d'améliorer la précision et la granularité des classifications, en intégrant non seulement le « quoi » (contenu), le « comment » (diffusion), mais aussi le « pourquoi » (intention).

3.7.7 Vers une Architecture Multimodale Étendue

L'intégration du Zero-Shot Learning dans l'approche hybride NLP+GNN ouvre la voie à des extensions multimodales. En effet, les fausses informations ne reposent pas uniquement sur le texte : elles s'appuient également sur des images, vidéos et contenus audio qui renforcent leur impact.

Une architecture multimodale peut associer plusieurs sources de données :

- **NLP** pour l'analyse textuelle,
- **Vision par ordinateur** pour le traitement des images et vidéos,
- **Speech-to-Text et analyse audio** pour les messages vocaux ou les podcasts,
- **Zero-Shot Learning** pour la détection d'intentions transversales à toutes ces modalités.

Par exemple, une vidéo contenant un discours conspirationniste peut être transcrite puis analysée par le module NLP, tandis que les images extraites sont vérifiées par un modèle de détection de deepfakes. En parallèle, le ZSL attribue une intention probable telle que « manipuler l'opinion publique ». Enfin, le GNN peut révéler que ce contenu circule intensivement dans des communautés homogènes, renforçant ainsi l'indice de désinformation.

Cette intégration multimodale enrichie par le Zero-Shot Learning offre une analyse plus robuste, capable de prendre en compte la diversité des signaux exploités par les acteurs malveillants. Elle constitue une étape clé vers la mise en place de systèmes de détection complets, adaptatifs et interprétables.

3.8 Défis, Limites et Perspectives Future

Malgré les avantages significatifs de l'approche hybride NLP+GNN pour la détection des fausses informations, plusieurs défis et limites subsistent. La recherche dans ce domaine est en constante évolution, et ces défis ouvrent la voie à de nouvelles perspectives de recherche et d'innovation.

3.8.1 Défis Liés à la Complexité et à la Dynamique des Données

Les réseaux sociaux sont des environnements extrêmement complexes et dynamiques, ce qui pose des défis majeurs pour la modélisation et l'analyse des graphes.

- **Taille et Densité des Graphes** : Les réseaux sociaux peuvent contenir des milliards de nœuds (utilisateurs, messages) et des trillions d'arêtes (interactions). La construction, le stockage et le traitement de graphes d'une telle envergure sont des défis computationnels et de mémoire considérables. Les GNN traditionnels peuvent avoir du mal à s'adapter à de très grands graphes, nécessitant des techniques d'échantillonnage ou de partitionnement.
- **Nature Dynamique et Évolutive** : Les réseaux sociaux évoluent constamment. De nouveaux utilisateurs rejoignent, de nouvelles relations se forment, et de nouveaux messages sont publiés en permanence. Les modèles doivent être capables de s'adapter à ces changements en temps réel, ce qui est difficile pour les architectures statiques. La détection des fausses informations doit être un processus continu et adaptatif.
- **Hétérogénéité des Données** : Comme mentionné précédemment, les réseaux sociaux sont hétérogènes, avec différents types de nœuds (utilisateurs, messages, images, vidéos) et d'arêtes (suivre, retweeter, aimer, commenter). La modélisation efficace de cette hétérogénéité tout en capturant les interactions complexes entre ces différents types d'entités reste un défi.

3.8.2 Coût Computationnel et Scalabilité

L'entraînement et le déploiement de modèles hybrides NLP+GNN peuvent être très coûteux en termes de ressources computationnelles.

- **Entraînement des Modèles Transformers** : Les modèles NLP basés sur les Transformers (comme BERT ou RoBERTa) sont gourmands en ressources, nécessitant des GPU puissants et de longues heures d'entraînement, même pour le fine-tuning. Leur utilisation dans des pipelines de détection en temps réel peut devenir un véritable goulot d'étranglement.
- **Opérations GNN sur Grands Graphes** : Les opérations d'agrégation et de combinaison dans les GNN peuvent devenir très coûteuses à mesure que la taille du graphe augmente. Bien que des techniques d'échantillonnage de voisins (comme GraphSAGE) aient été développées pour améliorer la scalabilité, le traitement de graphes massifs reste un défi. Le déploiement de ces modèles en production, où la latence est critique, nécessite des optimisations significatives.

3.8.3 Problématiques des Données Annotées et du Biais

La disponibilité de données de haute qualité est essentielle pour l'entraînement de modèles d'apprentissage automatique, mais elle pose des problèmes spécifiques dans le contexte de la détection des fausses informations.

- **Manque de Données Annotées à Grande Échelle** : L'annotation manuelle des fausses informations est un processus coûteux, chronophage et subjectif. Il est difficile de construire des jeux de données à grande échelle qui couvrent la diversité des fausses informations et des contextes de propagation. De plus, les étiquettes de vérité terrain peuvent être contestées ou évoluer avec le temps.
- **Biais dans les Données d'Entraînement** : Les jeux de données existants peuvent contenir des biais inhérents, reflétant par exemple les types de fausses informations qui ont été historiquement signalées ou les préjugés des annotateurs. Ces biais peuvent être appris par le modèle, conduisant à des performances inégales sur différents types de désinformation ou pour différentes communautés.

- **Problème du « Cold Start »** : Pour les nouvelles informations ou les nouveaux utilisateurs, il y a peu ou pas de données de propagation disponibles au début. Cela rend la détection précoce difficile, car le GNN a besoin d'un certain nombre d'interactions pour construire une représentation significative du graphe.

3.8.4 Évolution constante des tactiques de désinformation

Les acteurs malveillants derrière la création et la diffusion de fausses informations sont adaptatifs et innovants. Ils apprennent des systèmes de détection existants et développent de nouvelles tactiques pour les contourner.

- **Sophistication Linguistique** : Les fausses informations sont de plus en plus difficiles à distinguer des vraies en termes de style et de sémantique, rendant le travail du module NLP plus ardu.
- **Manipulation des Réseaux** : Les diffuseurs peuvent utiliser des stratégies complexes pour manipuler les schémas de propagation, comme l'utilisation de comptes dormants, la coordination discrète, ou l'exploitation de vulnérabilités dans les algorithmes de recommandation des plateformes.

Cela signifie que les modèles de détection doivent être continuellement mis à jour et réentraînés pour rester efficaces, ce qui pose un défi en termes de maintenance et de ressources.

3.8.5 Défis liés à la formulation des intentions dans le Zero-Shot Learning

La performance du Zero-Shot Learning (ZSL) dépend directement de la précision et de la cohérence des labels formulés en langage naturel. En effet, le modèle s'appuie sur la similarité sémantique entre la représentation du texte d'entrée et la description des intentions pour effectuer la classification. Or, dans le cas de la détection des fausses informations, la diversité et l'ambiguïté des intentions posent des contraintes majeures.

— **Variabilité des intentions** : les intentions peuvent aller de la mésinformation involontaire à la désinformation stratégique, en passant par la satire ou le contenu manipulé.

— **Sensibilité à la formulation** : une description imprécise ou culturellement biaisée du label peut réduire fortement la robustesse de la classification.

— **Alignement sémantique limité** : les modèles pré-entraînés sur des corpus généraux ne capturent pas toujours la subtilité des intentions spécifiques au contexte de la désinformation.

Ainsi, la conception et la normalisation des labels d'intention constituent un enjeu méthodologique crucial pour la fiabilité du ZSL.

3.8.6 Défis liés à l'adaptation contextuelle et culturelle

Un des points critiques du ZSL réside dans sa capacité limitée à prendre en compte la variabilité contextuelle et culturelle. Les modèles sémantiques sont majoritairement entraînés sur des données généralistes et anglophones, ce qui induit plusieurs limites :

— **Manque de contextualisation locale** : une même information peut être considérée comme trompeuse dans un contexte politique ou culturel donné, mais neutre dans un autre.

— **Variabilité linguistique** : les nuances de traduction, les idiomes et les spécificités lexicales influencent la qualité de l'inférence sémantique.

— **Propagation multi-communautés** : les dynamiques de diffusion varient fortement selon les groupes sociaux, et les modèles ZSL peinent à capturer ces différences.

Dans une approche hybride NLP+GNN, l'intégration de graphes sociaux enrichis (par exemple avec des métadonnées culturelles ou linguistiques) apparaît nécessaire pour dépasser ces limitations.

3.8.7 Défis computationnels et robustesse du Zero-Shot Learning

Le ZSL, lorsqu'il repose sur des architectures Transformers (BERT, RoBERTa, GPT, etc.), implique un coût computationnel particulièrement élevé, ce qui limite son application dans des environnements à grande échelle comme les réseaux sociaux.

— **Complexité des modèles** : l'inférence en Zero-Shot nécessite souvent une recherche sémantique entre le texte et un grand ensemble de labels, augmentant le temps de calcul.

— **Problème de scalabilité** : l'intégration du ZSL dans des pipelines temps réel se heurte aux contraintes de latence et de ressources GPU.

— **Vulnérabilité aux attaques adversariales** : de légères modifications lexicales ou syntaxiques (injection de synonymes, paraphrases adversariales) peuvent tromper le modèle.

Ces défis soulèvent la question de la résilience des systèmes hybrides, qui doivent combiner précision, scalabilité et robustesse face aux environnements adverses.

3.8.8 Perspectives de recherche (GNN hétérogènes, apprentissage auto-supervisé, explicabilité, multimodalité avancée)

Face à ces défis, plusieurs pistes de recherche prometteuses émergent pour améliorer l'approche hybride NLP+GNN.

- **GNN Hétérogènes Avancés** : Développer des architectures GNN plus sophistiquées capables de gérer nativement la complexité des graphes hétérogènes des réseaux sociaux. Cela inclut la modélisation de différents types de nœuds (utilisateurs, messages, hashtags, images) et d'arêtes (retweet, mention, réponse, relation d'amitié) de manière plus nuancée, en apprenant des représentations spécifiques pour chaque type et en les combinant efficacement.
- **Apprentissage Auto-supervisé sur Graphes** : Étant donné le manque de données annotées, l'apprentissage auto-supervisé (Self-Supervised Learning - SSL) sur graphes est une voie prometteuse. Les techniques de SSL peuvent apprendre des représentations de nœuds et de graphes à partir de données non étiquetées en résolvant des tâches auxiliaires (par exemple, prédiction de liens, reconstruction de caractéristiques masquées). Ces représentations pré-entraînées peuvent ensuite être fine-tunées avec un petit ensemble de données étiquetées, réduisant ainsi la dépendance aux annotations coûteuses.
- **Explicabilité des Modèles (XAI)** : Améliorer l'interprétabilité des modèles hybrides est crucial pour la confiance et l'adoption. La recherche se concentre sur le développement de techniques d'explicabilité de l'IA (XAI) qui peuvent non seulement dire si une information est fausse, mais aussi pourquoi elle est considérée comme telle. Cela peut inclure la mise en évidence des mots ou phrases clés dans le texte qui ont contribué à la décision, ou la visualisation des sous-graphes de propagation les plus influents.
- **Détection Multimodale Avancée** : Les fausses informations ne se limitent pas au texte. Elles peuvent inclure des images, des vidéos et des audios manipulés. L'intégration de modules d'analyse d'images (vision par ordinateur) et d'audio (traitement du signal) avec les modules NLP et GNN est une direction clé. Cela permettrait de construire des modèles véritablement multimodaux capables de détecter les fausses informations quel que soit leur format, en exploitant les incohérences entre les différentes modalités (par exemple, un texte affirmant un fait contredit par l'image associée).
- **Apprentissage Continu et Adaptatif** : Développer des systèmes capables d'apprendre et de s'adapter en continu aux nouvelles tactiques de désinformation. Cela pourrait impliquer des approches d'apprentissage en ligne (online learning) ou d'apprentissage par renforcement, où le modèle est constamment mis à jour à mesure que de nouvelles données et de nouvelles tactiques émergent.

Ces perspectives de recherche visent à rendre les systèmes de détection des fausses informations plus intelligents, plus robustes et plus adaptatifs, capables de faire face à la complexité croissante du paysage de la désinformation en ligne.

3.9 Problématique

Les fausses informations sur les réseaux sociaux constituent un défi persistant pour la recherche et l'ingénierie. Leur rapidité de diffusion, favorisée par les mécanismes viraux propres aux plateformes, rend les interventions tardives inefficaces. En parallèle, la sophistication croissante des contenus trompeurs — allant des *deepfakes* aux formulations linguistiques ambiguës — complexifie leur détection automatique.

Les approches actuelles présentent des limites structurelles :

— Les méthodes centrées sur le contenu textuel peinent à généraliser face au langage informel, aux abréviations, emojis ou manipulations contextuelles ;

— Les méthodes basées sur la propagation nécessitent un temps d'observation incompatible avec une détection précoce ;

— Les modèles multimodaux restent sensibles aux variations de domaine et aux contournements stratégiques.

Ainsi, la problématique centrale est la suivante :

comment concevoir des méthodes de détection capables d'opérer de manière fiable, rapide et robuste dans un environnement marqué par la diversité linguistique, la multimodalité et l'évolution constante des tactiques liées aux fausses informations ?

3.10 Proposition et contribution

3.11 Fondements conceptuels de l'approche hybride

Notre proposition repose sur l'idée que les fausses informations présentent des signaux distinctifs à plusieurs niveaux :

1. **Contenu textuel** : elles véhiculent des caractéristiques linguistiques, stylistiques et sémantiques spécifiques, détectables par des modèles NLP avancés.
2. **Propagation** : les schémas de diffusion sur les réseaux sociaux diffèrent souvent de ceux des informations fiables, produisant des structures graphiques que les GNN peuvent modéliser.
3. **Intention** : au-delà du texte explicite, l'auteur poursuit une finalité communicative (informative, persuasive, trompeuse, ironique). Détecter cette intention constitue un signal supplémentaire, particulièrement utile pour identifier des stratégies discursives de manipulation.
4. **Multimodalité** : les fausses informations exploitent diverses modalités (texte, image, vidéo, audio), rendant nécessaire une analyse conjointe afin de capter des incohérences inter-modales.

L'objectif de cette approche est de dépasser les limites d'une analyse centrée sur une seule dimension et de construire une architecture hybride capable de combiner contenu, diffusion, intention et multimodalité dans un cadre unifié.

3.12 Architecture du modèle hybride NLP-GNN multimodal

Notre architecture se compose de quatre modules principaux interconnectés :

Module NLP pour l'analyse du contenu

Ce module repose sur les modèles Transformer afin d'extraire des représentations contextuelles riches du texte :

— Prétraitement linguistique : normalisation, segmentation, lemmatisation et suppression des mots vides adaptées à la langue du contenu.

— Extraction de caractéristiques : indicateurs de subjectivité, complexité syntaxique, incohérences sémantiques, polarité émotionnelle.

— Codage sémantique profond : utilisation de modèles pré-entraînés (BERT, RoBERTa, CamemBERT) pour capturer les nuances lexicales et contextuelles.

— Vérification factuelle : comparaison avec des bases de connaissances externes pour repérer des contradictions avec des faits établis.

Module de détection de l'intention via Zero-shot Transformer

L'analyse de l'intention vise à identifier la finalité implicite du message : informer, manipuler, influencer ou tromper.

— Méthode Zero-shot : emploi d'un Transformer adapté (ex. XLM-R, BART) permettant de classifier les intentions sans jeu d'entraînement spécifique.

— Apport complémentaire : enrichit la représentation textuelle par un signal discursif, utile pour repérer les manipulations indirectes comme l'ironie ou les insinuations.

Module GNN pour l'étude de la diffusion

La propagation est représentée sous forme de graphe dynamique, enrichi de métadonnées :

— Construction du graphe de diffusion : nœuds = utilisateurs, arêtes = interactions (partages, commentaires, mentions), pondérées par des attributs temporels.

— Caractéristiques des nœuds : ancienneté du compte, centralité dans le réseau, historique d'activité.

— Modélisation : utilisation de GCN, GraphSAGE ou GAT pour apprendre des représentations capturant les motifs de diffusion atypiques associés aux fausses informations.

Module multimodal

Ce module intègre différents types de contenu associés au message :

— Texte : analyse NLP détaillée.

— Images : détection de manipulations visuelles, cohérence texte-image.

— Vidéos / audios : repérage de falsifications (deepfakes, vocaux synthétiques).

— Fusion multimodale : projection des représentations dans un espace latent commun pour combiner texte, image et audio de manière cohérente.

Module de combinaison et classification

Ce module fusionne les signaux issus des composants précédents :

— Attention croisée pour pondérer dynamiquement l'importance du contenu, de l'intention, de la diffusion et des modalités.

— Apprentissage de représentations conjointes via des couches non linéaires.

— Classification finale estimant la véracité des informations à partir de la représentation hybride.

— Interprétabilité grâce à l'inspection des poids d'attention, indiquant les segments textuels, indices multimodaux et structures de diffusion déterminants.

3.13 Avantages et contributions attendues

L'approche proposée présente plusieurs contributions majeures :

— Détection précoce améliorée : en combinant contenu, intention et propagation, le modèle peut repérer des signaux faibles dès les premières étapes de diffusion.

— Robustesse accrue : une manipulation visant uniquement une dimension (texte ou image) est compensée par l'analyse des autres modalités.

— Adaptabilité contextuelle : l'attention croisée permet au système de s'ajuster selon la nature de l'information (même langue, registre ou modalité).

— Meilleure généralisation : la diversité des signaux intégrés favorise une portabilité entre domaines et langues.

— Interprétabilité renforcée : l'analyse explicite des intentions et de la multimodalité améliore la transparence et la confiance dans le système.

3.14 Défis d'implémentation et perspectives

La mise en œuvre de cette architecture pose plusieurs défis :

— Fusion hétérogène : développer des techniques d'intégration optimales entre représentations textuelles, intentionnelles, multimodales et graphiques.

— Données limitées : rareté de jeux annotés combinant texte, intention et multimodalité, nécessitant des stratégies de transfert learning et d'augmentation de données.

— Scalabilité et temps réel : adapter l'architecture pour fonctionner à grande échelle sur des flux massifs.
— Évolution des tactiques adversariales : intégrer des mécanismes d'adaptation continue afin de résister aux nouvelles formes de manipulation (deepfakes plus réalistes, ironie sophistiquée, contournements des filtres).

3.15 Conclusion

Ce chapitre a montré la pertinence de l'approche hybride NLP+GNN pour détecter les fausses informations dans les réseaux sociaux. Elle combine l'analyse sémantique des contenus et l'étude des dynamiques de propagation, offrant une vision plus complète du phénomène. Cette synergie améliore la robustesse, la précision et la généralisation des systèmes, tout en renforçant leur interprétabilité. L'apport du Zero-Shot Learning ouvre de nouvelles perspectives pour la détection d'intentions et l'adaptabilité multilingue. Cependant, plusieurs défis persistent : coût computationnel, manque de données fiables, biais et évolution rapide des tactiques adversariales. Ces limites soulignent l'importance d'optimiser la scalabilité et l'adaptation continue aux contextes variés. En somme, l'approche hybride constitue une piste prometteuse pour concevoir des systèmes plus fiables et évolutifs contre la désinformation

Chapitre IV :Proposition d'un système hybride pour la détection de la désinformation sur les réseaux sociaux

4.1 Introduction

La désinformation constitue aujourd'hui un défi majeur pour les sociétés connectées : la facilité de production et de diffusion de contenus (texte, images, vidéos) permet à des informations erronées ou manipulées d'atteindre rapidement un large public, avec des conséquences parfois graves (panique sanitaire, manipulation électorale, préjudice individuel). Pour limiter ces risques, une approche consiste à détecter et filtrer, en amont, le contenu avant sa publication effective. Un tel système vise à évaluer la crédibilité d'un message au moment où l'auteur le soumet, puis à autoriser, mettre en quarantaine ou bloquer sa diffusion selon le score et les politiques en vigueur.

Un système de détection pré-publication ambitionne donc de réduire la propagation initiale des fausses informations, mais soulève plusieurs contraintes : coût de calcul à l'échelle (analyser chaque message), risques de faux positifs (censure abusive), et exigences éthiques (transparence, recours). Le présent chapitre décrit une proposition technique — un prototype hybride et multimodal — qui combine des modules de traitement du texte, de la vision, de la vidéo, de l'audio et de l'analyse de graphes pour maximiser la robustesse de la détection.

4.2 Architecture conceptuelle

4.2.1 Interfaces client

Les interfaces client rassemblent toutes les façons dont un auteur prépare et soumet un contenu à vérifier : application mobile, interface web, extension de navigateur, ou utilitaires de test (ligne de commande / GUI). Leur rôle principal est de collecter le texte et les fichiers multimédias (images, vidéos, audio) ainsi que des métadonnées contextuelles (ID utilisateur, horodatage, langue, provenance) puis d'initier une requête vers le serveur d'inférence.

Pour préserver l'expérience utilisateur, le client réalise une pré-validation légère : vérification des formats de fichiers (ex. .jpeg/.png pour images, .mp4 pour vidéos, .wav pour audio), contrôle des tailles maximales (par ex. 5 MB pour images, 50 MB pour vidéos dans le mode synchrone), et éventuellement un filtrage local simple (détection de langage, suppression de contenu manifestement interdit).

Le format d'échange côté client est une requête HTTP POST /predict en multipart/form-data contenant un champ texte (`text`) et des fichiers nommés (`image`, `video`, `audio`, `graph_json`). En plus du contenu, le client envoie des métadonnées minimales (par ex. `user_id`, `language`, `timestamp`) afin d'aider les modules de prétraitement et l'audit.

Enfin, le client doit afficher une rétroaction claire au rédacteur : si la vérification est synchrone, afficher le score et la recommandation (publier / mettre en quarantaine / corriger) ; si le traitement est asynchrone (vidéos lourdes), afficher un message indiquant que l'analyse est en cours et fournir un identifiant de requête pour consulter ultérieurement le résultat.

4.2.2 Serveur d'inférence (concept)

Le serveur d'inférence est le cœur du système : il reçoit les requêtes clients, exécute le prétraitement par modalité, appelle les encodeurs spécialisés, effectue la fusion des représentations et renvoie une décision accompagnée d'un score et d'un diagnostic.

Le serveur applique des contrôles de sécurité et de format (types MIME, taille), stocke temporairement les fichiers si nécessaire et transforme les flux bruts en formats prêts pour les modèles (par ex. `input_ids` pour BERT, tenseurs images normalisés, séquences de frames échantillonnées pour la vidéo, spectrogrammes pour l'audio, structures `edge_index` pour le graphe).

Le prétraitement est conçu pour tolérer des entrées partielles : si une modalité est absente (texte seul, pas d'image), le serveur exécute le pipeline avec les modalités disponibles et utilise des vecteurs nuls ou des embeddings de substitution pour les modalités manquantes.

Le moteur de décision applique des règles métier (seuils configurables) pour déterminer l'action recommandée : publier automatiquement, mettre en quarantaine pour revue humaine, ou bloquer provisoirement.

4.2.3 Pipeline d'apprentissage et stockage

Le pipeline d'apprentissage est une infrastructure hors-ligne responsable de la préparation des jeux de données annotés, de l'entraînement, de la validation et du versioning des artefacts (modèles, tokenizers,

normalisations).

Il commence par la collecte et le nettoyage des données (texte, multimédia, graphes), la normalisation et la création d'un mapping d'étiquettes cohérent (par ex. binaire Fake / Real).

L'entraînement combine transfert learning (poids pré-entraînés) et fine-tuning si ressources suffisantes. Les résultats (rapports de classification, matrices de confusion, courbes ROC/PR) sont sauvegardés et associés aux artefacts.

Une boucle de rétroaction intègre périodiquement les cas vérifiés par des humains dans le dataset pour améliorer le modèle (apprentissage continu).

4.3 Présentation d'Anaconda comme outil d'exécution du projet

Le système repose sur de nombreuses bibliothèques de machine learning (PyTorch, Transformers, Scikit-learn), de traitement multimédia (OpenCV, MoviePy) et de communication réseau (Flask). Pour garantir une installation simple, reproductible et compatible, Anaconda a été utilisé comme gestionnaire d'environnements.

Anaconda est une distribution libre et multiplateforme spécialisée pour la data science. Elle intègre Python et R, un gestionnaire de paquets (`conda`), ainsi qu'un système d'environnements virtuels isolés. Elle fournit aussi Jupyter Notebook, qui permet de tester et visualiser des résultats de manière interactive.

4.4 L'apprentissage automatique

L'apprentissage automatique (ou *machine learning*) est une branche de l'intelligence artificielle qui consiste à développer des systèmes capables d'apprendre à partir de données, plutôt que d'être explicitement programmés.

4.4.1 Fonctionnement général

Un système de machine learning repose sur plusieurs étapes : - collecte et préparation des données, - choix du modèle adapté, - entraînement (optimisation des paramètres), - évaluation (sur données de test), - déploiement (dans une application réelle).

4.4.2 Types d'apprentissage

On distingue : - l'apprentissage supervisé (à partir de données annotées), - l'apprentissage non supervisé (structure dans données non étiquetées), - l'apprentissage par renforcement (essais-erreurs avec récompenses).

4.4.3 Importance et applications

Le machine learning est au cœur de : détection de désinformation, recommandations (Netflix, YouTube), reconnaissance vocale, diagnostic médical, cybersécurité, véhicules autonomes.

4.4.4 Avantages et limites

Avantages : traitement massif de données, automatisation, adaptabilité.

Limites : dépendance aux données de qualité, coût matériel élevé (GPU), risque de biais et manque de transparence.

4.5 Encodeurs : détails techniques et recommandations

4.5.1 Texte

BERT (bert-base-uncased, 768-d) projeté en 256 dimensions via `nn.Linear`. Option légère : DistilBERT. Recommandation : geler BERT dans un premier temps, fine-tuner seulement si ressources suffisantes.

4.5.2 Image

ResNet18 pré-entraîné sur ImageNet. La dernière couche FC est supprimée (vecteur 512 → ´ en 256). Fine-tuning recommandé uniquement sur les derniers blocs.

4.5.3 Vidéo

R3D-18 (entrées : B,3,T,H,W avec T=8–16). Produit un vecteur de dimension 512 projeté en 256. Poids pré-entraînés conseillés.

4.5.4 Audio

Extraction MFCC (40 coefficients), suivi d'un petit CNN (Conv2D → → → † ´ en 64 dimensions.

4.5.5 Graphe

GCN avec deux couches GCNConv (→64, ⋆ ∞ , , 64. † absence de PyG, fallback en vecteur nul.

4.5.6 Fusion

Concaténation des vecteurs : texte (256) + image (256) + vidéo (256) + graphe (64) + audio (64) = 896. Passage dans un MLP (896 → 512 → 2 f. 0.3 ´gularisation.

4.6 Racine du projet et fichiers principaux

4.6.1 Arborescence

```
multimodal_project/
 train.py
 server.py
 client_cli.py
 client_gui.py
 prepare_liar.py
 datasets/LIAR/{train.tsv,valid.tsv,test.tsv}
 models/
    text_encoder.py
    image_encoder.py
    video_encoder.py
    audio_encoder.py
    graph_encoder.py
    fusion.py
 outputs/
     model.pt
```

4.6.2 Description des fichiers

client_gui.py :

train.py : Script principal d'entraînement. Il charge les données (via `prepare_liar.py`), initialise les encodeurs (texte, image, vidéo, audio, graphe), entraîne le modèle multimodal et sauvegarde les poids entraînés dans `outputs/model.pt`.

server.py : Met en place un serveur (Flask ou équivalent) qui charge le modèle entraîné et reçoit des requêtes des clients. Son rôle est d'effectuer l'inférence (détection de la désinformation) en temps réel.

client_cli.py : Client en mode terminal (CLI) permettant à un utilisateur d'envoyer un message ou un fichier multimédia au serveur pour vérification.

client_gui.py : Client avec interface graphique (GUI) pour interagir facilement avec le système et tester le fonctionnement comme un utilisateur réel.

prepare_liar.py : Contient les fonctions de préparation du dataset LIAR (lecture des fichiers TSV, nettoyage, création des DataLoaders pour entraînement, validation et test).

datasets/LIAR : Contient `train/valid/test` :

- `train.tsv` : données d'entraînement,
- `valid.tsv` : données de validation,
- `test.tsv` : données de test.

Chaque ligne correspond à une déclaration annotée comme vraie ou fausse, utilisée pour apprendre et évaluer le modèle.

models/ : Ensemble des encodeurs texte, image, vidéo, audio, graphe et module de fusion.

models/text_encoder.py : Définit l'encodeur texte (BERT ou DistilBERT), qui transforme une phrase en vecteur de caractéristiques.

models/image_encoder.py : Définit l'encodeur image (ResNet18), qui extrait des représentations visuelles à partir d'images.

models/video_encoder.py : Définit l'encodeur vidéo (R3D-18), qui capture l'information temporelle et spatiale des vidéos.

models/audio_encoder.py : Définit l'encodeur audio basé sur MFCC et CNN, qui extrait des caractéristiques spectrales et temporelles des fichiers audio.

models/graph_encoder.py : Définit l'encodeur graphe (GCN), qui permet de traiter des données structurées sous forme de graphes.

models/fusion.py : Contient le module de fusion multimodale, qui concatène les vecteurs issus des différents encodeurs et applique un MLP pour prédire la classe (désinformation ou information fiable).

outputs/model.pt : Fichier de sortie contenant les poids du modèle entraîné, chargé par `server.py` pour faire de l'inférence.

4.7 Codes principaux

Code serveur (`server.py`)

```

import os, io, tempfile, json
from fastapi import FastAPI, UploadFile, File, Form
from fastapi.responses import JSONResponse
import uvicorn
import torch
import numpy as np
from PIL import Image
import cv2

from models.text_encoder import TextEncoder
from models.image_encoder import ImageEncoder
from models.video_encoder import VideoEncoder
from models.audio_encoder import AudioEncoder
from models.graph_encoder import GraphEncoder
from models.fusion import FusionClassifier
from transformers import BertTokenizer

app = FastAPI()
DEVICE = torch.device("cuda" if torch.cuda.is_available() else "cpu")
MODEL_PATH = "outputs/model.pt"

# Preprocessors
from torchvision import transforms
image_transform = transforms.Compose([
    transforms.Resize((224,224)),
    transforms.ToTensor(),
```

```
    transforms.Normalize(mean=[0.485,0.456,0.406], std=[0.229,0.224,0.225])
])
video_frame_transform = transforms.Compose([
    transforms.Resize((112,112)),
    transforms.ToTensor()
])

# Instantiate encoders
text_enc = TextEncoder(out_dim=256).to(DEVICE)
img_enc = ImageEncoder(out_dim=256).to(DEVICE)
vid_enc = VideoEncoder(out_dim=256).to(DEVICE)
aud_enc = AudioEncoder(out_dim=64).to(DEVICE)
graph_enc= GraphEncoder()
fusion  = FusionClassifier(tdim=256,idim=256,vdim=256,gdim=64,adim=64,
                           hidden=512,nclass=2).to(DEVICE)
tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")

# Load state
if os.path.exists(MODEL_PATH):
    state = torch.load(MODEL_PATH, map_location=DEVICE)
    try:
        text_enc.proj.load_state_dict(state["text_proj"])
        img_enc.proj.load_state_dict(state["img_proj"])
        vid_enc.proj.load_state_dict(state["vid_proj"])
        aud_enc.load_state_dict(state["audio_state"])
        fusion.load_state_dict(state["fusion"])
        print("Loaded saved model state.")
    except Exception as e:
        print("Failed to load full state:", e)
else:
    print("No trained model found - using untrained fusion.")

def read_video_frames_from_file(path, max_frames=16):
    cap = cv2.VideoCapture(path)
    frames=[]
    total = int(cap.get(cv2.CAP_PROP_FRAME_COUNT))
    if total <=0:
        cap.release(); return None
    indices = np.linspace(0, total-1, min(max_frames,total), dtype=int)
    idx_set = set(indices.tolist())
    i=0; fetched=0
    while True:
        ret, frame = cap.read()
        if not ret:
            break
        if i in idx_set:
            frame = cv2.cvtColor(frame, cv2.COLOR_BGR2RGB)
            pil = Image.fromarray(frame)
            t = video_frame_transform(pil) # C,H,W
            frames.append(t)
            fetched +=1
        i+=1
        if fetched>=len(idx_set):
            break
    cap.release()
    if len(frames)==0: return None
    stack = torch.stack(frames, dim=1)
    return stack

@app.post("/predict")
async def predict(
```

```
    text: str = Form(None),
    image: UploadFile = File(None),
    video: UploadFile = File(None),
    audio: UploadFile = File(None),
    graph: UploadFile = File(None)
):
    try:
        # TEXT
        if text:
            enc = tokenizer(text, truncation=True, padding=True,
                            max_length=128, return_tensors="pt")
            input_ids = enc["input_ids"].to(DEVICE)
            attn = enc["attention_mask"].to(DEVICE)
            with torch.no_grad():
                t_feat = text_enc(input_ids, attn)
        else:
            t_feat = torch.zeros(1,256, device=DEVICE)

        # IMAGE
        if image:
            b = await image.read()
            pil = Image.open(io.BytesIO(b)).convert("RGB")
            img_t = image_transform(pil).unsqueeze(0).to(DEVICE)
            with torch.no_grad():
                i_feat = img_enc(img_t)
        else:
            i_feat = torch.zeros(1,256, device=DEVICE)

        # VIDEO
        if video:
            tmp = tempfile.NamedTemporaryFile(delete=False, suffix=".mp4")
            tmp.write(await video.read()); tmp.close()
            frames = read_video_frames_from_file(tmp.name, max_frames=16)
            os.unlink(tmp.name)
            if frames is not None:
                v_input = frames.unsqueeze(0).to(DEVICE).float()
                with torch.no_grad():
                    v_feat = vid_enc(v_input)
            else:
                v_feat = torch.zeros(1,256, device=DEVICE)
        else:
            v_feat = torch.zeros(1,256, device=DEVICE)

        # AUDIO
        if audio:
            tmp = tempfile.NamedTemporaryFile(delete=False, suffix=".wav")
            tmp.write(await audio.read()); tmp.close()
            try:
                import torchaudio
                wav, sr = torchaudio.load(tmp.name)
                if sr != aud_enc.sample_rate:
                    resampler = torchaudio.transforms.Resample(
                                orig_freq=sr, new_freq=aud_enc.sample_rate)
                    wav = resampler(wav)
                wav_mono = wav.mean(dim=0, keepdim=True)
                with torch.no_grad():
                    a_feat = aud_enc(wav_mono.to(DEVICE))
            except Exception as e:
                print("Audio read error:", e)
                a_feat = torch.zeros(1,64, device=DEVICE)
            finally:
```

```
                os.unlink(tmp.name)
        else:
            a_feat = torch.zeros(1,64, device=DEVICE)

        # GRAPH
        g_feat = torch.zeros(1,64, device=DEVICE)

        # Fusion & predict
        with torch.no_grad():
            logits = fusion(t_feat, i_feat, v_feat, g_feat, a_feat)
            probs = torch.softmax(logits, dim=1).cpu().numpy()[0].tolist()
            pred = int(torch.argmax(logits, dim=1).item())
        label = "Real" if pred==1 else "Fake"
        return JSONResponse({"prediction": label,
                             "pred": pred,
                             "probs": probs})
    except Exception as e:
        return JSONResponse({"error": str(e)}, status_code=500)

if __name__ == "__main__":
    print("Starting server on http://0.0.0.0:8000")
    uvicorn.run(app, host="0.0.0.0", port=8000)
```

Code client (client_gui.py)

```
import tkinter as tk
from tkinter import filedialog, messagebox
import requests, os

URL = "http://127.0.0.1:8000/predict"

class App:
    def __init__(self, root):
        self.root = root
        root.title("Client Multimodal Upload")
        root.geometry("600x450")

        tk.Label(root, text="Texte:").pack()
        self.text = tk.Text(root, height=4, width=70)
        self.text.pack()

        self.img_var = tk.StringVar()
        tk.Button(root, text="Choisir Image", command=self.choose_image).pack()
        tk.Label(root, textvariable=self.img_var).pack()

        self.vid_var = tk.StringVar()
        tk.Button(root, text="Choisir Vido", command=self.choose_video).pack()
        tk.Label(root, textvariable=self.vid_var).pack()

        self.aud_var = tk.StringVar()
        tk.Button(root, text="Choisir Audio", command=self.choose_audio).pack()
        tk.Label(root, textvariable=self.aud_var).pack()

        self.graph_var = tk.StringVar()
        tk.Button(root, text="Choisir Graphe (JSON)", command=self.choose_graph).pack()
        tk.Label(root, textvariable=self.graph_var).pack()

        tk.Button(root, text="Envoyer au serveur",
                  bg="green", fg="white",
                  command=self.send).pack(pady=10)
```

```
36
37          self.output = tk.Text(root, height=6, width=70)
38          self.output.pack()
39
40      def choose_image(self):
41          p = filedialog.askopenfilename(filetypes=[("Images","*.jpg *.jpeg *.png")])
42          if p: self.img_var.set(p)
43      def choose_video(self):
44          p = filedialog.askopenfilename(filetypes=[("Vidos","*.mp4 *.avi *.mov")])
45          if p: self.vid_var.set(p)
46      def choose_audio(self):
47          p = filedialog.askopenfilename(filetypes=[("Audio","*.wav *.mp3")])
48          if p: self.aud_var.set(p)
49      def choose_graph(self):
50          p = filedialog.askopenfilename(filetypes=[("JSON","*.json")])
51          if p: self.graph_var.set(p)
52
53      def send(self):
54          text = self.text.get("1.0","end-1c").strip()
55          files = {}
56          data = {}
57          if text: data["text"] = text
58          try:
59              if self.img_var.get():
60                  files["image"] = (os.path.basename(self.img_var.get()),
61                                    open(self.img_var.get(),"rb"),
62                                    "image/jpeg")
63              if self.vid_var.get():
64                  files["video"] = (os.path.basename(self.vid_var.get()),
65                                    open(self.vid_var.get(),"rb"),
66                                    "video/mp4")
67              if self.aud_var.get():
68                  files["audio"] = (os.path.basename(self.aud_var.get()),
69                                    open(self.aud_var.get(),"rb"),
70                                    "audio/wav")
71              if self.graph_var.get():
72                  files["graph"] = (os.path.basename(self.graph_var.get()),
73                                    open(self.graph_var.get(),"rb"),
74                                    "application/json")
75              resp = requests.post(URL, data=data, files=files, timeout=180)
76              if resp.status_code==200:
77                  self.output.delete(1.0, tk.END)
78                  self.output.insert(tk.END, resp.text)
79              else:
80                  messagebox.showerror("Erreur",
81                                       f"{resp.status_code}: {resp.text}")
82          except Exception as e:
83              messagebox.showerror("Erreur", str(e))
84          finally:
85              for f in files.values():
86                  try: f[1].close()
87                  except: pass
88
89  if __name__ == "__main__":
90      root = tk.Tk()
91      App(root)
92      root.mainloop()
```

4.8 Avantages et limites

4.8.1 Avantages

Amélioration de la précision Grâce à l'approche multimodale (texte, image, vidéo, audio, graphe), le système exploite plusieurs sources d'information. Cela permet de mieux distinguer les vraies informations des fake news, même lorsqu'elles sont déguisées sous forme de vidéos ou d'images manipulées.

Détection en temps réel (ou quasi temps réel) Avec le serveur d'inférence (server.py), les messages envoyés par les clients , peuvent être vérifiés immédiatement avant d'être publiés. Cela limite la propagation de la désinformation avant qu'elle ne devienne virale.

Flexibilité et extensibilité Le système est modulaire : chaque encodeur (texte, image, vidéo, audio, graphe) est indépendant. On peut donc ajouter, remplacer ou améliorer un encodeur sans modifier toute l'architecture.

Apprentissage automatique continu Le pipeline d'entraînement (train.py) permet de réentraîner le système dès que de nouvelles données annotées sont disponibles. Cela améliore la robustesse face à l'évolution des fake news.

Application pratique Avec les clients (CLI ou GUI), le système peut être testé directement comme un vrai outil de détection utilisable dans un réseau social.

4.8.2 Inconvénients et limites

Coût computationnel élevé Les encodeurs comme BERT (texte) et ResNet (image/vidéo) nécessitent beaucoup de ressources (GPU, mémoire). Cela limite le déploiement à grande échelle, surtout en temps réel.

Dépendance aux données d'entraînement,le modèle apprend à partir du dataset (ex. LIAR). Si les données sont limitées ou biaisées, les résultats seront eux aussi biaisés et moins généralisables.

Difficulté d'accès aux bases de faits en temps réel Vérifier les informations contre des bases externes (fact-checkers, sources fiables) est complexe et nécessite des mises à jour constantes.

Problèmes éthiques et juridiques,risque de censure abusive si un contenu légitime est bloqué.

Problème de transparence : les utilisateurs doivent comprendre pourquoi leur contenu est rejeté. Responsabilité légale en cas d'erreur de classification.

4.9 Conclusion

La proposition d'un système hybride de détection de la désinformation sur les réseaux sociaux montre qu'il est possible d'agir en amont, avant même que les fausses informations ne soient diffusées à grande échelle. Grâce à une approche multimodale combinant l'analyse du texte, de l'image, de la vidéo, de l'audio et des graphes sociaux, ce système peut offrir une vision plus complète et donc une détection plus fiable que les approches classiques.

Toutefois, cette avancée s'accompagne de limites techniques et éthiques : le besoin en ressources matérielles est important, la dépendance aux données annotées reste forte, et la frontière entre modération efficace et censure abusive demeure délicate à gérer.

En somme, ce type de système ne doit pas être vu comme une solution parfaite, mais comme un outil d'aide à la décision destiné à renforcer la lutte contre la désinformation. Son efficacité dépendra d'un équilibre entre l'automatisation et la supervision humaine, ainsi que de l'intégration de mécanismes de transparence et de mise à jour continue.

5 Conclusion générale

La désinformation constitue aujourd'hui l'un des plus grands défis de l'ère numérique. Avec l'essor des réseaux sociaux, les fake news connaissent une diffusion rapide et massive, bouleversant l'accès à une information fiable et menaçant la stabilité sociale, économique et politique. Ce mémoire s'est inscrit dans cette problématique en explorant, à travers une approche progressive, les dimensions théoriques, techniques et conceptuelles de la détection des fausses informations.

Dans le premier chapitre, nous avons mis en lumière la nature et les formes variées des fausses nouvelles. Cette partie a permis d'établir une typologie claire des différents types de désinformation (rumeurs, propagande, satire trompeuse, clickbait, deepfakes, etc.), tout en soulignant leurs caractéristiques spécifiques. L'analyse a montré que les fake news ne sont pas de simples erreurs ou malentendus, mais bien des constructions délibérées ayant un impact profond sur la société et nécessitant des outils adaptés pour être combattues.

Le deuxième chapitre s'est focalisé sur les réseaux sociaux, principaux vecteurs de diffusion de la désinformation. Nous avons examiné leur fonctionnement, leur classification et leurs usages, en montrant comment leurs mécanismes internes (algorithmes de recommandation, viralité, bulles informationnelles) favorisent parfois la propagation des contenus trompeurs. Nous avons également étudié les enjeux associés (sociaux, économiques et politiques), ainsi que les stratégies déjà mises en place par les plateformes et les chercheurs pour détecter les fausses nouvelles. Toutefois, l'analyse a révélé les limites de ces approches traditionnelles, souvent centrées sur une seule modalité (texte) ou ne prenant pas en compte la complexité des dynamiques sociales.

Dans le troisième chapitre, nous avons présenté les approches hybrides, en mettant en évidence l'apport du Traitement Automatique du Langage Naturel (NLP) pour l'analyse du contenu textuel et celui des Réseaux de Neurones Graphiques (GNN) pour l'étude des interactions sociales et des graphes de propagation. Nous avons également montré l'intérêt d'intégrer des techniques multimodales capables de traiter simultanément le texte, les images, les vidéos et les audios. Cette approche hybride et multimodale s'est révélée particulièrement prometteuse, car elle permet non seulement d'améliorer la précision des détections, mais aussi d'anticiper des formes de désinformation de plus en plus sophistiquées, comme les deepfakes.

Enfin, le quatrième chapitre a été consacré à la proposition d'une architecture conceptuelle de système hybride de détection. Nous avons détaillé ses principaux modules : collecte et préparation des données, analyse multimodale, détection par NLP et GNN, puis décision et réponse. Cette proposition, bien que théorique, offre une vision claire et structurée d'un système capable de répondre aux enjeux posés par la désinformation dans les environnements numériques modernes.

En conclusion, ce travail a montré qu'une approche hybride et multimodale constitue une réponse pertinente au défi de la désinformation sur les réseaux sociaux. Toutefois, des perspectives restent ouvertes pour des recherches ultérieures, notamment :

l'amélioration de la robustesse des modèles face aux manipulations toujours plus sophistiquées ;

l'optimisation des performances pour un déploiement en temps réel à grande échelle ;

l'intégration de mécanismes éthiques et juridiques afin de respecter la liberté d'expression tout en luttant efficacement contre la désinformation ;

l'adaptation de ces systèmes à des contextes multilingues et culturels variés.

Ainsi, le présent mémoire ambitionne de contribuer à l'avancement des recherches dans ce domaine, tout en offrant une base théorique et conceptuelle pour la conception de systèmes de détection plus performants, capables de protéger l'espace numérique et de renforcer la confiance dans l'information

Références

[1] Mathieu-Robert Sauvé. *Les fake news dans les médias du Québec : perceptions des journalistes*. Mémoire de maîtrise, Université de Sherbrooke, page 15, 2019.

[2] Claire Wardle. *Fake news. Itś complicated*. First Draft News, 2017.

[3] Gary D. Bond and Adrienne Y. Lee. Language of lies in prison : Linguistic classification of prisonerśtruthful and deceptive natural language. *Applied Cognitive Psychology*, 19(3) :313–329, 2005.

[4] Glynis Bogaard, Ewout H. Meijer, Aldert Vrij, and Harald Merckelbach. Scientific Content Analysis (SCAN) cannot distinguish between truthful and fabricated accounts of a negative event. *Frontiers in Psychology*, 7 :243, 2016.

[5] Melissa Zimdars. *False, Misleading, Clickbait-y, and Satirical "News"Sources*. Google Docs, 2016.

[6] Jennifer Golbeck et al. Fake News vs Satire : A Dataset and Analysis. In *Proc. of the 10th ACM Conference on Web Science*, pages 17–21, 2018.

[7] J. Bernstein. *Bad News*. Harperś Magazine, septembre 2021. https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation/

[8] Juan Cao et al. Exploring the role of visual content in fake news detection. In *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161, 2020.

[9] Alvaro Figueira and Luciana Oliveira. The current state of fake news : challenges and opportunities. *Procedia Computer Science*, 121 :817–825, 2017.

[10] M. Vasilogambros. *Election Disinformation Fears Came True For State Officials*. Pew, 2020. https://www.pewtrusts.org/en/research-and-analysis/blogs/stateline/2020/11/20/election-disinformation-fears-came-true-for-state-officials

[11] Alexis Kochel. *Misinformation variation ? Looking through the gendered lens*, 2023.

[12] Dorje C. Brody and David M. Meier. How to model fake news. *arXiv preprint arXiv :1809.00964*, 2018.

[13] M. Westerlund. The Emergence of Deepfake Technology : A Review. *Technology Innovation Management Review*, 9(11) :40–53, 2019. https://timreview.ca/article/1282

[14] B. Paris and J. Donovan. Deepfakes and Cheap Fakes : The Manipulation of Audio and Visual Evidence. Data Society, 2019. https://datasociety.net/library/deepfakes-and-cheap-fakes/

[15] R. Chesney and D. Citron. Deep Fakes : A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, 107(6) :1753–1819, 2019. https://doi.org/10.2139/ssrn.3213954

[16] Ian Goodfellow et al. Generative adversarial networks. *Communications of the ACM*, 63(11) :139–144, 2020.

[17] [Fouad Jabiri. *Applications de méthodes de classification non supervisées à la détection d'anomalies*. PhD thesis, Université Laval, 2020.

[18] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2) :211–236, 2017.

[19] Srijan Kumar and Neil Shah. False information on web and social media : A survey. *arXiv preprint arXiv :1804.08559*, 2018.

[20] Misako Takayasu et al. Rumor diffusion and convergence during the 3.11 earthquake : a Twitter case study. *PLoS One*, 10(4) :e0121443, 2015.

[21] Carlos Carvalho, Nicholas Klagge, and Emanuel Moench. The persistent effects of a false news shock. *Journal of Empirical Finance*, 18(4) :597–615, 2011.

[22] Kiran Garimella et al. Balancing information exposure in social networks. In *Advances in Neural Information Processing Systems*, pages 4663–4671, 2017.

[23] Michael Barthel, Amy Mitchell, and Jesse Holcomb. Many Americans believe fake news is sowing confusion. *Pew Research Center*, 15 :12, 2016.

[24] Kai Shu et al. Fake News Detection on Social Media : A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter*, 19(1) :22–36, 2017.

[25] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. *Me, My Echo Chamber, and I : Introspection on Social Media Polarization*. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, pages 823–831, 2018.

[26] Zhou, X., Jain, A., Phoha, V. V., Zafarani, R. Fake News Early Detection : An Interdisciplinary Study. *arXiv preprint arXiv :1904.11679*, 2019.

[27] N. Diakopoulos. Algorithmic accountability. *Digital Journalism*, 3(3) :398–415, 2015.

[28] S. Vosoughi, D. Roy, and S. Aral. The spread of true and false news online. *Science*, 359(6380) :1146–1151, 2018.

[29] B. D. Horne and S. Adali. This just in : Fake news packs a lot in title, uses simpler, repetitive language. 2017.

[30] Svitlana Volkova and Jin Yea Jang. Misleading or Falsification : Inferring Deceptive Strategies and Types in Online News and Social Media. In *Companion of The Web Conference 2018*, pages 575–583.

[31] Rada Mihalcea and Carlo Strapparava. The lie detector : Explorations in the automatic recognition of deceptive language. In *ACL-IJCNLP 2009 Conference Short Papers*, pages 309–312.

[32] Natali Ruchansky, Sungyong Seo, and Yan Liu. CSI : A Hybrid Deep Model for Fake News Detection. In *Proc. of the 2017 ACM Conference on Information and Knowledge Management*, pages 797–806.

[33] Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web (WWW)*, pp. 675–684. ACM.

[34] Feng Qian et al. Neural User Response Generator : Fake News Detection with Collective User Intelligence. In *IJCAI-18*, pages 3834–3840. https://doi.org/10.24963/ijcai.2018/533

[35] Kai Shu et al. FakeNewsNet : A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv preprint arXiv :1809.01286*, 2018.

[36] Clayton Allen Davis et al. Botornot : A System to Evaluate Social Bots. In *Proc. of the 25th International Conference Companion on World Wide Web*, pages 273–274, 2016.

[37] Wang, W. Y., Yuan, Y. (2020). Fake news detection on social media : A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1), 22–36.

[38] William Yang Wang. "Liar, Liar Pants on Fire" : A New Benchmark Dataset for Fake News Detection. In *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 422–426, 2017. https://doi.org/10.18653/v1/P17-2067

[39] Mehrdad Farajtabar et al. Fake News Mitigation via Point Process Based Intervention. In *Proc. of the 34th ICML*, Vol. 70, pages 1097–1106, 2017.

[40] Jooyeon Kim et al. Leveraging the Crowd to Detect and Reduce the Spread of Fake News and Misinformation. In *Proc. of the 11th ACM WSDM*, 2018.

[41] Pennycook, G., Rand, D. G. (2021). The Psychology of Fake News. Trends in Cognitive Sciences, 25(5), 388–402.

[42] Liu, Y., Wu, Y. (2018). Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 354–361). AAAI.

[43] Kotteti, S., Roy, A., Bansal, M. (2020). Stance-aware rumour verification in social media using propagation trees. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1431–1441). ACL.

[44] Zhu, Z., Peng, W., Wang, J., Huang, L. (2015). Measuring the semantic consistency of rumors for detection. In *Proceedings of the 24th International Conference on World Wide Web Companion* (pp. 909–914). ACM.

[45] Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K. F., Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence* (pp. 3818–3824).

[46] Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., Cook, J. (2012). Misinformation and its correction : Continued influence and successful debiasing. Psychological Science in the Public Interest, 13(3), 106–131.

[47] Nyhan, B., Reifler, J. (2010). When Corrections Fail : The Persistence of Political Misperceptions. Political Behavior, 32(2), 303–330.

[48] Bruns, Debra P., Kraguljac, Nina V., Bruns, Thomas R. (2020). COVID-19 : Facts, Cultural Considerations, and Risk of Stigmatization. Journal of Transcultural Nursing, 31(4), 326–332.

[49] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H. E., Quattrociocchi, W. (2016). The spreading of misinformation online. Proceedings of the National Academy of Sciences, 113(3), 554–559.

[50] University of Southern California. (2021)

[51] Russell, S., & Norvig, P. (2010). *Artificial Intelligence : A Modern Approach* (3rd ed.). Prentice Hall.//

[52] Caplan, R., Boyd, D. (2018). "I tweet honestly, I tweet passionately" : Twitter users, context collapse, and the imagined audience. New Media Society, 13(1), 114–133.

[53] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[54] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[55] Starbird, K. (2019). Disinformation's orchestrated chaos. Journal of Communication, 69(3), 327–329.

[56] Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., Menczer, F. (2018). The spread of low-credibility content by social bots. Nature Communications, 9(1), 4787.

[57] Ferrara, E. (2020). Bots, elections, and social media : A brief overview. In S. Bashir A. Campolo (Eds.), Disinformation, Misinformation, and Fake News in Social Media (pp. 95–114). Springer.

[58] Boyd, D. M., Ellison, N. B. (2007). Social network sites : Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1), 210–230.

[59] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[60] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

[61] Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv :1301.3781.

[62] Pennington, J., Socher, R., Manning, C. (2014). GloVe : Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543.

[63] Kaplan, A. M., Haenlein, M. (2010). Users of the world, unite ! The challenges and opportunities of Social Media. *Business Horizons*, 53(1), 59–68.

[64] Sunstein, C. R. (2018). Republic : Divided democracy in the age of social media. Princeton University Press.

[65] Z. Zhang, Q. Lv, X. Jia, W. Yun, G. Miao, Z. Mao and G. Wu, "GBCA : Graph Convolution Network and BERT combined with Co-Attention for fake news detection," *Pattern Recognition Letters*, vol. 180, pp. 26–32, 2024. doi :10.1016/j.patrec.2024.02.014.

[66] W. Jin, Y. Gao, T. Tao, X. Wang, N. Wang, B. Wu and B. Zhao, "Veracity-Oriented Context-Aware Large Language Models–Based Prompting Optimization for Fake News Detection," *International Journal of Intelligent Systems*, vol. 2025, no. 1, pp. 5920142, 2025. doi :10.1155/INT/5920142.

[67] Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring Minds : Early Detection of Rumors in Social Media from Enquiry Posts. *Proceedings of the 24th International Conference on World Wide Web (WWW)*, 1395–1405.

[68] Z. Yi, C. Tang and S. Lu, "User Comment-Guided Cross-Modal Attention for Interpretable Multimodal Fake News Detection," *Applied Sciences*, vol. 15, no. 14, p. 7904, 2025. doi :10.3390/app15147904.

[69] Wenpeng Yin, Daniel Khashabi, Chris Callison-Burch, and Matt Gardner. *Benchmarking Zero-shot Text Classification : Datasets, Evaluation and Entailment Approach*. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3914–3923, 2019.

[70] Yin, W., & Schütze, H. (2017). Attentive Zero-Shot Text Classification with Label Representation Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[71] Brown, T., et al. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*.

[72] Yin, W., Hay, J., & Roth, D. (2019). Benchmarking Zero-shot Text Classification : Datasets, Evaluation and Entailment Approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[73] Centre de crise de Belgique. *Types des fausses informations*. Disponible sur : https://centredecrise.be/fr/risques-en-belgique/risques-pour-la-securite/desinformation/desinformation

[74] Aleteia. *Fausse image : deepfake et désinformation*. Publié le 13 juin 2024. Disponible sur : https://fr.aleteia.org/2024/06/13/deep-fake-chatbot-six-fois-ou-les-fideles-catholiques-ont-ete-abuse-par-lia

[75] Laval Virtual. *Fausse vidéo : comment détecter des deepfakes de plus en plus réalistes*. Disponible sur : https://blog.laval-virtual.com/comment-detecter-des-deepfake-de-plus-en-plus-realistes/

[76] Presidio. *Architecture du réseau GAN*. Disponible sur : https://www.presidio.com/exploring-the-power-of-generative-adversarial-networks-gans-with-azure/

[77] Inria. *Architecture de l'auto-encodeur*. Disponible sur : https://sciml.gitlabpages.inria.fr/scimllectures/chapAPsec5.html

[78] Marketing91. *Types des réseaux sociaux*. Disponible sur : https://www.marketing91.com/7-types-of-social-media-channels

[79] Humanities and Social Sciences Communications (Nature Publishing Group). *Enjeux des réseaux sociaux*. Disponible sur : https://www.nature.com/articles/s41599-023-02441-z

[80] ResearchGate. *Architecture CNN + GNN (CNN-RNN architecture)*. Disponible sur : https://www.researchgate.net/figure/CNN-RNN-architecture_fig1_353254139

[81] Scientific Reports (Nature Publishing Group). *Approche ultimodale*. Disponible sur : https://www.nature.com/articles/s41598-025-05702-w

[82] Fraidoon Omarzai (GoPenAI). *Schéma comparatif des techniques d'embedding lexical*. Disponible sur : https://blog.gopenai.com/all-word-embedding-techniques-in-depth-768780914f6c

[83] DataScientest. *Réseaux de neurones graphiques – tout savoir*. Disponible sur : https://datascientest.com/graph-neural-networks-tout-savoir

[84] Gowri Shankar. *Graph Convolution Network – A Practical Implementation of Vertex Classifier and Its Mathematical Basis*. Disponible sur : https://gowrishankar.info/blog/graph-convolution-network-a-practical-implementation-of-vertex-classifier-and-its-mathematical-basis/

[85] Epichka. *GAT Paper Explained – Graph Attention Network*. Disponible sur : https://epichka.com/blog/2023/gat-paper-explained/

[86] Illustration of GraphSAGE (ResearchGate). *GraphsSAGE : modèle avec stratégie d'échantillonnage*. Disponible sur : https://www.researchgate.net/figure/Illustration-of-GraphSAGE-model-with-a-sampling-strategy-Here-the-maximum-number-of_fig4_355873169