Dédicace

Je dédie ce travail à tous ceux qui me sont chers,

A

Ma mère,

Qui m'a soutenu toutes les fois quand je n'étais pas à l'aise. Elle est d'une importance primordiale puisque son amour est inconditionnel et inébranlable. J'espère ne jamais te décevoir ni trahir ta confiance et tes sacrifices.

A

Mon père,

De tous les pères. Tu es le meilleur tu as été et tu seras toujours un exemple pour moi par tes qualités humaines, ta persévérance et perfectionnisme. La relation amicale entre nous n'a pas de prix. . En ce jour, j'espère réaliser l'un de tes rêves.

A

Mon cher frère,

Tu étais toujours là pour me soutenir et m'encourager. Je vous souhaite une vie pleine de bonheur et de succès et que Dieu vous protège.

Au final aucune dédicace ne saurait exprimer mes respects, ma reconnaissance et tout Mon amour. Puisse dieu vous préserver et vous procurer santé et bonheur.

Rana

Remerciements

La construction de ce mémoire n'aurait été possible sans l'intervention de certaines personnes.

Je tiens à exprimer ici mes plus sincères remerciements pour leurs précieux conseils.

Je souhaite tout d'abord adresser mes vifs et chaleureux remerciements à Madame Souhir BOUAZIZ, mon encadrante académique, pour l'accompagnement tout au long de ce travail, qui n'aurait pu être mené à bien sans son soutien et ses encouragements. Ses conseils avisés, ses critiques constructives et ses qualités humaines m'ont été d'une aide précieuse pour mener à terme ce projet.

Je remercie également les membres du jury pour le temps qu'ils ont consacré à l'examen de ce mémoire et j'espère qu'ils y trouveront les qualités de clarté et de motivation attendues.

Enfin, j'exprime ma profonde gratitude envers toutes les personnes qui m'ont aidé, de près ou de loin, à accomplir ce travail.

Tables des matières

Liste des figures

Liste des Tableaux

Liste des abréviations

AE-GCN	**AutoEncoder Graph-Convolutional Network**
BERT	**Bidirectional Encoder Representation from Transformer**
Bi-LSTM	**Bidirectional Long Short-Term Memory**
BOW	**Bag Of Words**
CBOW	**Continuous Bag-Of-Words**
CNN	**Convolutional Neural Network**
CRF	**Conditional Random Field**
CV	**Count Vectorizer**
DDGCN	**Dual Dynamic Graph Convolutional Network**
DL	**Deep Learning**
DNN	**Deep Neural Network**
FND	**Fake News Detection**
FNR	**Fake News Revealer**
GAE	**Graph AutoEncoder**
GAL	**Generative Adversarial Learning**
GAN	**Generative Adversarial Network**
GCN	**Graph Convolutional Network**
GLAN	**Graph-based Linear Assignment Network**

GLOVE	**Global Vectors for Word Representation**
GRU	**Gated Recurrent Unit**
HAN	**Hierarchical Attention Networks**
H-MCAN	**Hierarchical Multi-modal Contextual Attention Network**
HMM	**Hidden Markov Models**
KG	**knowledge Graphs**
KMAGCN	**Knowledge-aware Multi-modal Adaptive Graph Convolutional Network**
k-NN	**k-Nearest Neighbors**
LSTM	**Long Short-Term Memory**
MCAN	**Multi-modal Co-Attention Network**
ML	**Machine Learning**
MLM	**Masked Language Model**
MSCCNN	**Memristor-Based Sparse Compact Convolutional Neural Network**
NLP	**Natural Language Processing**
NLTK	**Natural Language Toolkit**
NSP	**Next Sentence Prediction**
PLM	**Pre-trained Language Models**
RNN	**Recurrent Neural Network**
RoBERTa	**Robustly optimized BERT approach**
SA	**Sentiment Analysis**

SMAN	**Multi-head Structure-aware Attention Network**
SVM	**Support Vector Machine**
TF-IDF	**Term Frequency-Inverse Document Frequency**
UGC	**User-generated Content**
UPFD	**User Preference-aware Fake Detection**
VAE-GCN	**Variational AutoEncoder- Graph Convolutional Network**
VGAE	**Variational Graph Auto-Encoder**
WMD	**Word Mover's Distance**

Introduction générale

Les fausses nouvelles sont des informations délibérément créées pour être trompeuses ou mensongères. Elles sont conçues pour diffuser de la désinformation, et leur succès repose presque entièrement sur les lecteurs, qui ont le pouvoir de les partager ou de les arrêter (Özgöbek & Gulla, 2017). Les réseaux sociaux représentent actuellement l'allié le plus puissant des fausses nouvelles : un espace non supervisé et non réglementé où l'information est accessible à l'échelle mondiale et instantanément, et où les gens passent généralement seulement quelques secondes sur chaque publication. Ils constituent un moyen de diffusion parfaitement adapté pour propager la désinformation (Tacchini, et al., 2017) (Vedova, et al., 2018). Bien sûr, en raison de leur taille, les réseaux sociaux resteront non supervisés dans un avenir proche, du moins pas par des humains. Ainsi, les solutions envisagées sont (1) éduquer les utilisateurs à distinguer les fausses nouvelles des vraies ou (2) développer un système de détection des fausses nouvelles capable de bloquer ces dernières ou d'avertir les utilisateurs en cas de possible manque de véracité.

Cependant, avec la popularité croissante des médias sociaux largement utilisés à des fins politiques, le problème des fausses nouvelles a pris plus d'importance ces dernières années, posant ainsi un grand défi de détection. La diffusion rapide de fausses nouvelles peut causer des dommages considérables aux individus, à la société, voire aux pays. Dans ce contexte, le sentiment public exerce une influence croissante sur la société. Comprendre l'opinion publique est souvent nécessaire dans de nombreux domaines : de la stratégie de marque à la prédiction d'événements tels que les résultats d'élections. La capacité à mesurer précisément le sentiment public en temps quasi réel est donc très souhaitable. L'analyse des sentiments est l'une des approches permettant d'évaluer le sentiment public.

L'analyse des sentiments consiste à examiner les opinions, perceptions, attitudes, pensées et émotions des individus partagées sur différentes plateformes de médias sociaux. Plus précisément, l'analyse des sentiments vise à classer un texte particulier comme étant un sentiment neutre, positif ou négatif. Dang et al. (Dang, et al., 2020) ont identifié trois principales approches dans l'analyse des sentiments : basée sur l'apprentissage automatique, basée sur des lexiques et l'approche hybride. La détection des fausses nouvelles est difficile, et les études sur ce sujet en sont encore à leurs débuts.

Cependant, la prolifération des fausses nouvelles est en pleine expansion. Cela contribue à développer et à explorer des pistes de recherches permettant d’actualiser et d’améliorer les techniques d’identification de ces fausses nouvelles. De très nombreuses études sur l’état de l’art consacrées à l’identification des fausses nouvelles sur les réseaux sociaux exploitent soit une, soit plusieurs des caractéristiques d’une nouvelle comme le contenu, la propagation en réseau ou l’utilisateur. Les sentiments véhiculés dans les titres des nouvelles pourraient, en revanche, être au cœur de l’identification des fausses nouvelles en apportant un indice de la crédibilité de l’information. Dans ce sens, les techniques d’apprentissage profond récentes contribuent à l’amélioration, la classification, la prédiction ou l’analyse du contenu textuel grâce à leur capacité d’apprentissage efficace, d’extraction de caractéristiques et de modélisation des structures complices.

Dans ce travail, nous proposons un modèle pour la détection des fausses nouvelles qui s’appuie sur une analyse de sentiment prise comme un des points principaux permettant d’améliorer le

modèle de détection proposé. Ce modèle basé sur une approche d'apprentissage profond permet de traiter un vaste corpus de fausses nouvelles pour en extraire, grâce à l'analyse des sentiments, des caractéristiques susceptibles d'enrichir le modèle proposé par une information complémentaire basée sur l'analyse des sentiments des articles d'actualité. Ces caractéristiques viennent compléter les caractéristiques du contenu des nouvelles au sein du modèle de détection proposé pour permettre une plus grande performance du détecteur.

Cette approche se distingue de l'existant en ce qu'elle permet une prise en compte plus fine des émotions et sentiments, souvent ignorés dans les méthodes classiques de détection, apportant ainsi une dimension supplémentaire à la classification. De plus, grâce à l'utilisation de modèles avancés, il devient possible de traiter efficacement de grandes quantités de données, un aspect crucial dans la détection des fausses nouvelles où les ensembles de données sont massifs et diversifiés.

Dans le cadre de l'analyse des sentiments, nous avons calculé les polarités du texte. Les résultats de cette analyse sont ensuite fusionnés avec ceux de la tâche de classification, qui repose sur des modèles d'apprentissage profond tels que CNN, Bi-LSTM et BERT permettent l'extraction des caractéristiques complexes et contextuelles des articles afin de renforcer le modèle de détection avec des informations à la fois sémantiques et émotionnelles.

Dans ce projet, nous utilisons deux ensembles de données tel que ISOT et GossipCop qui reflètent respectivement les vraies et fausses nouvelles et les ressources officielles (vraies).

Les résultats expérimentaux confirment l'efficacité de notre approche : notre modèle basé sur BERT a atteint des performances élevées, notamment une précision de 99,34 % sur l'ensemble de données ISOT et de 97,54 % sur GossipCop. Ces résultats témoignent de la robustesse de BERT pour saisir les nuances complexes des contenus textuels, renforcée par l'ajout de l'analyse de sentiment, qui enrichit la compréhension contextuelle et émotionnelle des informations.

La structure de ce mémoire est organisée comme suit :

- Nous commençons par le chapitre 1 qui présente un aperçu des méthodes utilisées pour identifier les fausses nouvelles, en mettant l'accent sur celles qui se propagent sur les réseaux sociaux.
- Au chapitre 2, nous explorons l'état actuelle de la technologie permettant de détecter les fausses nouvelles et ainsi un accent particulier y est mis sur l'analyse de sentiments.
- Nous exposons au chapitre 3 notre principale contribution à la résolution du problème à l'étude.
- Enfin, au chapitre 4 nous présentons les résultats expérimentaux de notre système de détection de fausses informations, en explorant, analysant et comparant les performances de différents modèles d'apprentissage profond.

Chapitre 1 : Les fausses nouvelles dans les médias sociaux : Définitions et approches

1. Introduction

Ce chapitre vise à dresser un bilan des méthodes déployées pour reconnaître les fausses nouvelles, en particulier celles diffusées via les réseaux sociaux. Dans un premier temps, nous allons donner la définition des fausses nouvelles et, dans un deuxième temps, nous allons présenter leurs spécificités dans la façon dont elles circulent sur les médias sociaux. Dans un troisième temps, nous allons investiguer les diverses méthodes de détection retenues, notamment celles qui sont dites supervisées, faiblement supervisées et non supervisées. Enfin, nous conclurons ce chapitre en résumant les principaux points abordés.

2. Les fausses nouvelles

« Fake News » a été annoncé comme le mot officiel de l'année du Collins Dictionary pour 2017 (Liu & Wu, 2020). Les fausses nouvelles peuvent être définies comme des articles de presse publiés qui contiennent de fausses informations pour induire intentionnellement les lecteurs en erreur et à des fins malveillantes (A. Alkhodair, et al., 2020).

Ainsi, nous pouvons identifier trois aspects clés des fausses nouvelles : sa forme en tant qu'article de presse, son intention trompeuse et la vérifiabilité de son contenu comme complètement ou partiellement faux. (Wardle, 2017) a déconstruit les fausses nouvelles en sept catégories :

- une fausse connexion, lorsque les titres, les éléments visuels ou les légendes n'appuient pas le contenu;
- faux contexte, correspondant à un contenu authentique partagé avec de fausses informations contextuelles;
- le contenu manipulé, c'est-à-dire l'information authentique manipulée pour tromper;
- le contenu trompeur, qui implique une utilisation trompeuse de l'information pour encadrer une question ou une personne;
- le contenu d'imposteur, lorsque des sources authentiques sont usurpées;
- contenu fabriqué, entièrement faux, conçu pour tromper et nuire;
- satire/parodie, avec un potentiel de duperie mais aucune intention de nuire. Compte tenu de la nature non nuisible de ces nouvelles et parce qu'elles sont facilement identifiables comme parodiques , ce type de nouvelles n'est généralement pas considéré pour la détection de fausses nouvelles, bien que la satire puisse être utilisée comme excuse pour éviter l'accusation de répandre de fausses nouvelles.

La diffusion de fausses nouvelles sur Internet est plus rapide que la diffusion de vraies nouvelles (Ammara , et al., 2019), parce que les gens sont curieux de connaître des nouvelles, et ont tendance à partager les dernières informations, surtout en partageant des nouvelles de dernière heure sans vérifier leur véracité (Al-Rakhami & Al-Amri, 2020). Certaines fausses nouvelles sont publiées ou partagées involontairement et sont appelées désinformation (A. Alkhodair, et al., 2020). La visualisation répétée de fausses nouvelles les rend familières au destinataire, les rend crédibles et les fait circuler comme de vraies nouvelles. Cependant, il devient difficile de faire la distinction entre les fausses nouvelles et les vraies, car la recherche scientifique a révélé que la capacité humaine à distinguer les informations vraies et fausses est relativement faible,

avec un taux d'environ 54% (de Oliveira, et al., 2020). De plus, des chercheurs de l'Université Stanford ont mené des recherches sur la fiabilité des informations publiées sur Internet (Kumar, et al., 2020). Beaucoup de fausses nouvelles sont souvent liées à des événements ou à des crises qui se sont produits récemment et qui n'ont pas été vérifiés (Ammara , et al., 2019). Cependant, les fausses nouvelles disparaissent souvent d'Internet, y compris des réseaux sociaux, après un certain temps, mais ces fausses nouvelles peuvent avoir laissé un impact profondément négatif (Kapusta, et al., 2020). Retenir des fausses nouvelles ou des rumeurs sans les réfuter ou révéler la vérité peut avoir l'effet inverse de rendre les gens confus. Cela peut les inciter à spéculer ou à exagérer les nouvelles (Shrivastava, et al., 2020). Depuis que les fausses nouvelles sont devenues un défi mondial et une menace majeure pour la démocratie, l'économie et la coexistence pacifique (Zhou, et al., 2020a), les organisations non gouvernementales, les organisations de la société civile, les journalistes, les politiciens et les chercheurs s'efforcent de réduire les risques (Ammara , et al., 2019). En conséquence, des entreprises telles que Facebook, Twitter et Google ont accordé une attention particulière à la lutte contre la propagation des fausses nouvelles. Ces entreprises ont mené de nombreuses recherches sur cet aspect (Kumar, et al., 2020).

3. Les fausses nouvelles dans les médias sociaux

De nos jours, beaucoup de gens passent leur temps sur les réseaux sociaux pour se connecter, obtenir des informations et des nouvelles, et les partager, plutôt que de regarder les médias traditionnels (Machová, et al., 2022). La principale raison d'utiliser les médias sociaux par les utilisateurs du monde entier est d'obtenir des nouvelles et de suivre l'actualité. Cela est basé sur le rapport Digital 2021 Global Digital Overview, comme le montre la figure 1-1 (Hamed, et al., 2023).

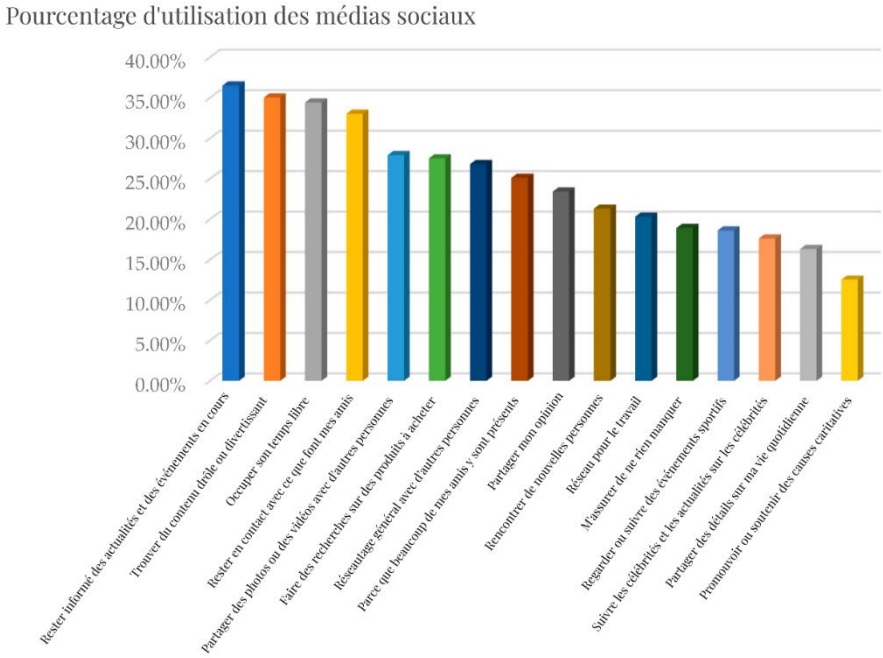

Figure 1-1 : Les raisons d'utiliser les médias sociaux selon le rapport Digital 2021 Global Digital Overview (Hamed, et al., 2023)

Beaucoup de gens utilisent les réseaux sociaux pour publier des nouvelles et des informations par le biais de leurs comptes ou pages parce que la publication d'informations sur ces plateformes diffère des médias traditionnels en ce sens que cela ne prend pas longtemps, n'a pas de coût et n'est pas soumis à des restrictions de vérification (Ammara , et al., 2019). La nature de la structure des plateformes de médias sociaux permet la diffusion des nouvelles en temps réel et rapidement, quelle que soit la crédibilité de ces nouvelles (Bahad, et al., 2019).

Le site Statista a présenté une statistique (Misinformation and who has the responsibility to stop it U.S. by politics 2022 | Statista) le 27 août 2019 basée sur une enquête menée aux États-Unis en 2018 sur la façon dont les réseaux sociaux sont responsables de la propagation de fausses nouvelles, et la conclusion de cette enquête est que 29% des participants ont déclaré que les médias sociaux sont principalement responsables de la diffusion de fausses nouvelles, tandis que 60% d'entre eux ont indiqué que ces plateformes sont en partie responsables de la diffusion de fausses nouvelles, comme le montre la figure 1-2 suivante (Hamed, et al., 2023).

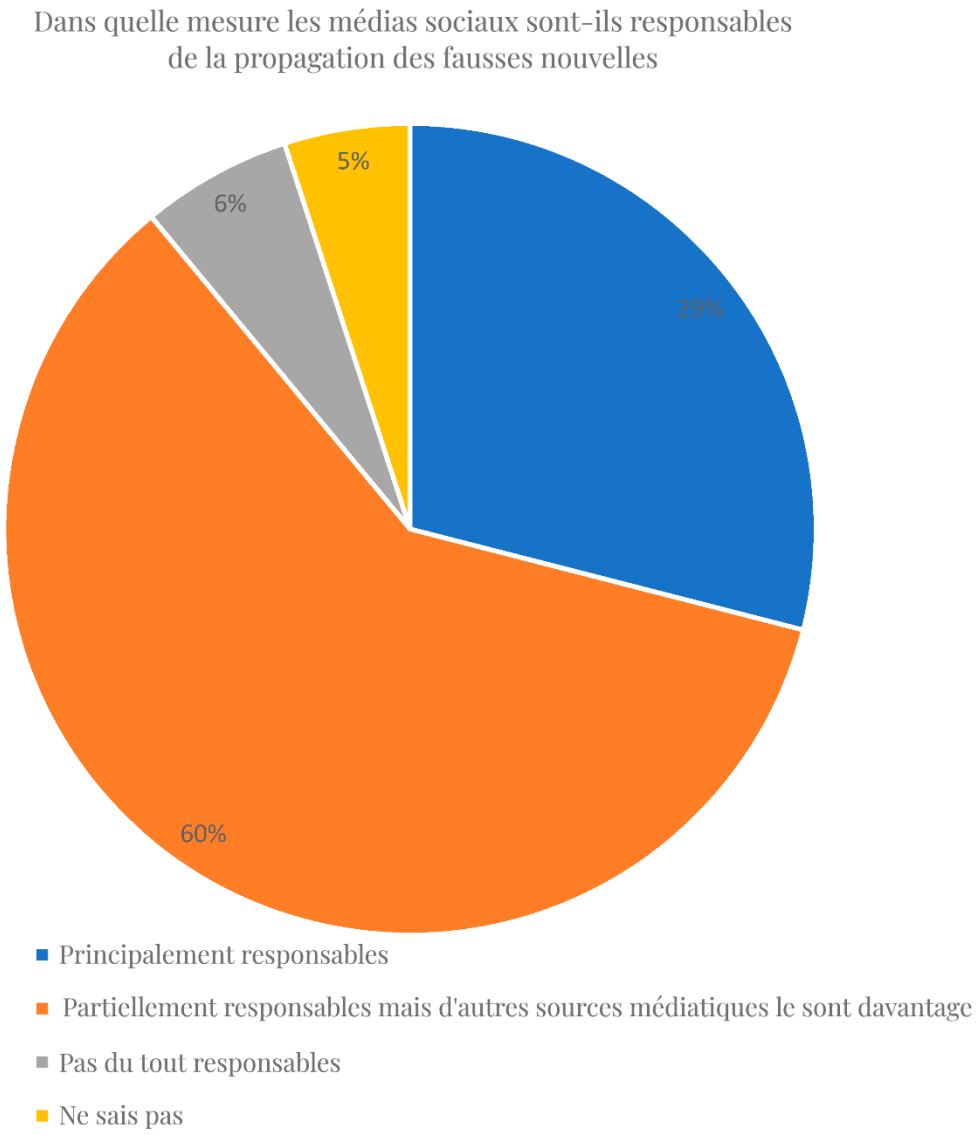

Figure 1-2 : Le rôle des médias sociaux dans la diffusion de fausses nouvelles (Hamed, et al., 2023)

Cependant, malgré toutes ces caractéristiques positives des réseaux sociaux, il y a des aspects négatifs à ces plateformes, qui sont exploitées par des individus ou des groupes pour diffuser de la désinformation et des fausses nouvelles à des fins malveillantes, ce qui peut être à des fins financières, répandre la haine fondée sur des motifs extrémistes, manipuler l'esprit des gens pour des raisons politiques ou créer des opinions biaisées pour des raisons électorales (Al-Makhadmeh & Tolba , 2020). Ces aspects négatifs des réseaux sociaux, représentés par la diffusion de fausses nouvelles, annoncent un grave danger qui affecte négativement la société et la vie des citoyens. Cela nécessite l'existence de modèles qui détectent les fausses nouvelles et limitent leur propagation (Lin & Chen, 2020).

4. Les méthodes de détection des fausses nouvelles

L'enquête de Zhou et Zafarani (Zhou & Zafarani, 2018), a divisé les méthodes de détection des fausses nouvelles du point de vue des caractéristiques. Notez que les données étiquetées limitées attirent de plus en plus de chercheurs pour lutter contre les fausses nouvelles avec peu ou pas de données étiquetées (Zhou & Zafarani, 2020). D'après la littérature, les méthodes de détection des fausses nouvelles sont catégorisées en approches supervisées, faiblement supervisées et non supervisées. La figure 1-3 montre la taxonomie de classification des méthodes de détection des fausses nouvelles basée sur l'apprentissage profond (Deep Learning : DL). En outre, il convient de noter que les méthodes disparates le long de ces trois dimensions mentionnées ci-dessus se concentrent sur des caractéristiques différentes, sur lesquelles nous reviendrons dans les sous-sections suivantes.

- *Méthodes supervisées :* Les méthodes supervisées apprennent avec des données étiquetées. La principale préoccupation est la façon dont les modèles DL utilisent ces riches informations sur les fonctionnalités. Par conséquent, ces méthodes peuvent être classées en : contenu d'actualités, contexte social et connaissances externes.
- *Méthodes faiblement/non supervisées :* Les méthodes faiblement supervisées supposent que seules des données étiquetées limitées sont disponibles au stade de l'apprentissage. La solution commune consiste à dériver des informations de supervision faibles de l'information disponible. Selon la façon d'obtenir une supervision faible, les méthodes semi-supervisées sont classées en une supervision faible du contenu et une supervision sociale faible. Les méthodes non supervisées apprennent avec des données totalement non étiquetées. Certains chercheurs recourent à des méthodes génératives ou utilisent des connaissances probabilistes.

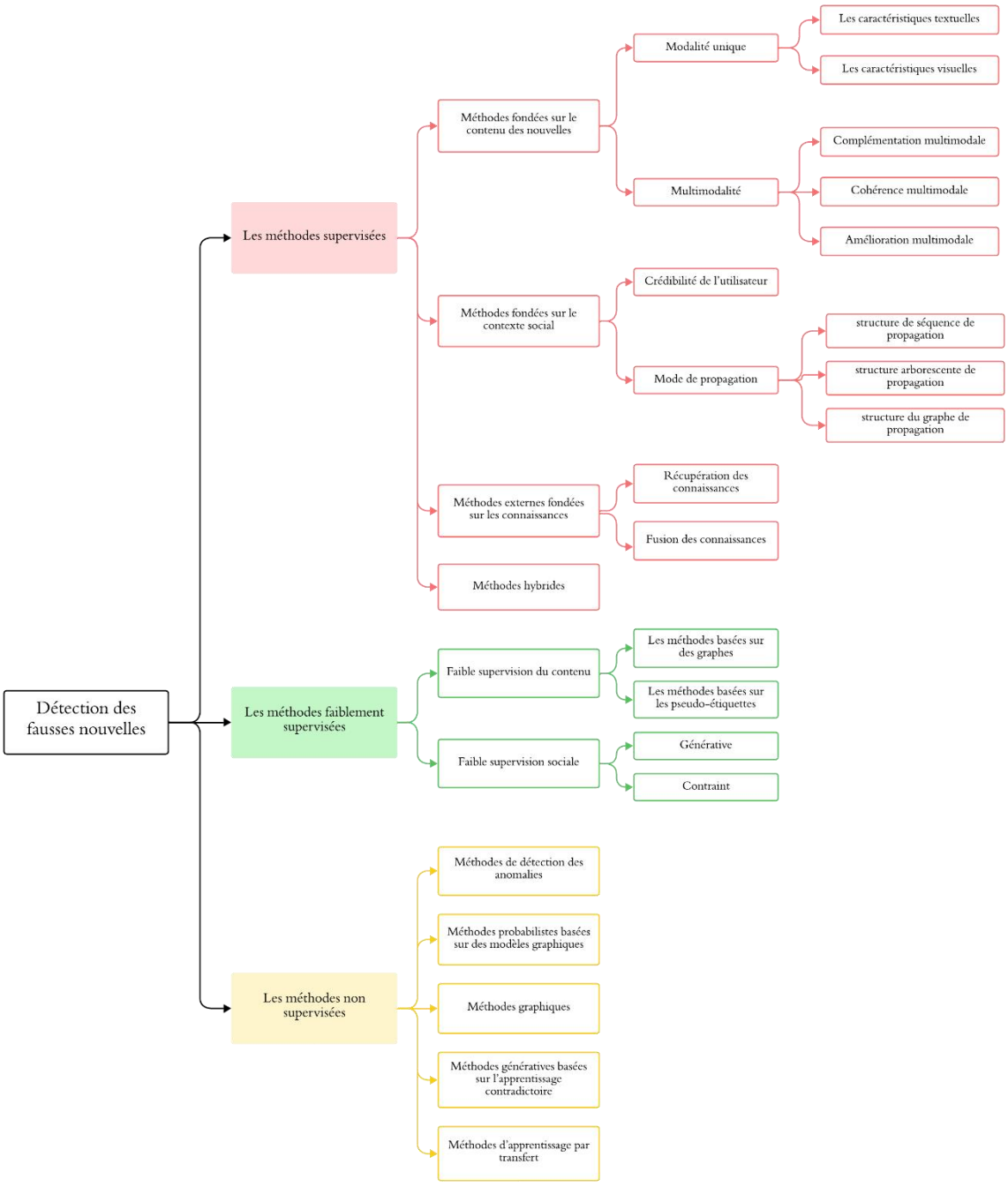

Figure 1-3 : Taxonomie des méthodes de détection des fausses nouvelles

4.1.Les méthodes supervisées

Les méthodes supervisées ont tendance à apprendre de diverses fonctionnalités avec des exemples étiquetés. Afin de réaliser de meilleures performances, des quantités de techniques telles que l'introduction d'informations multimodales, de connaissances externes et de stratégie d'intégration ont été explorées. Selon les caractéristiques utilisées, nous les divisons davantage comme suit.

4.1.1. Méthodes fondées sur le contenu des nouvelles

Le contenu des nouvelles désigne les informations explicites que les nouvelles contiennent à l'origine, telles que le texte des articles ou des images qui y sont attachés. Généralement, les méthodes basées sur le contenu des nouvelles utilisent ces caractéristiques textuelles / visuelles.

4.1.1.1. Modalité unique

La modalité unique désigne l'utilisation exclusive de caractéristiques provenant d'un seul mode d'information, comme le texte ou l'image, afin de différencier les nouvelles authentiques des fausses.

Les caractéristiques textuelles : peuvent être classées comme des caractéristiques génériques et des caractéristiques latentes. Les premiers sont souvent utilisés dans un cadre de l'apprentissage automatique (Machine Learning : ML) traditionnel, qui décrit le contenu textuel à partir de niveaux linguistiques : lexique, syntaxe, discours, sémantique, etc. Les caractéristiques textuelles latentes se réfèrent à l'intégration de texte de nouvelles. L'incorporation de texte peut être dérivée au niveau des mots, des phrases et des documents. En cela, un article de presse peut être représenté par des vecteurs latents, qui peuvent être utilisés immédiatement comme entrée pour les classificateurs (comme les Support Vector Machines : SVMs) ou ensuite intégrés dans des structures de réseau de neurones. Les réseaux de neurones récurrents (Recurrent Neural Network : RNN) sont hautement capables de modéliser des données séquentielles. En outre, afin de détecter les fausses nouvelles le plus tôt possible, certains chercheurs ont supposé que l'information multiforme ne soit pas disponible avant qu'un article de presse ne soit déjà devenu populaire. Ces œuvres utilisaient des fonctionnalités textuelles uniquement. Par exemple, Giachanou et al. (Giachanou, et al., 2019) ont examiné le rôle des signaux émotionnels et ont proposé un modèle de réseau de neurones à mémoire à long terme (Long Short-Term Memory : LSTM) qui intègre des signaux émotionnels obtenus à partir du texte des revendications afin de distinguer les vraies des fausses nouvelles.

Caractéristiques visuelles : Avec le développement du multimédia, l'actualité des réseaux sociaux contient non seulement des informations textuelles, mais aussi des images, des vidéos et d'autres informations visuelles qui impliquent une sémantique riche. En raison de l'hétérogénéité entre l'information textuelle et visuelle, il est difficile pour les approches textuelles basées sur des caractéristiques de capturer l'information visuelle. De nombreux chercheurs ont proposé d'utiliser des caractéristiques visuelles pour détecter les fausses nouvelles. Pour mieux utiliser l'information pertinente à la tâche comme les caractéristiques intrinsèques des images de fausses nouvelles, Qi et al. (Qi, et al., 2019) ont proposé un cadre visuel neuronal multi-domaines qui combinait le domaine de fréquence et l'information visuelle du domaine pixel pour distinguer les nouvelles réelles des fausses par les caractéristiques visuelles. Les images de fausses nouvelles peuvent être de mauvaise qualité, apparaissant dans le domaine de fréquence. Le modèle proposé capture

automatiquement les caractéristiques de qualité d'image dans le domaine de fréquence en utilisant un réseau de neurones convolutif (Convolutional Neural Network : CNN) et extrait automatiquement les caractéristiques sémantiques d'image dans le domaine de pixels en utilisant un CNN-RNN.

4.1.1.2. Multimodalité

Les caractéristiques textuelles et visuelles sont efficaces dans les tâches de détection de fausses nouvelles, respectivement. Les nouvelles dans les réseaux sociaux existants contiennent souvent des informations textuelles et visuelles. C'est une idée naturelle de les combiner pour une meilleure performance. Nous illustrerons des méthodes multimodales sur trois axes selon les différentes perspectives multimodales qu'elles adoptent pour faciliter la détection des fausses nouvelles (Fake News Detection : FND).

Complémentation multimodale : Certaines études considèrent l'information visuelle comme un complément aux textes de fausses nouvelles. Ils ont utilisé un encodeur de texte pour extraire des caractéristiques de texte et un encodeur visuel pour extraire des caractéristiques visuelles, et les ont simplement concaténées comme caractéristique des nouvelles. Dans ce cadre, Wang et al. (Wang, et al., 2021) ont proposé une méthode pour détecter les fausses nouvelles multimodales sur les événements émergents par le biais du processus méta neuronal. Avec le développement du modèle pré-entraîné actuel, Singhal et al. (Singhal, et al., 2019) ont d'abord introduit des modèles de langage pré-entraînés tels que (Pre-training of Deep Bidirectional Transformers : BERT) et XLnet pour encoder des fonctionnalités de texte, puis les ont complétés avec des fonctionnalités visuelles. Malgré le succès obtenu par ces travaux, ils ne tiennent pas compte des corrélations complexes entre les modes contenus dans les fake news, qui limitent l'efficacité de la détection de contenu multimodal.

Cohérence multimodale : Les images non pertinentes sont des caractéristiques des fausses nouvelles multimodales. Par conséquent, certains travaux ont prêté attention à la mesure de la cohérence multimodale dans la détection.
Zhou et al. (Zhou, et al., 2020b) ont utilisé le modèle de sous-titrage d'image pour transformer les images en phrases, puis ont évalué la similitude des phrases entre le texte original des nouvelles et les légendes d'image produites pour calculer l'incohérence multimodale. Cependant, en raison des disparités entre l'ensemble de données de formation du modèle de légende d'image et le corpus de nouvelles réel, la performance du modèle est limitée. Inspirés par l'excellente performance du transformateur dans la représentation visuelle, Ghorbanpour et al. (Ghorbanpour, et al., 2023) ont proposé une méthode (Fake News Revealer : FNR) qui utilisait un transformateur de vision (Dosovitskiy, et al., 2020) et BERT (Devlin, et al., 2019) pour extraire les fonctions d'image et de texte séparément. Et puis, FNR a utilisé la perte contrastive pour déterminer la similitude d'image et de texte.

Amélioration multimodale : Plutôt que de modéliser directement les images de manière grossière lors de la fusion d'informations, le texte et les images sont liés dans une sémantique de haut niveau, dont les parties alignées indiquent généralement les caractéristiques importantes des nouvelles. Par conséquent, certains travaux se concentrent sur l'extraction de fonctionnalités dans les images et les textes et sur l'amélioration mutuelle pour mieux détecter les fausses nouvelles. Qian et al. (Qian, et al., 2021a) ont proposé un réseau (Hierarchical Multi-modal Contextual Attention Network : H-MCAN) pour modéliser conjointement l'information contextuelle multimodale et la sémantique hiérarchique du texte dans un cadre unifié et profond pour la détection des fausses nouvelles. Le contexte multimodal de chaque nouvelle est modélisé par le réseau d'attention contextuelle multimodal. Wu et al. (Wu, et al., 2021) ont proposé MCAN, qui a extrait les caractéristiques du domaine spatial et du domaine de fréquence de l'image et les caractéristiques textuelles du texte. MCAN a également développé une nouvelle approche de fusion avec de multiples couches de co-attention pour apprendre les relations d'intermodalité, qui fusionnait d'abord les caractéristiques visuelles, puis les caractéristiques textuelles. La représentation fusionnée obtenue à partir de la dernière couche de co-attention a été utilisée pour la détection de fausses nouvelles.

Pour résumer, il y a trois biais inductifs précieux lorsque l'on considère les corrélations texte-image dans la tâche multimodale de détection de fausses nouvelles : Les images apportent des informations supplémentaires au texte original, ce qui nécessite un complément multimodal. Le texte et les images avec des éléments incohérents sont un signal possible pour la détection multimodale de fausses nouvelles. Le texte et les images s'enrichissent mutuellement en repérant les caractéristiques essentielles.

4.1.2. Méthodes fondées sur le contexte social

Dans certains cas, il n'est pas satisfaisant de repérer les fausses nouvelles simplement à partir du contenu des nouvelles, car les fausses nouvelles ont été écrites intentionnellement pour confondre le public. Les caractéristiques de réseautage et d'interconnexion des plateformes de médias sociaux fournissent des renseignements supplémentaires, notamment les caractéristiques du contexte social. Il représente les engagements des utilisateurs et les comportements sociaux des utilisateurs sur les médias sociaux. Les méthodes peuvent être séparées en catégories fondées sur la crédibilité et sur la propagation.

4.1.2.1. Crédibilité de l'utilisateur

Les techniques basées sur la crédibilité visent à utiliser la fiabilité des utilisateurs pour aider à identifier les fausses nouvelles. Normalement, les informations de crédibilité peuvent être collectées soit à partir de la description explicite de l'utilisateur, soit en analysant les relations entre les articles de presse et d'autres composants tels que les utilisateurs, les éditeurs et les publications.

Jiang et al. (Jiang, et al., 2019) ont utilisé l'apprentissage de représentations de réseau attribuées pour explorer les corrélations possibles entre les utilisateurs dans le réseau d'amitié en se basant sur les attributs des utilisateurs et reconstruire le réseau utilisateur-nouvelles, afin d'améliorer les représentations des nouvelles et des utilisateurs dans le réseau de propagation des nouvelles. L'émotion joue également un rôle essentiel dans la détection des fausses nouvelles en ligne. (Zhang, et al., 2021) ont proposé BERT-EMO pour représenter l'émotion de l'éditeur et l'émotion sociale et ont également considéré la relation entre l'émotion de l'éditeur et l'émotion sociale (double émotion) pour exposer les signaux émotionnels distinctifs pour détecter les fausses nouvelles.

Bien que les modèles puissent obtenir de bonnes représentations des utilisateurs à l'aide de réseaux de neurones profonds, les limites se manifestent principalement dans les questions de confidentialité. En d'autres termes, de nombreux utilisateurs ne sont pas disposés à montrer des informations personnelles exactes, ce qui conduit à l'introduction d'informations bruyantes.

4.1.2.2. Mode de propagation

Les fausses nouvelles se propagent souvent différemment des vraies nouvelles sur les réseaux sociaux. Les méthodes basées sur la propagation visent à modéliser le chemin de propagation des nouvelles et à analyser la différence entre la diffusion de nouvelles réelles et la diffusion de fausses nouvelles pour identifier les fausses nouvelles.

Certaines approches récentes utilisent des caractéristiques temporelles et linguistiques extraites d'une séquence de commentaires d'utilisateurs pour détecter les fausses nouvelles. Ils modélisent le processus de diffusion des nouvelles comme une structure séquentielle.

Bien que la modélisation de structure de séquence obtienne de bons résultats, la structure de séquence a du mal à capturer la relation structurelle entre les retweets de nouvelles et les commentaires. Afin de mieux connaître les informations structurelles entre les nouvelles et ses séquences retweetées ou commentées. Certains chercheurs essaient de modéliser le processus de diffusion des nouvelles comme une structure arborescente. Inspirés par l'idée d'améliorer la puissance de représentation des objets structurés à l'aide d'un transformateur, Ma et Gao (Ma & Gao, 2020) ont proposé un modèle de détection de fausses nouvelles basé sur un transformateur d'arbre, qui utilise un mécanisme d'auto-attention pour modéliser les interactions sémantiques post-niveau à l'intérieur et entre les sous-arbres. Pour renforcer la robustesse du modèle, Ma et al (Ma, et al., 2021) ont proposé une approche de type réseau générateur antagoniste (Generative Adversarial Network : GAN), dans laquelle un générateur à transformateur a été conçu pour produire des voix incertaines ou contradictoires, polarisant davantage le fil de conversation original pour faire pression sur le discriminateur afin d'apprendre des

caractéristiques indicatives de rumeurs plus substantielles à partir des exemples augmentés et plus difficiles.

Toutefois, l'efficacité des méthodes citées ci-dessus est trop faible pour appréhender les caractéristiques de la propagation, et les caractéristiques de la structure générale de la dispersion des rumeurs sont ignorées. D'autres chercheurs ont tenté de représenter le processus de diffusion des nouvelles comme une structure graphique, convertissant ainsi le problème de détection des fausses nouvelles en un problème de classification de graphique. La structure de propagation des nouvelles est modélisée en utilisant des graphes homogènes et hétérogènes.

Les réseaux homogènes sont composés de nœuds et de bords d'un seul type (Zhou & Zafarani, 2019). Selon l'enquête Vosoughi et al (Vosoughi, et al., 2018), les fausses informations se répandent plus rapidement, plus loin et plus largement que les vraies informations. Grâce à la réussite du modèle de codeur automatique dans le domaine de l'apprentissage de l'information latente, Lin et al. (Lin, et al., 2020) ont proposé (AutoEncoder Graph-Convolutional Network : AE-GCN) et (Variational AutoEncoder- Graph Convolutional Network : VAE-GCN), qui ont utilisé (Graph AutoEncoder : GAE) et sa variante (Variational AutoEncoder- Graph Convolutional Network : VGAE) pour apprendre des informations sur la structure graphique sur la détection des rumeurs. Plutôt que de se concentrer sur les réseaux statiques et de supposer que toute la structure du réseau de propagation de l'information est accessible avant de mettre en œuvre des algorithmes d'apprentissage, Song et ses collègues (Song, et al., 2021) ont proposé un cadre de détection dynamique basé sur les graphes pour représenter le processus d'évolution temporelle des actualités dans le monde réel en tant qu'évolution graphique dans une perspective de temps continu.

Les techniques utilisant le graphe homogène ne peuvent représenter qu'un type particulier de nœud ou de bord, tandis qu'elles ne peuvent pas prendre en compte les informations provenant de plusieurs nœuds ou de relations multiples. Cependant, pour diffuser les nouvelles, il est souvent nécessaire d'utiliser divers types d'informations sur les nœuds, tels que les utilisateurs et les commentaires, ce qui nécessite la mise en place d'un réseau de neurones hétérogènes pour combiner les informations provenant de plusieurs nœuds ou arêtes.

Les réseaux hétérogènes sont constitués de plusieurs types de nœuds ou de bords. Les chercheurs Zhou et Zafarani (Zhou & Zafarani, 2019) et Huang et al. (Huang, et al., 2019) ont présenté des réseaux d'attention graphique hétérogène basé sur des méta-chemins pour encoder les relations sémantiques globales entre le comportement des utilisateurs. Afin d'accroître la robustesse du modèle, Yang et al. (Yang, et al., 2021) ont d'abord utilisé un cadre d'apprentissage contradictoire graphique particulier pour apprendre des caractéristiques de structure plus caractéristiques. L'étude de Nguyen et ses collègues (Nguyen, et al., 2020) a suggéré l'utilisation d'un réseau de neurones à graphes inductifs hétérogènes afin de modéliser efficacement l'article, le retwitter et les utilisateurs. Les travaux

mentionnés ci-dessus ignorent les informations locales et ne prennent en compte que les informations globales du graphe de propagation. Cependant, l'information locale d'un graphe de propagation implique généralement une information sémantique riche. Yuan et al. (Yuan, et al., 2019) ont créé un graphique varié en utilisant des tweets sources, des retweets et des utilisateurs. Par la suite, ils ont suggéré l'utilisation d'un réseau (Graph-based Linear Assignment Network : GLAN) qui peut codage en même temps des informations sémantiques locales et générales. Dans un premier temps, ils ont employé un mécanisme d'attention afin de combiner les retweets liés aux nouvelles avec les nouvelles originales afin d'obtenir une représentation des nouvelles locales. Ensuite, ils ont représenté les relations mondiales entre tous les tweets sources, les retweets et les utilisateurs afin d'obtenir une représentation des nouvelles mondiales.

En résumé, les algorithmes basés sur le contexte social utilisent les données de profil utilisateur et les données de propagation afin de repérer les fausses nouvelles. Ils ont révélé des résultats satisfaisants lorsque les informations sont adéquates. Toutefois, la protection de la vie privée rend difficile l'obtention des informations personnelles des utilisateurs. De plus, on ne peut pas obtenir des données exhaustives sur la propagation au début de la diffusion des nouvelles. Ainsi, du point de vue de la détection précoce, les approches basées sur le contexte social ne sont pas suffisamment adaptées pour être mises en œuvre dans la réalité.

4.1.3. Méthodes externes fondées sur les connaissances

La plupart des méthodes ci-dessus reposent fortement sur les caractéristiques linguistiques et sémantiques du contenu des nouvelles ou des caractéristiques du contexte social. Cependant, ils ne parviennent pas à exploiter efficacement les connaissances externes, ce qui pourrait aider à déterminer si le document de nouvelles est fiable. Le contenu des nouvelles est très condensé et comprend de nombreuses mentions d'entités. En raison de problèmes avec les alias, les abréviations et les orthographes alternatives, il n'est pas toujours possible de comprendre directement le contenu des nouvelles. Ainsi, plusieurs études introduisent des connaissances externes pour aider à améliorer la performance de la détection des fausses nouvelles, dont l'efficacité a également été analysée dans Ahmed et al. (Ahmed, et al., 2019b).

Une source typique de connaissances préalables est les Graphes de Connaissances (knowledge Graphs : KGs), qui décrivent les entités et leurs relations sous forme de graphique. En particulier, les KGs incluent des informations obtenues dans de nombreux domaines.

Récupération des connaissances : Le processus de récupération des connaissances consiste à trouver un ensemble de concepts connexes C, compte tenu d'un texte de poste. Il est composé de deux étapes « la liaison d'entités et la conceptualisation d'entités ».

Fusion des connaissances : Après avoir obtenu les concepts de connaissances connexes dans le KG, les modèles doivent fusionner les informations du texte d'actualité et les

concepts de connaissances de base pour obtenir la représentation de chaque message et détecter les fausses nouvelles.

Les méthodes existantes pour fusionner l'information des textes d'actualité et les connaissances externes peuvent être largement divisées en deux volets.

Les approches basées sur l'attention : fusionnent l'information du texte et les concepts de connaissances externes par le mécanisme d'attention. Dun et al. (Dun, et al., 2021) se sont efforcés d'intégrer la connaissance des entités et des contextes d'entités à partir du graphique des connaissances pour détecter les fausses nouvelles. Premièrement, ils ont identifié les entités mentionnées dans le contenu des nouvelles et les ont alignées sur celles du graphique des connaissances. Ensuite, les entités et leurs contextes ont été utilisés comme une connaissance externe pour fournir des informations supplémentaires. Enfin, ils ont conçu un mécanisme d'attention conscient de la connaissance pour mesurer l'importance de la connaissance.

Les approches basées sur les graphes : utilisent la structure de topologie des graphes pour la fusion d'informations. Plus précisément, ces méthodes construisent des entités de connaissances externes et des entités de texte en graphes hétérogènes pour chaque nouvelle et utilisent des méthodes de réseau de neurones pour la fusion d'informations. De cette façon, le problème de détection des fausses nouvelles est converti en classification graphique pour identifier les fausses nouvelles. Qian et al. (Qian, et al., 2021b) ont proposé un réseau (Knowledge-aware Multi-modal Adaptive Graph Convolutional Network : KMAGCN) qui modélise les messages sous forme de graphes pour obtenir les représentations sémantiques à long terme, et ont fusionné les textes, les concepts de connaissances et les images dans un cadre unifié.

4.1.4. Méthodes hybrides

La détection des fausses nouvelles repose sur le contenu des nouvelles, les informations de contexte social et les connaissances externes impliquées. C'est pourquoi de nombreuses études ont tenté d'utiliser la méthode hybride pour obtenir de meilleurs résultats.

Les utilisateurs expriment leurs opinions sur l'événement en utilisant des informations interactives telles que leurs likes et commentaires sur les articles de presse. Différentes recherches ont cherché à repérer les fausses informations en combinant le contenu des nouvelles avec les informations interactives des utilisateurs. Shu et al. (Shu, et al., 2019a) ont utilisé le réseau de co-attention pour combiner les informations de contenu des nouvelles avec les informations de commentaires des utilisateurs pour détecter les fausses nouvelles. Ils ont également cherché la première importance des phrases de contenu de nouvelles et des commentaires des utilisateurs pour fournir une interprétabilité pour la détection de fausses nouvelles.

Le rôle de la crédibilité des écrivains de nouvelles est également crucial pour détecter les fausses informations. Une partie des recherches a essayé de compromettre la crédibilité des utilisateurs en se basant sur des profils d'utilisateurs et des relations sociales. De plus, ils combinent les données issues de la crédibilité des utilisateurs et du contenu des nouvelles afin de détecter les fausses nouvelles. Dou et al. (Dou, et al., 2021) ont étudié l'impact des préférences des utilisateurs sur la détection des fausses nouvelles et ont proposé une méthode appelée (Détection des fausses nouvelles basée sur les préférences des utilisateurs : UPFD). L'UPFD a utilisé les publications précédentes des éditeurs de nouvelles afin d'obtenir une représentation de leurs préférences et a combiné les préférences intrinsèques des éditeurs de nouvelles avec la propagation externe des nouvelles pour repérer les fausses nouvelles.

Les connaissances externes impliquées dans les nouvelles peuvent aider les lecteurs à comprendre les nouvelles. Certaines études ont tenté de détecter les fausses nouvelles en utilisant des connaissances externes sur les nouvelles, le contenu des nouvelles et le contexte social. Dans leur étude, Sun et ses collègues (Sun, et al., 2022) ont examiné non seulement les connaissances externes liées au contenu des nouvelles, mais également les connaissances externes liées aux commentaires des nouvelles. Les chercheurs ont développé un réseau de neurones convolutif à graphes dynamiques doubles (Dual Dynamic Graph Convolutional Network : DDGCN) afin de repérer les rumeurs sur les réseaux sociaux. L'utilisation des connaissances externes dans les nouvelles et les commentaires par DDGCN permet d'améliorer la compréhension du contenu et des commentaires. De plus, il modélise le processus de propagation des nouvelles en utilisant des réseaux de neurones de graphes dynamiques, ce qui a donné des résultats positifs.

Le tableau 1-1 présente une synthèse des approches supervisées utilisées pour la détection des fausses nouvelles.

Catégorie	Sous-catégorie	Description	Exemples/Références
Méthodes fondées sur le contenu	Modalité unique	Utilisation des caractéristiques textuelles ou visuelles pour la détection des fausses nouvelles.	Textuelles : SVM, RNN et LSTM proposé par (Giachanou, et al., 2019) qui intègre des signaux émotionnels pour détecter les fausses nouvelles. Visuelles : CNN, CNN-RNN pour capturer les caractéristiques sémantiques et de qualité des images proposé par (Qi, et al., 2019).

	Multimodalité	Combinaison de textes et d'images pour une meilleure détection. Les approches sont : complémentation, cohérence, amélioration multimodale.	Complémentation : concaténation des caractéristiques ex : BERT, XLNet proposé par (Singhal, et al., 2019). Cohérence : mesure des incohérences ex. : sous-titrage d'image proposé par (Zhou, et al., 2020b). Amélioration : H-MCAN proposé par (Qian, et al., 2021a), MCAN proposé par (Wu, et al., 2021) pour modéliser des liens texte-image.
Méthodes fondées sur le contexte social	Crédibilité de l'utilisateur	Modélise les interactions des utilisateurs et leur fiabilité via des réseaux attribués ou des relations émotionnelles.	Analyse des émotions ex : BERT-EMO proposé par (Zhang, et al., 2021).
	Mode de propagation	Modélisation de la diffusion des nouvelles pour différencier vraies/fausses nouvelles via des structures séquentielles, ou graphiques.	Transformateur d'arbre proposé par (Ma & Gao, 2020), GAN proposé par (Ma, et al., 2021), les graphes homogènes comme AE-GCN et VAE-GCN proposé par (Lin, et al., 2020), les graphes hétérogènes comme GLAN proposé par (Yuan, et al., 2019) pour modéliser les interactions et la structure de propagation.
Méthodes fondées sur les connaissances externes	Récupération des connaissances	Utilisation de graphes de connaissances pour enrichir la compréhension des nouvelles et combler les lacunes liées aux entités mentionnées.	KG : concepts liés aux entités via liaison et conceptualisation.
	Fusion des connaissances	Combine le texte des nouvelles et les concepts	Approches basées sur l'attention ex : mécanisme

		des KG pour obtenir une représentation enrichie.	d'attention conscient des connaissances proposé par (Dun, et al., 2021). Approches basées sur les graphes ex : KMAGCN proposé par (Qian, et al., 2021b).
Méthodes hybrides	Combinaison des approches	Intègre le contenu des nouvelles, les données sociales (ex : likes, commentaires, et les connaissances externes pour une détection optimale.	Réseaux de co-attention proposé par (Shu, et al., 2019a). Analyse des préférences des utilisateurs ex : UPFD proposé par (Dou, et al., 2021).

Tableau 1-1 : Les méthodes supervisées pour la détection des Fausses Nouvelles

4.2.Les méthodes faiblement supervisées

L'apprentissage faiblement supervisé est une solution prometteuse pour utiliser des modèles d'apprentissage profond pour la détection des fausses nouvelles, qui apprend de l'expérience ne contenant que de la supervision faible (comme des informations supervisées incomplètes, inexactes ou bruyantes). Traditionnellement, selon que l'oracle ou l'intervention humaine (expert en la matière) est utilisé, cela peut être classé en (Hu, et al., 2022):

Apprentissage semi-supervisé : Il fait référence à l'apprentissage à partir d'un petit nombre d'échantillons étiquetés et (généralement un grand nombre d'échantillons non étiquetés).

Apprentissage actif : Il sélectionne des données informatives non étiquetées pour interroger un oracle pour la sortie y.

À notre connaissance, la plupart des travaux existants dans FND choisissent la méthode semi-supervisée. Selon la source des données des médias sociaux pour dériver une supervision faible, nous classons la méthode semi-supervisée comme une supervision de contenu faible et une supervision sociale faible.

4.2.1. Faible supervision du contenu

Les techniques de supervision de contenu faible se basent sur des données de texte d'actualité partiellement étiquetées en entrée, en utilisant les informations linguistiques des articles d'actualité étiquetés, tout en explorant les modèles dissimulés dans les données non étiquetées. Il est possible de classer les méthodes de supervision de contenu faible en deux catégories principales : les méthodes basées sur des graphes et les méthodes basées sur des étiquettes.

Les méthodes basées sur des graphes : exploitent la similitude entre les textes de nouvelles afin de représenter les nouvelles étiquetées et les nouvelles non étiquetées en un seul graphe. La contrainte fondamentale réside dans le fait que les nœuds du graphique qui sont proches les uns des autres ont souvent le même nom. La détection de fausses nouvelles semi-supervisée a été transformée en un problème de classification de nœud, et les problèmes de données étiquetées rares sont résolus en diffusant des étiquettes connues sur un graphique afin de déterminer des étiquettes inconnues. Hu et al. (Hu, et al., 2019) ont proposé d'apprendre les représentations des nœuds d'actualités via l'intégration de graphes et ont utilisé des blocs GCN multi-profondeur pour capturer des informations de voisins multi-échelles combinées par un mécanisme d'attention. Meel et Vishwakarma (Meel & Vishwakarma, 2021) ont mis au point une technique semi-supervisée de détection des fausses nouvelles basée sur un réseau convolutif de graphes. L'architecture recommandée comprend trois composantes essentielles : la collecte des données d'intégration de mots à partir des articles de presse dans des ensembles de données utilisant GloVe, la construction d'un graphe de similarité en utilisant (Word Mover's Distance : WMD), et l'application de GCN pour la classification binaire des articles de presse dans un paradigme semi-supervisé.

Les méthodes basées sur les pseudo-étiquettes : utilisent les données étiquetées pour former un modèle supervisé, puis appliquent le modèle supervisé pour étiqueter les données non étiquetées avec des pseudo-étiquettes, et enfin utilisent les étiquettes réelles et les pseudo-étiquettes pour former le modèle non supervisé ou recycler le modèle supervisé. Dong et al. (Dong, et al., 2019) ont proposé un modèle appelé réseaux de neurones à deux chemins convolutifs, dans lequel un chemin adopte l'apprentissage supervisé, et l'autre chemin adopte l'apprentissage non supervisé pour FND. Plus précisément, la méthode supervisée a besoin de données étiquetées pour entraîner par perte d'entropie croisée. Le modèle supervisé entraîné fournit des pseudo-étiquettes pour les données non étiquetées, puis le chemin non supervisé est entraîné en utilisant la perte d'erreur quadratique moyenne entre le pseudo-label du chemin supervisé et le résultat prévu du chemin non supervisé. Ensuite, la somme pondérée des deux pertes est utilisée comme perte totale pour former conjointement l'architecture globale du réseau. Li et al. (Li, et al., 2021) ont proposé un mécanisme d'auto-apprentissage pour effectuer la détection semi-supervisée des fausses nouvelles. Ils ont proposé une couche de réseau de confiance pour évaluer la confiance des pseudo-étiquettes et ont retracé le modèle en utilisant des pseudo-étiquettes avec une grande confiance et les données originales de l'étiquette.

Toutes les techniques mentionnées précédemment utilisent un modèle entraîné afin de générer des pseudo-étiquettes pour les données non étiquetées, et utilisent les étiquettes réelles et les pseudo-étiquettes pour créer un autre modèle ou recycler le modèle. Toutefois, ils ne sont pas conscients que les données puissent être classées à partir de différentes perspectives, ce qui permet de former différents classificateurs à partir de différentes perspectives.

4.2.2. Faible supervision sociale

Par rapport au contenu des nouvelles, les informations de contexte social ont des propriétés uniques qui les rendent appropriées pour dériver une supervision faible. D'une part, l'information véhiculée par le contenu des nouvelles est limitée, tandis que la quantité d'information produite par les utilisateurs (par exemple, les commentaires, les comportements) n'est pas proscrite, elle fournit donc des sources abondantes pour obtenir des fonctionnalités utiles. D'autre part, lorsqu'il n'y a pas de données étiquetées explicites pour optimiser les modèles, l'information sur le contexte social (p. ex., les attitudes des utilisateurs à l'égard des nouvelles ainsi que leur crédibilité) devient une référence auxiliaire essentielle pour juger de la véracité des allégations de nouvelles (Shu et al. (Shu, et al., 2020a)).

Utilisateurs : Les renseignements personnels et les caractéristiques des utilisateurs pourraient être saisis pour FND, comme l'âge d'inscription, le nombre d'abonnés ou de partisans et le nombre de tweets publiés par les utilisateurs. La contrainte injectée d'une supervision faible est que les auteurs de fausses et vraies nouvelles peuvent créer des groupes distincts avec des traits distincts que les signaux de profil utilisateur pourraient illustrer.

Les publications provenant des commentaires ou des positions des utilisateurs sont également essentielles pour déduire l'authenticité des nouvelles. La contrainte sous-jacente injectée est que la crédibilité d'une nouvelle est fortement corrélée à la crédibilité des nouvelles connexes précédentes publiées par cet éditeur de nouvelles.

Réseau : Il existe différents types de réseaux (tels que les réseaux d'amitié, les réseaux de diffusion et les réseaux d'interaction), et chacun reflète certaines caractéristiques qui distinguent les fausses nouvelles des vraies (Shu & Liu, 2019). Par exemple, la diffusion des nouvelles implique la participation en temps réel d'un grand nombre d'utilisateurs sur les médias sociaux. Les fausses nouvelles pourraient apparaître rapidement, tandis que les vraies nouvelles montrent une tendance constante (Shu & Liu, 2019). Les réseaux d'interaction montrent les liens entre diverses entités, telles que les éditeurs, les articles de presse et les lecteurs, qui pourraient être utilisés pour extraire les caractéristiques du réseau des entités pertinentes et prédire la crédibilité de l'information en fonction de leur association.

Il y a deux manières d'utiliser une faible supervision sociale :

Générative : cela signifie générer des étiquettes faibles, puis apprendre directement avec des étiquettes faibles.

Contraint : cela signifie représenter une supervision faible basée sur les données des médias sociaux comme des contraintes.

Le travail suivant (Shu, et al., 2019b) est un travail typique pour exploiter l'information du contexte social comme contraintes. Les règles pour dériver les contraintes sont les suivantes : Premièrement, les utilisateurs liés sont plus susceptibles de partager des articles de presse similaires. Deuxièmement, les éditeurs politiquement biaisés sont plus

susceptibles de créer de fausses histoires. Troisièmement, les utilisateurs peu fiables sont plus susceptibles de propager de fausses nouvelles. De cette façon, la distribution des étiquettes est estimée en injectant des contraintes dans le cadre d'intégration du réseau hétérogène pour l'apprentissage des représentations de l'actualité : d'une part, pour la relation de publication, la présentation des actualités doit prendre en considération l'orientation politique de l'éditeur ; d'autre part, pour la relation de propagation, en contraignant la présentation des actualités et la représentation de l'utilisateur à être proches l'une de l'autre si l'actualité est fausse et l'utilisateur est peu crédible.

En outre, certains travaux Konkobo et al. (Konkobo, et al., 2020), Yuan et al. (Yuan, et al., 2020) se concentrent sur le problème de la détection précoce basée sur le cadre de la faible supervision sociale. Plus précisément, Konkobo et al. (Konkobo, et al., 2020) ont d'abord construit un modèle pour extraire les opinions des utilisateurs exprimées dans les commentaires, puis ils ont utilisé l'algorithme CredRank pour évaluer la crédibilité des utilisateurs et ont construit un petit réseau d'utilisateurs impliqués dans la diffusion de nouvelles données. Yuan et al. (Yuan, et al., 2020) ont construit un réseau d'information hétérogène composé d'éditeurs de nouvelles, de contenus d'actualités et d'utilisateurs de nouvelles, et ont présenté une approche de détection précoce des fausses nouvelles (Multi-head Structure-aware Attention Network : SMAN). SMAN traite la crédibilité des créateurs et des lecteurs de nouvelles comme des signaux faiblement supervisés. Ils ont utilisé la crédibilité des utilisateurs ainsi que des informations textuelles pour détecter rapidement les fausses nouvelles.

4.3.Les méthodes non supervisées

À la recherche d'une alternative aux méthodes supervisées, de plus en plus de chercheurs envisagent de détecter les fausses nouvelles de manière non supervisée.

4.3.1. Méthodes de détection des anomalies

D'après la psychologie sociale et la dynamique de la communication sociale, les utilisateurs qui diffusent des rumeurs ont des comportements différents de ceux qui diffusent des faits authentiques. Pour tirer parti de ces disparités afin de faciliter la détection des rumeurs, certains chercheurs considèrent les messages de rumeurs comme des anomalies sur les réseaux sociaux et considèrent que la détection de fausses nouvelles est une détection d'anomalie.

La détection d'anomalies non supervisée est utilisée dans cette méthode, qui regroupe les publications régulières de l'historique des publications des utilisateurs et les projette dans l'espace d'intégration profond en utilisant des méthodes de codage non supervisées. Dans la phase de test, les modèles calculent la distance entre les données du jeu de test et les poteaux réguliers, et sélectionnent la valeur aberrante avec une distance plus considérable que la rumeur (Hu, et al., 2022).

4.3.2. Méthodes probabilistes basées sur des modèles graphiques

L'approche probabiliste basée sur des graphes traite le problème de la détection des fausses nouvelles comme un problème probabiliste. Dans l'étude de Yang et al. (Yang, et al., 2019) la véracité des nouvelles et la crédibilité des utilisateurs ont été considérées comme des facteurs cachés. Selon eux, la véracité des données est plus liée à la crédibilité des utilisateurs. Dans un premier temps, ils ont extrait les données des commentaires des utilisateurs pour obtenir leur avis sur les nouvelles. Ensuite, ils ont élaboré un modèle graphique probabiliste bayésien afin de capturer l'ensemble du processus de génération de la vérité des nouvelles et des opinions des utilisateurs. Ils suggèrent en même temps une méthode d'échantillonnage de Gibbs effondré afin de résoudre le problème d'inférence.

4.3.3. Méthodes graphiques

Les approches basées sur des graphes non supervisés pour détecter les fausses nouvelles exploitent la propriété qui ferme les nœuds d'un réseau et ont tendance à avoir des étiquettes similaires. Gangireddy et al. (Gangireddy, et al., 2020) ont présenté une approche non supervisée basée sur des graphes pour détecter les fausses nouvelles appelée GTUT. Utilisant les nouvelles comme nœuds et les similitudes entre les nouvelles comme bords, ils ont développé un graphique des nouvelles. Il commence par l'identification d'un ensemble de graines d'articles faux et vrais en utilisant des observations de haut niveau sur le comportement inter-utilisateur dans la propagation de fausses nouvelles. GTUT utilise ensuite la similitude entre les nœuds de nouvelles et étend les étiquettes à toutes les nouvelles dans le graphique. En utilisant des informations textuelles provenant d'articles de presse et d'informations sociales, GTUT a réussi. GTUT est inefficace dans les situations où le contexte social n'est pas toujours présent.

4.3.4. Méthodes génératives basées sur l'apprentissage contradictoire

L'apprentissage antagoniste génératif (Generative Adversarial Learning : GAL) (Chen, et al., 2019) était adapté à la réalisation d'un apprentissage adaptatif pour des scénarios non supervisés. Le GAL contenait deux composants : un générateur et un discriminateur. Le premier générait des échantillons inconnus et le second déterminait si les échantillons générés étaient proches des échantillons réels. L'entraînement contradictoire entre eux devait produire des résultats optimaux. Guo et al. (Guo, et al., 2021) ont présenté un modèle de réseau génératif antagoniste (Graph-GAN) qui repose sur l'intégration de graphes. En premier lieu, il a créé des espaces à grain fin en utilisant un code sur le graphique. De plus, il a mis en place une interaction contradictoire continue entre un générateur et un discriminateur pour le décodage non supervisé, ce qui lui permet d'apprendre activement les règles des espaces de caractéristiques. La mise en place du schéma en deux étapes a non seulement réussi à détecter les rumeurs floues dans des situations non supervisées, mais a également amélioré la robustesse de la formation non supervisée.

4.3.5. Méthodes d'apprentissage par transfert

L'apprentissage par transfert vise à acquérir des connaissances généralisables à partir des données du domaine source avec une étiquetage riche, puis à les transférer à la tâche cible, ce qui contribue à la formation de la tâche cible. Les documents anglais sont souvent remplis de données annotées dans les tâches de détection de fausses nouvelles, tandis que d'autres langues mineures sont moins annotées. Le modèle est éduqué à partir de données en anglais afin d'acquérir des connaissances préalables généralisées, puis il est transféré à la tâche de repérer les fausses nouvelles dans d'autres langues. Selon les recherches de Du et al. (Du, et al., 2021) un cadre d'apprentissage profond nommé CrossFake a été développé pour former BERT sur les ensembles de données d'actualités en anglais étiquetés. Ce cadre permettrait de repérer la plupart des fausses nouvelles chinoises non étiquetées après la traduction. Tian et al. (Tian, et al., 2021) ont suggéré un cadre d'apprentissage par transfert interculturel sans prise de vue afin de créer un système de détection des fausses nouvelles qui ne requiert aucune annotation pour une langue nouvelle. Ce système peut être utilisé dans deux langues car il peut repérer des rumeurs en se basant sur un modèle unique. Tout d'abord, il a amélioré un modèle linguistique multilingue pré-entraîné (tel que le BERT multilingue) pour détecter les fausses nouvelles à l'aide de données annotées dans la langue source (telle que l'anglais). Il a ensuite utilisé ce modèle pour classer les fausses nouvelles dans une autre langue cible, comme l'anglais, et a créé une étiquette de rumeur « argent » dans cette langue cible. L'utilisation de ces étiquettes argentées permet d'affiner le modèle multilingue pour l'ajuster à la langue visée. Il est important de souligner que dans le cadre non supervisé où les étiquettes de vérité fondamentale ne sont pas accessibles, les modèles peuvent donc prendre en compte certaines informations implicites sous forme d'indices de jugement référentiels.

Le tableau 1-2 met en évidence les principales catégories, approches et exemples pour comprendre et appliquer les méthodes non supervisées dans la détection des fausses nouvelles.

Catégorie	Sous-catégorie	Description	Exemple / Méthode
Méthodes non supervisées	Méthodes de détection des anomalies	Identifie les rumeurs comme des anomalies en analysant les comportements divergents des utilisateurs sur les réseaux sociaux.	Utilisation de modèles non supervisés pour détecter les anomalies : (Hu, et al., 2022).
	Méthodes probabilistes basées sur des graphes	Modélisation des relations entre véracité des nouvelles et crédibilité des utilisateurs comme un problème probabiliste via des graphes bayésiens.	Modèle graphique bayésien et échantillonnage de Gibbs effondré : (Yang, et al., 2019).

	Méthodes graphiques	Exploite la propriété selon laquelle des nœuds proches dans un graphe partagent souvent des étiquettes similaires.	GTUT pour la propagation des étiquettes dans les graphes : (Gangireddy, et al., 2020).
	Méthodes génératives basées sur l'apprentissage contradictoire	Utilise un générateur et un discriminateur en interaction pour améliorer les performances de la détection non supervisée.	Graph-GAN intégrant des graphes pour des scénarios non supervisés : (Guo, et al., 2021).
	Méthodes d'apprentissage par transfert	Transfert des connaissances acquises sur un domaine source (souvent richement étiqueté) vers un domaine cible (souvent peu ou non étiqueté).	CrossFake pour détecter des fausses nouvelles dans plusieurs langues : (Du, et al., 2021). Apprentissage par transfert interculturel : (Tian, et al., 2021).

Tableau 1-2 : Les méthodes non supervisées pour la détection des Fausses Nouvelles

5. Conclusion

Ce chapitre met en évidence l'ampleur du problème des fausses nouvelles sur les réseaux sociaux et les défis associés à leur détection. Il est crucial de faire des avancées dans l'analyse des sentiments et la technologie de détection des fausses nouvelles afin de créer des outils efficaces pour combattre la propagation de la désinformation sur Internet.

Chapitre 2 : La détection des fausses nouvelles basée sur l'analyse des sentiments : Etat de l'art

1. Introduction

L'objectif de ce chapitre est d'explorer l'état actuellement de la technologie permettant de détecter les fausses nouvelles et ainsi un accent particulier y est mis sur l'analyse de sentiments. Plus précisément, ce chapitre commence avec une introduction à l'analyse des sentiments dans le traitement du langage naturel. Il s'agit d'une étape importante pour identifier les émotions exprimées dans le texte. Ensuite, ce chapitre présente les différentes approches et techniques d'analyse de sentiment dans le domaine du traitement du langage naturel. Enfin, ce chapitre examine les différentes méthodes de détection des fausses nouvelles utilisant l'analyse des sentiments. En mettant en évidence les approches les plus prometteuses et les tendances actuelles.

2. Analyse des sentiments dans le traitement du langage naturel

L'analyse de sentiment (Sentiment Analysis : SA) est l'une des sections appartenant au domaine du traitement automatique du langage naturel (Natural Language Processing : NLP) et est chargée de concevoir et d'appliquer des modèles, des techniques et des approches pour identifier si un texte traite d'informations objectives ou subjectives et, dans ce dernier cas, pour identifier si ces informations ont été exprimées de manière négative, neutre ou positive, ainsi que pour déterminer si elles sont fortes ou faibles. La méthode d'analyse de sentiment est utilisée dans de nombreux domaines, notamment dans les médias sociaux, comme la classification des opinions des utilisateurs sur les publications des médias sociaux, ou la connaissance des tendances de masse lors des élections et la prédiction des résultats finaux, en plus de contrôler l'opinion publique en comprenant les attitudes du public grâce à l'analyse des opinions des utilisateurs sur certaines situations. De plus, elle contribue au marketing commercial en explorant les désirs des consommateurs à l'égard des biens offerts sur les plateformes de médias sociaux. L'analyse de sentiment est également utilisée pour détecter les fausses nouvelles et elle est un facteur influent dans la détermination des informations trompeuses, en fournissant des informations cruciales sur leur contenu. Comme une partie significative du public des fausses nouvelles ne lit pas au-delà des titres, les éditeurs utilisent délibérément des combinaisons de polarité (positive et négative) ou de valence émotionnelle et d'excitation (faible et forte) pour tromper les lecteurs. Par conséquent, les titres doivent susciter la curiosité des lecteurs et les attirer émotionnellement afin d'augmenter la propagation des fausses nouvelles. Les résultats de la recherche menée par Paschen et Management montrent que les titres de fausses nouvelles sont significativement plus négatifs que les titres de vraies nouvelles. Cela indique que les titres sont un puissant différenciateur émotionnel entre les fausses et les vraies nouvelles. L'analyse fine des sentiments du contenu subjectif peut être abordée de manière positive grâce aux études dans ce domaine. L'analyse de sentiment opère généralement à un niveau plus général dans les efforts de recherche. Elle se concentre davantage sur l'identification de la subjectivité ou de la position sémantique d'un texte que sur l'identification d'une émotion particulière. Il est souvent important de prêter attention à la manière dont quelqu'un se sent après avoir été provoqué. Par exemple, même si la tristesse et la peur sont des émotions négatives, pouvoir les distinguer peut-être crucial. En cas de catastrophe, la peur peut être utilisée pour identifier le début de la tragédie, tandis que la tristesse peut être associée à ses phases finales. De nombreuses études antérieures ont porté sur le domaine de l'analyse des sentiments dans les médias sociaux, mais

peu d'études ont examiné l'analyse des émotions. L'analyse des émotions consiste à classer les données en fonction des sentiments qu'elles véhiculent, tels que la joie, la surprise, la colère, la peur, la tristesse et le dégoût. En général, un texte contient plusieurs mots particuliers (généralement trouvés dans des lexiques émotionnels) pour transmettre des émotions spécifiques. Par conséquent, les études utilisent des lexiques émotionnels annotés par des experts pour extraire des caractéristiques basées sur les émotions du texte. Il existe plusieurs lexiques émotionnels qui ont été développés de manière psychologiquement bien structurée, notamment ceux proposés par Magda Arnold, Robert Plutchik et Gerrod Parrot. La plupart des classifications d'émotions trouvées dans les études antérieures découlent des classifications du modèle de Plutchik, telles que celles illustrées dans la Figure 2-1. La question de la détection des fausses nouvelles a été étudiée récemment, et la grande majorité de ces études utilisent uniquement des caractéristiques basées sur le texte. Les fausses nouvelles ont pour objectif délibéré de susciter les émotions des lecteurs afin d'être cru et partagé sur les médias sociaux. C'est l'une de ses caractéristiques déterminantes. L'analyse émotionnelle joue un rôle clé dans la détermination du comportement de l'utilisateur vis-à-vis d'un sujet spécifique. Vosoughi, Roy a présenté une étude qui examinait les émotions liées aux rumeurs. Ils ont étudié la validité des rumeurs répandues sur Twitter et ont découvert que les fausses rumeurs provoquaient des réactions de peur, de dégoût et de surprise chez les gens. En revanche, les vraies rumeurs provoquaient des réactions de joie, de tristesse, de confiance et d'anticipation (Hamed, et al., 2023).

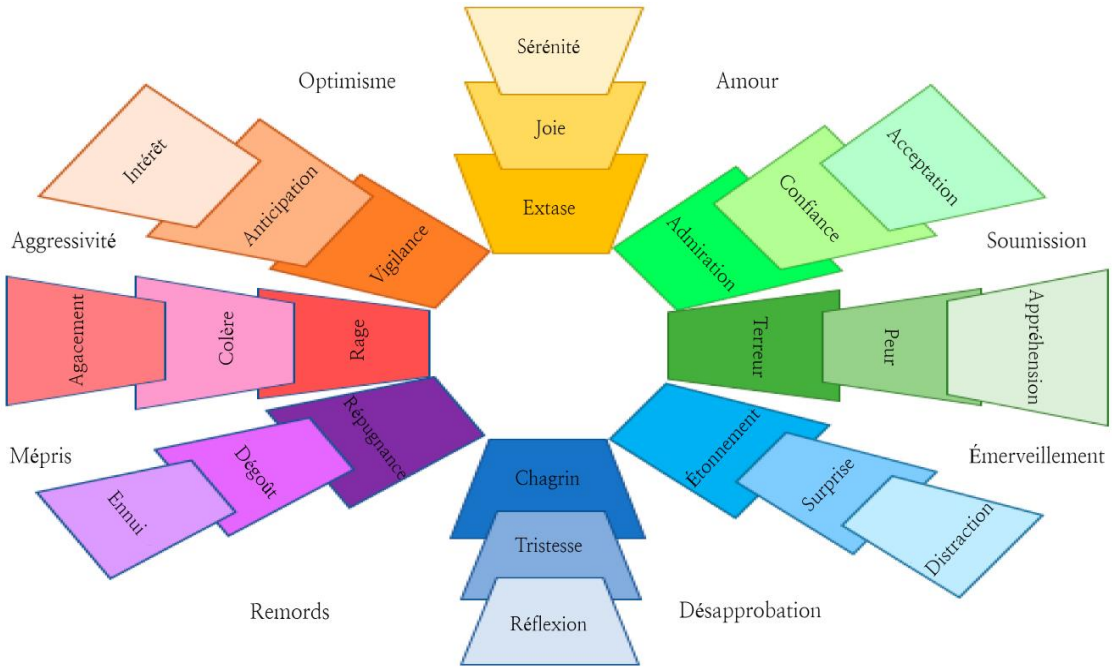

Figure 2-1 : La roue de l'émotion Plutchik (Hamed, et al., 2023)

En effectuant l'analyse de sentiment à l'aide de l'apprentissage profond ou d'autres techniques d'apprentissage automatique classique, les données textuelles d'entraînement doivent être nettoyées avant d'être utilisées pour induire le modèle de classification. Les titres des actualités présentent généralement des espaces blancs, des caractères de ponctuation, des caractères non alphabétiques ou encore des liens préfixés et des mots vides. Ces caractères sont généralement

retirés à l'aide des bibliothèques telles que BeautifulSoup car ils ne portent aucune information utile à l'analyse de sentiment. Ensuite, les titres sont séparés en mots individuels, ceux-ci sont ramenés à leur forme de base par lemmatisation, puis transformés en vecteurs numériques en utilisant l'incorporation de mots ou la fréquence de terme-fréquence inverse du document (TF-IDF : Term Frequency-Inverse Document Frequency).

L'incorporation de mots c'est la technique de modélisation de langage et d'apprentissage de caractéristiques, dans laquelle chaque mot est attribué à un vecteur de valeurs réelles, de sorte que les mots avec des significations proches présentent une représentation similaire. L'apprentissage de valeurs peut être effectué en utilisant les réseaux neuronaux. Un système largement utilisé qui intègre des mots est Word2vec (GloVe ou Gensim), qui contient des modèles tels que Skip-gram et le sac de mots continu (CBOW : Continuous Bag-Of-Words). Dans les deux cas, les modèles font appel à une probabilité que les mots successifs se rapprochent. Skip-gram commence par un mot et prédit les mots susceptibles. CBOW inverse, prédit un mot en fonction des mots de contexte reçus.

TF-IDF est une mesure statistique reflétant l'importance d'un mot pour un document dans une collection ou un corpus. Cette métrique prend en compte la fréquence du mot dans le document cible, ainsi que la fréquence dans les autres documents du corpus. Plus la fréquence d'un mot dans un document cible est élevée et plus sa fréquence dans les autres documents est faible, plus son importance est grande. La classe vectorizer dans la bibliothèque scikit-learn est généralement utilisée pour calculer TF-IDF. Les techniques d'incorporation de mots et de TF-IDF sont toutes deux utilisées comme caractéristiques d'entrée des algorithmes d'apprentissage profond en traitement automatique du langage naturel. Les tâches d'analyse de sentiment transforment des collections de données brutes en vecteurs de nombres réels continus. Divers types d'analyse ou de tâches relatifs à la tâche incluent la classification objective ou subjective, la détection de la polarité des sentiments et l'analyse de sentiment basée sur les caractéristiques ou les aspects. En fonction du contexte, le degré de subjectivité peut varier et un document objectif peut encore énoncer des phrases subjectives. L'analyse de sentiment basée sur les aspects correspond aux sentiments exprimés envers des aspects distincts des entités concernées; par exemple, le taux, la chambre, l'emplacement, la propreté ou le service. La polarité et l'intensité sont deux composantes utilisées pour noter l'analyse de sentiment. La polarité décide si c'est négatif, neutre ou positif. L'intensité indique la force relative du sentiment.

3. Approches et techniques d'analyse de sentiment dans le traitement du langage naturel

L'analyse de sentiment est un processus d'extraction d'informations sur une entité et d'identification automatique de toute subjectivité de cette entité. L'objectif est de déterminer si le texte généré par les utilisateurs véhicule leurs opinions positives, négatives ou neutres. La classification des sentiments peut être effectuée à trois niveaux d'extraction : le niveau de l'aspect ou de la caractéristique, le niveau de la phrase et le niveau du document. Actuellement, il existe trois approches pour aborder le problème de l'analyse de sentiment : (1) les techniques basées sur des lexiques, (2) les techniques basées sur l'apprentissage automatique et (3) les approches hybrides.

La figure 2-2 illustre une taxonomie des méthodes basées sur l'apprentissage profond pour l'analyse de sentiment.

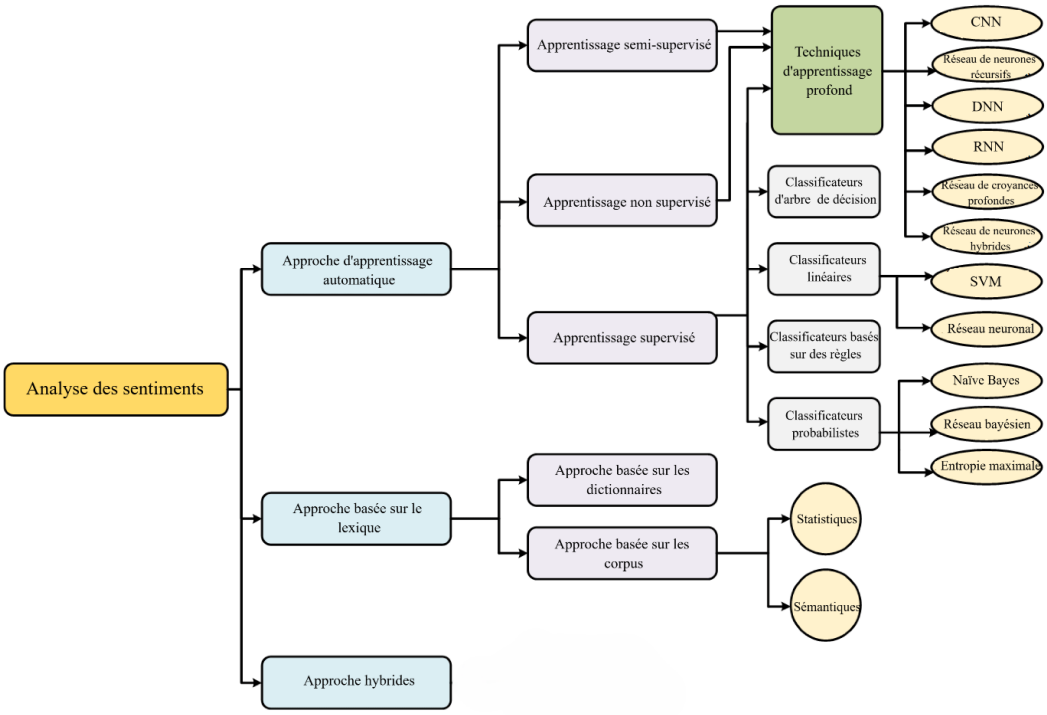

Figure 2-2 : Taxonomie des techniques d'analyse de sentiment

3.1.Les techniques basées sur les lexiques

Les techniques basées sur les lexiques (Lexicon-based) ont été les premières à être utilisées pour l'analyse de sentiment. Elles se divisent en deux approches : basées sur des dictionnaires et basées sur des corpus. Dans le premier type, la classification des sentiments est effectuée en utilisant un dictionnaire de termes, tels que ceux trouvés dans SentiWordNet (Esuli & Sebastiani, 2006) et WordNet[1]. Néanmoins, l'analyse de sentiment basée sur les corpus ne repose pas sur un dictionnaire prédéfini mais sur une analyse statistique du contenu d'une collection de documents, en utilisant des techniques basées sur les k plus proches voisins (k-NN : k-Nearest Neighbors), le champ aléatoire conditionnel (CRF : Conditional Random Field) et les modèles de Markov cachés (HMM : Hidden Markov Models) (Soni & Sharaff, 2015), entre autres. De nombreuses études ont utilisé des approches basées sur des lexiques, telles que celle de (Jurek, et al., 2015) qui ont développé un algorithme d'analyse des sentiments se concentrant sur l'analyse en temps réel du contenu Twitter. Cette méthode comprend deux composants principaux : une fonction de combinaison basée sur les preuves et la normalisation des sentiments, qui sont utilisés pour estimer l'intensité des sentiments.

Al-Khalisy et Jehlol (Alkhalisy & Jehlol, 2018) ont proposé une approche basée sur les dictionnaires pour extraire des informations significatives de la propagande terroriste, telles que le nom du compte, la localisation et les données des supporters. Cette méthode utilise le bag-of-words (BOW) pour calculer les scores globaux de chaque tweet représentant les

[1] Princeton University, 2010. Available at: https://wordnet.princeton.edu/

données d'entraînement et pour analyser la polarité, la liste de mots créée comprenant des antonymes et des synonymes du dictionnaire.

Chalothorn et Ellman (Chalothorn & Ellman, 2012) ont suggéré l'utilisation de ressources lexicales telles que NLTK toolkit, SentiWordNet, et WordNet pour l'analyse des publications radicales en ligne. La polarité et l'intensité du texte sont calculées pour analyser le sentiment. Pour cela, le corpus de texte a été initialement acquis à partir de diverses plateformes web comme Qawem et Montada, et après un prétraitement des données essentiel, diverses mesures basées sur les attributs ont été employées pour identifier et gérer le contenu extrémiste et religieux.

Dans le travail de (Al-Bakri, et al., 2022) les auteurs ont proposé un modèle pour évaluer les entreprises de e-tourisme en utilisant des avis en dialecte irakien recueillis sur Facebook. Ces avis sont analysés par des méthodes de text mining pour la classification des sentiments. Les mots sentimentaux générés sont classés en posts « positifs », « négatifs » et « neutres » en utilisant la théorie des ensembles approximatifs, la méthode des k plus proches voisins (K-NN) et les techniques de Naïve Bayes. Les résultats expérimentaux ont testé 71 entreprises de tourisme irakiennes, avec 28 % de ces entreprises ayant une très bonne évaluation, 26 % une bonne évaluation, 31 % une évaluation moyenne, 4 % une évaluation acceptable et 11 % une très mauvaise évaluation. Ces résultats ont aidé les entreprises à améliorer leur travail et leurs programmes pour répondre de manière adéquate et rapide aux demandes des clients.

3.2.Les techniques basées sur l'apprentissage automatique

Les techniques basées sur l'apprentissage automatique (Machine-learning-based) proposées pour les problèmes d'analyse de sentiment peuvent être divisées en deux groupes : (1) les modèles traditionnels et (2) les modèles d'apprentissage profond.

3.2.1. les modèles traditionnels

Les modèles traditionnels font référence à des techniques classiques d'apprentissage automatique, telles que le classifieur Naïve Bayes, le classifieur d'entropie maximale ou les SVMs. Les entrées de ces algorithmes comprennent des caractéristiques lexicales, des caractéristiques basées sur les lexiques de sentiment, des parties du discours, ou des adjectifs et des adverbes. La précision de ces systèmes dépend des caractéristiques choisies.

Dans l'enquête de (Anto, et al., 2016), les auteurs ont proposé une technique de rétroaction automatique basée sur les données de Twitter. Différents classificateurs comme SVM, Naïve Bayes et entropie maximale ont été utilisés sur les commentaires Twitter. Parmi ces classificateurs, la performance basée sur SVM était la plus élevée.

Ali Hassan (Hasan, et al., 2018) a réalisé une analyse des sentiments avec les deux principaux algorithmes d'apprentissage automatique, à savoir Naïve Bayes et SVM. Ils ont collecté un ensemble de données de 100 000 tweets et, après prétraitement, il restait 6250 tweets. Les résultats de Naïve Bayes ont montré une précision maximale de 79 %

avec le calcul de polarité W-WSD. En revanche, SVM a atteint une précision maximale de 62,33 % en utilisant W-WSD. L'étude avait une collection de données très limitée et n'utilisait que deux algorithmes d'apprentissage automatique pour l'analyse.

Les auteurs (Kumar, et al., 2019), (Sharma & Moh, 2016) ont analysé les sentiments exprimés via SVM et Naïve Bayes concernant des événements politiques pour la prédiction des élections en Inde et aux États-Unis. L'étude a été réalisée en 2016 et a prédit les probabilités électorales du BJP avec une précision de 78,4 % en utilisant le test SVM, ce qui était 16,3 % plus élevé que la performance de Naïve Bayes.

(Anis, et al., 2020) a mis en œuvre une analyse des sentiments sur les avis des utilisateurs en utilisant trois types de classificateurs : Naïve Bayes, Random Forest et SVM. Après avoir calculé la matrice de confusion de chacun, il a trouvé que le SVM performe mieux que les autres classificateurs avec une précision d'environ 81,6 % et un F1-score de 66,5 %. Les avis sont classés en labels positifs, négatifs ou neutres.

Omar et al. (Omar, et al., 2021) ont identifié la relation entre les discours haineux et les sujets présents sur les plateformes sociales en ligne en se basant sur une méthode d'apprentissage automatique. Cette approche utilise la classification multi-étiquette en employant les classificateurs de régression logistique, SVM linéaire et forêt aléatoire. Pour classer le sentiment des textes en positif, neutre ou négatif, les auteurs ont utilisé des représentations de caractéristiques qui incluent TF-IDF, N-gram et BOW.

Rehman et al. (Rehman, et al., 2021) ont proposé une méthode pour détecter le texte radical sur Twitter, où le langage religieux joue un rôle significatif dans la radicalisation. Les auteurs ont utilisé à la fois des caractéristiques radicales et religieuses pour entraîner le modèle et ont appliqué TF-IDF pour l'ingénierie des caractéristiques à intégrer dans les classificateurs ML, y compris la forêt aléatoire, SVM et Naïve Bayes pour détecter la polarité des sentiments.

(Al-Mashhadany, et al., 2022) ont développé une méthode de SA pour classer les centres de beauté en Irak en catégories sains et malsains. Les chercheurs ont utilisé les commentaires des centres de beauté sur Facebook pour mettre en œuvre l'évaluation. Les méthodologies comprenaient deux méthodes : basées sur le lexique et basées sur l'apprentissage automatique. Trois mécanismes d'apprentissage automatique ont été mis en œuvre : la théorie des ensembles approximatifs, Naïve Bayes et les K-NNs. Il est à noter que la théorie des ensembles approximatifs est meilleure par rapport aux deux autres, atteignant une précision de 95,2 %, tandis que Naïve Bayes atteint 87,5 % et K-NN atteint 78 %.

3.2.2. les modèles d'apprentissage profond

L'apprentissage profond adapte une approche à plusieurs couches pour les couches cachées du réseau neuronal. Dans les approches traditionnelles de l'apprentissage automatique, les caractéristiques sont définies et extraites manuellement ou en utilisant des méthodes de sélection de caractéristiques. Cependant, dans les modèles d'apprentissage profond, les caractéristiques sont apprises et extraites automatiquement,

ce qui permet d'obtenir une meilleure précision et performance. En général, les hyperparamètres des modèles de classification sont également mesurés automatiquement. La Figure 2-3 montre les différences dans la classification de la polarité des sentiments entre les deux approches : apprentissage automatique traditionnel (SVM, réseaux bayésiens ou arbres de décision) et apprentissage profond (Dang, et al., 2020). Les réseaux neuronaux artificiels et l'apprentissage profond fournissent actuellement les meilleures solutions à de nombreux problèmes dans les domaines de la reconnaissance d'images et de la parole, ainsi que dans le traitement du langage naturel. Plusieurs types de techniques d'apprentissage profond sont discutés dans cette section.

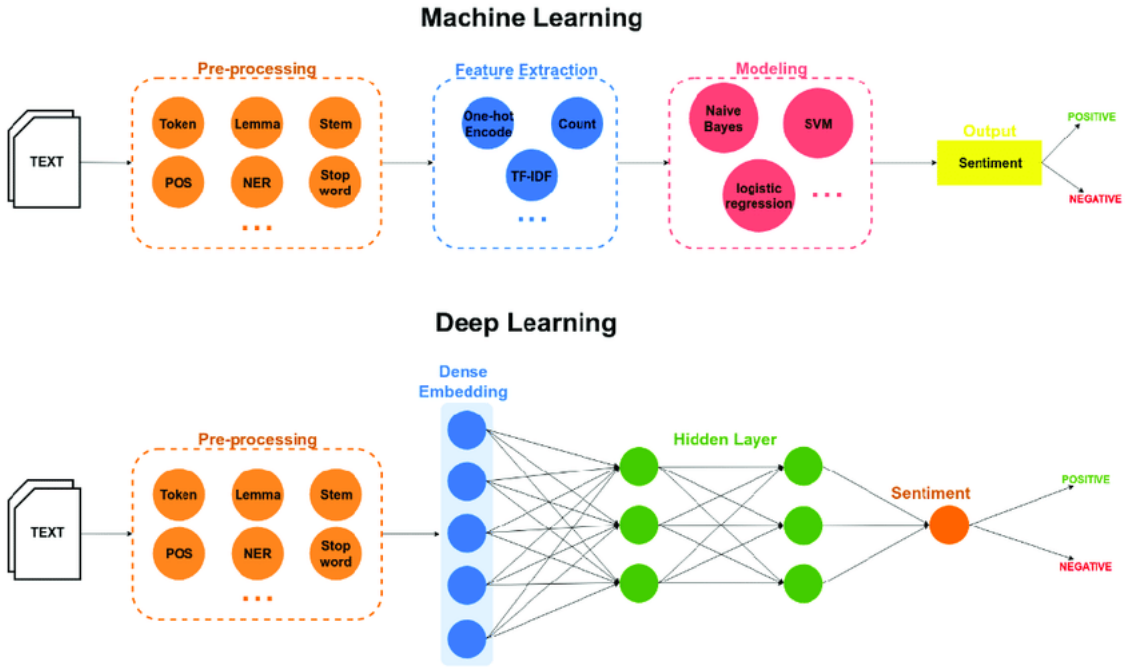

Figure 2-3 : Différences entre deux approches de classification de la polarité des sentiments, apprentissage automatique et apprentissage profond (Dang, et al., 2020)

Réseaux Neuronaux Profonds (DNN)

Un réseau neuronal profond (Aggarwal, 2018) est un réseau neuronal avec plus de deux couches, dont certaines sont des couches cachées.

Comme le montre la Figure 2-4, l'architecture de DNN possède davantage de couches cachées (au moins deux couches théoriquement). L'idée principale de DNN est de considérer la sortie de la couche cachée précédente comme l'entrée de la couche cachée actuelle afin d'obtenir des caractéristiques de haut niveau plus abstraites.

Afin d'éviter la disparition ou l'explosion du gradient, l'initialisation des poids dans les réseaux de neurones est également une tâche importante. Les approches d'initialisation existantes incluent principalement l'initialisation par distribution uniforme, l'initialisation par distribution gaussienne, l'initialisation Xavier (Glorot & Bengio,

2010), et l'initialisation Kaiming (He, et al., 2015). (Liu, et al., 2022) ont proposé une méthode améliorée et étendue d'initialisation des poids avec une fonction d'activation asymétrique, qui peut étendre la gamme de sélection de la fonction d'activation et améliorer les performances du réseau.

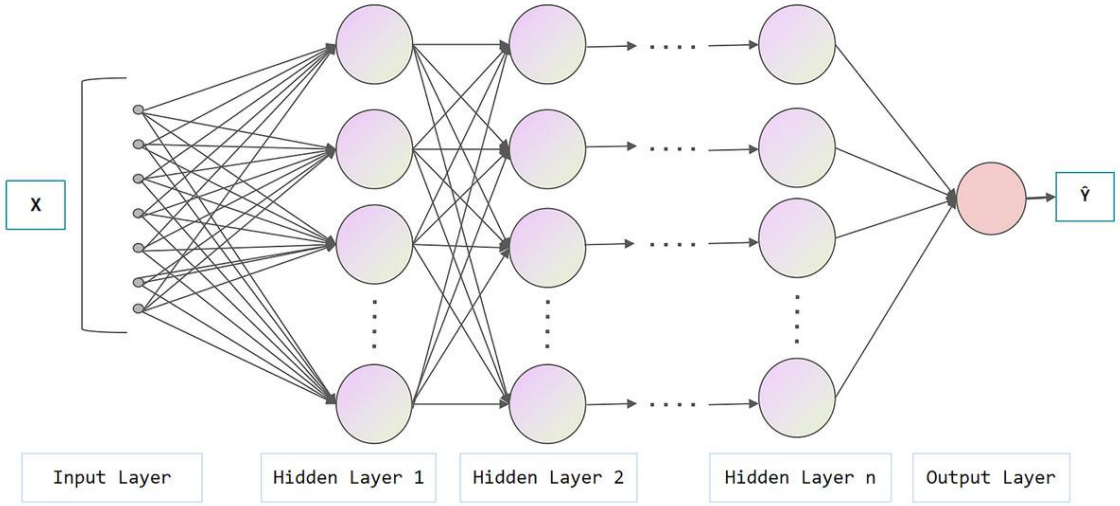

Figure 2-4 : Réseaux Neuronaux Profonds (DNN) (Zhu, et al., 2022)

Réseaux de Neurones Convolutifs (CNN)

Un réseau de neurones convolutifs est un type particulier de réseau de neurones feed-forward initialement utilisé dans des domaines tels que la vision par ordinateur, les systèmes de recommandation et le traitement du langage naturel. Il s'agit d'une architecture de réseau de neurones profond (Zhang, et al., 2018), typiquement composée de couches de convolution et de couches de pooling ou de sous-échantillonnage pour fournir des entrées à une couche de classification entièrement connectée. Les couches de convolution filtrent leurs entrées pour extraire des caractéristiques ; les sorties de plusieurs filtres peuvent être combinées. Les couches de pooling ou de sous-échantillonnage réduisent la résolution des caractéristiques, ce qui peut augmenter la robustesse des CNN au bruit et aux distorsions. Les couches entièrement connectées exécutent les tâches de classification. Un exemple d'architecture de CNN peut être vu dans la Figure 2-5. Les données d'entrée ont été prétraitées pour les remodeler en une matrice d'embedding.

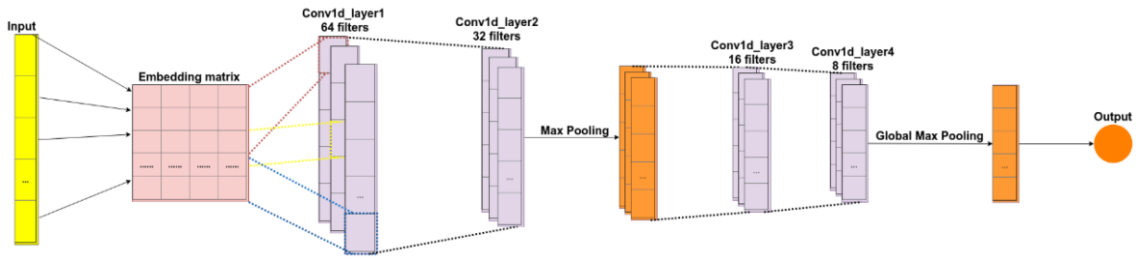

Figure 2-5 : Réseaux de Neurones Convolutifs (CNN) (Dang, et al., 2020)

La figure 2-5 montre une matrice d'embedding d'entrée traitée par quatre couches de convolution et deux couches de max pooling. Les deux premières couches de convolution possèdent respectivement 64 et 32 filtres, qui sont utilisés pour entraîner différentes caractéristiques ; celles-ci sont suivies par une couche de max pooling, utilisée pour réduire la complexité de la sortie et pour éviter le surapprentissage des données. Les troisième et quatrième couches de convolution possèdent respectivement 16 et 8 filtres, qui sont également suivies par une couche de max pooling. La couche finale est une couche entièrement connectée qui réduira le vecteur de hauteur 8 à un vecteur de sortie d'un, étant donné qu'il y a deux classes à prédire (Positif, Négatif).

CNN a été utilisé dans de nombreuses analyses de sentiments. Cependant, dans le travail de (Zhou, et al., 2023) les auteurs ont exploré un nouveau modèle CNN basé sur des réseaux convolutifs position-gated res2net, incorporant des caractéristiques de fusion sélective pour l'analyse des sentiments en utilisant la technologie des réseaux résiduels et des mécanismes d'attention.

Dans l'enquête de (Cao, et al., 2023) les auteurs ont proposé une méthode Multi-Scale Concatenation CNN (MSCCNN) avec des caractéristiques mixtes et un CNN amélioré pour reconnaître l'état de santé des machines tournantes, visant à améliorer la supériorité et la capacité de généralisation du modèle .

Hanyun Li et al. (Li, et al., 2023) ont introduit un algorithme à double canal qui intègre CNN et Bidirectional Long Short-Term Memory (BiLSTM) avec un mécanisme d'attention (DC-CBLA).

Réseaux de Neurones Récurrents (RNN)

Les réseaux de neurones récurrents (Britz, 2015) sont une classe de réseaux de neurones dont les connexions entre les neurones forment un cycle dirigé, créant des boucles de rétroaction au sein du RNN.

La fonction principale des RNN est le traitement des informations séquentielles sur la base de la mémoire interne capturée par les cycles dirigés. Contrairement aux réseaux de neurones traditionnels, les RNN peuvent se souvenir des calculs précédents et les réutiliser en les appliquant au prochain élément dans la séquence des entrées.

Un type particulier de RNN est le LSTM, capable d'utiliser une mémoire longue comme entrée des fonctions d'activation dans la couche cachée. La Figure 2-6 illustre un exemple de l'architecture LSTM. Les données d'entrée sont prétraitées pour être remodelées en une matrice d'embedding (le processus est similaire à celui décrit pour les CNN). La couche suivante est le LSTM, qui comprend 200 cellules. La couche finale est une couche entièrement connectée, comprenant 128 cellules pour la classification de texte. La dernière couche utilise la fonction d'activation sigmoïde pour réduire le vecteur de hauteur 128 à un vecteur de sortie d'un, étant donné qu'il y a deux classes à prédire (positif, négatif).

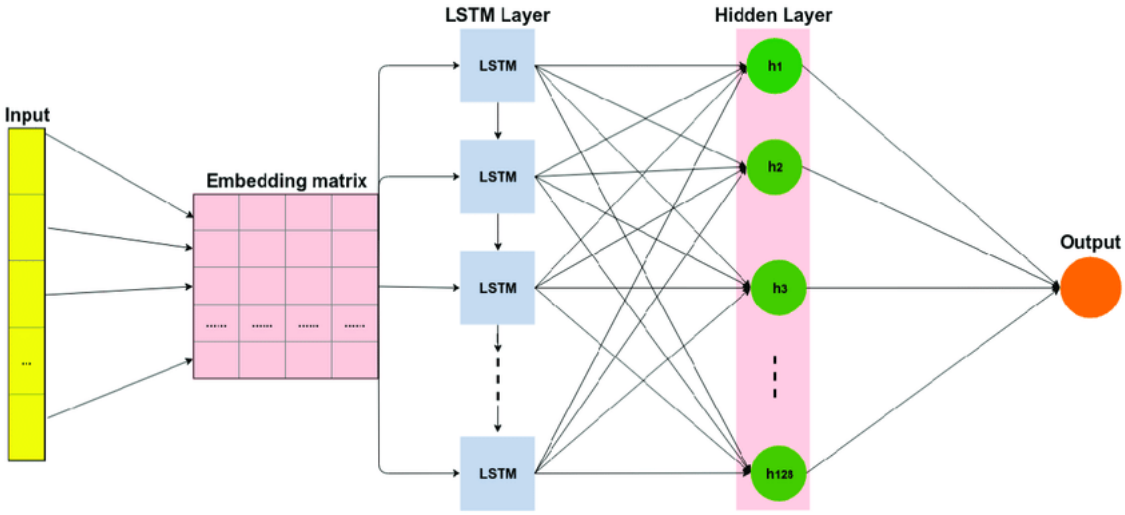

Figure 2-6 : Réseaux à Mémoire à Long Court Terme (LSTM) (Dang, et al., 2020)

Dans le travail de (Habbat, et al., 2023) les auteurs ont proposé une stratégie de combinaison multi-modèles, intégrant différents modèles RNN tels que LSTM, Bidirectional LSTM et GRU, en utilisant diverses techniques d'embedding de mots. Cette approche a atteint des performances significatives sur des données de tweets non structurés.

Ainapure et al. (Ainapure, et al., 2023) ont employé les techniques Bi-LSTM et GRU pour l'analyse des sentiments des commentaires postés sur la plateforme Twitter, atteignant des précisions de 92,70 % et 91,24 %, respectivement, sur un ensemble de données COVID-19.

Guesbaya et al. (Guesbaya, et al., 2023) ont introduit un transfert souple de capteur de réseau neuronal récurrent à mémoire à long terme (LSTM-R), estimant l'ouverture de ventilation en fonction des mesures des variables climatiques intérieures et extérieures.

3.3.Les approches hybrides

Les approches hybrides combinent des approches basées sur des lexiques et sur l'apprentissage automatique. Les lexiques de sentiment jouent généralement un rôle clé dans la majorité de ces stratégies.

Pour combiner les forces des différents modèles individuels, plusieurs chercheurs ont adopté des modèles hybrides qui intègrent deux ou plusieurs approches, telles que l'approche lexicale et l'apprentissage automatique, ou l'approche lexicale et l'apprentissage profond, et ainsi de suite. La combinaison de ces méthodes permet de surmonter les limitations de chaque approche (Dang, et al., 2020). L'avantage de combiner des approches basées sur l'apprentissage et des lexiques est qu'elle élimine le besoin d'un étiquetage manuel des données d'entraînement et permet la mesure et la détection de la polarité au niveau conceptuel. Ngoge (Ngoge, 2016) a développé une approche hybride combinant des techniques d'apprentissage automatique avec des méthodes lexiques pour classifier les sentiments afin d'identifier des activités terroristes. Cette approche utilise SVM, le

classificateur Naïve Bayes, et les méthodes d'entropie maximale en combinaison avec des méthodes lexiques pour prédire des modèles dans les tweets liés aux attaques terroristes au Kenya. Gupta et Joshi (Gupta & Joshi, 2020) ont proposé un modèle hybride qui extrait des vecteurs de caractéristiques à partir de SentiWordNet pour construire un classificateur SVM pour l'analyse des sentiments sur Twitter. Du et al. (Du, et al., 2017) ont appliqué un apprentissage automatique hiérarchique pour extraire les sentiments des opinions sur les vaccins HPV sur Twitter et ont conclu que la méthode était très efficace.

Bien que cela ait montré de bons résultats en utilisant la combinaison des méthodes lexique et d'apprentissage automatique, il y a plusieurs limitations à cela qui semblent dépendre de la qualité du lexique, qui n'est pas bien entendue par exemple pour gérer des contextes sémantiques au niveau d'humour, ou filtrer les mots non pertinents qui ajoutent du bruit aux avis. Pour surmonter la dépendance à la qualité du lexique, Diverses études utilisent l'apprentissage profond adoptant des modèles hybrides des mots. Il améliore considérablement les performances des tâches d'analyse des sentiments en créant de meilleurs modèles de mots pour des tâches complexes. Singh et al. (Singh, et al., 2022) ont développé un modèle hybride d'apprentissage profond intégrant des modèles LSTM et RNN avec des couches d'attention pour prédire les sentiments des données Twitter liées au COVID-19. (Ahmad, et al., 2019a) ont présenté une approche conjointe des modèles LSTM et CNN pour classifier les tweets liés à l'extrémisme. (Salur & Aydin, 2020) ont proposé l'amalgamation de divers embeddings avec plusieurs modèles d'apprentissage profond, y compris LSTM, CNN, BiLSTM, et GRU, pour extraire des caractéristiques des embeddings de mots et les fusionner pour la classification des sentiments. (Tam, et al., 2021) ont suggéré un modèle ConvBiLSTM, qui intègre Bi-LSTM et CNN pour classifier les sentiments en utilisant Word2Vec et GloVe pour obtenir des embeddings de tweets. Shehu et al. (Shehu, et al., 2021) ont appliqué trois méthodes d'augmentation des données pour augmenter la taille d'entraînement des données Twitter turques tronquées et ont ensuite utilisé RNN, le réseau d'attention hiérarchique (HAN), et CNN pour l'analyse des sentiments.

Dans les études (Lalji & Deshmukh, 2016), (Basiri & Kabiri, 2018) & (Verma & Thakur, 2018) une approche hybride a été appliquée pour améliorer l'efficacité de la classification des sentiments, et une efficacité particulièrement élevée a été atteinte dans l'étude de (Nandi & Agrawal, 2016).

3.4.Les techniques basés sur les Transformateurs

L'analyse des travaux existants révèle que peu d'études ont utilisé des modèles de langage pré-entraînés (PLM : Pre-trained Language Models) pour détecter les fausses nouvelles, et peu de recherches ont exploré comment exploiter au mieux ces PLMs pour cette tâche. Il devient extrêmement difficile de traiter manuellement des quantités massives de contenu généré par les utilisateurs (UGC : User-generated Content). Par conséquent, des systèmes automatisés capables de détecter les contenus falsifiés sont essentiels. Cependant, détecter les fausses nouvelles sur les réseaux sociaux est une tâche complexe, car celles-ci sont délibérément écrites pour tromper les lecteurs, et l'UGC est généralement de mauvaise qualité. Pour relever ces défis, les chercheurs ont proposé diverses méthodes pour interpréter le sens d'un mot à travers des vecteurs d'incorporation. Les méthodes basées sur

les réseaux neuronaux, telles que Word2Vec et GloVe, sont couramment utilisées pour apprendre les incorporations de mots à partir de grands corpus. Ces modèles d'incorporation présentent l'inconvénient d'être sans contexte, car le contexte est négligé et les incorporations statiques pour les mots sont générées indépendamment de leurs contextes. Pour obtenir une performance plus fine, un modèle doit être capable de capturer les motifs sémantiques et contextuels. De plus, un modèle d'apprentissage automatique ou d'apprentissage profond peut automatiquement extraire des informations sémantiques à partir d'une entrée donnée pour détecter les contenus falsifiés, mais ils ne peuvent pas reconnaître avec précision ces contenus sans une compréhension approfondie du texte. Il y a eu un intérêt croissant pour le paradigme de l'attention ces dernières années.

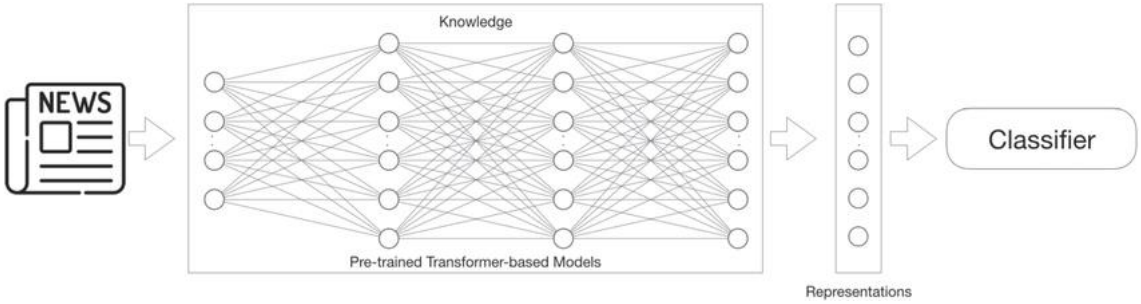

Figure 2-7 : La structure des modèles PLMs (Alghamdi, et al., 2023)

Un changement de paradigme global se produit dans la communauté NLP, visant à développer un ensemble de modèles qui non seulement améliorent la précision mais répondent également au problème du manque de données annotées, une question de longue date dans la communauté scientifique. De plus, il est urgent de détecter automatiquement les fausses nouvelles ; cependant, c'est une tâche difficile, car les modèles existants d'apprentissage automatique et d'apprentissage profond (avant l'avènement des modèles transformateurs) n'offrent pas une compréhension sémantique plus profonde des entrées textuelles. Cela a conduit la recherche en NLP à faire de grands progrès en introduisant des modèles de langage pré-entraînés basés sur les transformateurs. L'utilisation de PLM entraînés sur des données massives non annotées pour les tâches de classification de texte devient de plus en plus populaire. Pour s'adapter à la tâche en aval, de nouvelles couches de réseaux neuronaux sont ajoutées au-dessus des couches pré-entraînées dans le PLM. Comme le montre la Figure 2-7, une couche complètement connectée (FC) est ajoutée au-dessus des PLMs pour la classification. L'adoption des PLMs dans un cadre d'apprentissage par transfert facilite l'utilisation de leurs connaissances acquises (désignées comme connaissance dans la Figure 2-7). Cette connaissance peut être efficacement exploitée pour améliorer les performances sur des tâches spécifiques en utilisant des techniques telles que le fine-tuning ou l'extraction de caractéristiques (représentées comme Représentations). Ensuite, un classificateur (représenté comme Classificateur) peut être appliqué à ces représentations pour atteindre les objectifs de la tâche souhaitée. Une approche sophistiquée est nécessaire pour détecter les fausses nouvelles, car il devient de plus en plus difficile de distinguer entre le contenu faux et réel.

Cette section introduit trois PLM : BERT, DistilBERT et RoBERTa, qui sont considérés comme des percées majeures grâce à leurs performances impressionnantes sur une suite de tâches NLP, en grande partie grâce à leurs représentations puissantes de la langue apprises à partir de quantités massives de corpus de texte. Ces modèles peuvent être facilement ajustés sur une tâche en aval spécifique grâce à ce que l'on appelle l'apprentissage par transfert.

BERT : BERT, qui signifie Bidirectional Encoder Representation from Transformer, introduit par (Devlin, et al., 2019) est la première représentation de langage profondément bidirectionnelle et non supervisée qui joue un rôle d'encodeur de transformateur bidirectionnel multi-couches (réalise l'auto-attention dans les deux directions) qui conditionne conjointement les contextes gauche et droit dans toutes les couches. Ainsi, BERT génère des incorporations contextuelles. De plus, pour supprimer la contrainte de l'unidirectionnalité, BERT effectue un pré-entraînement en utilisant une tâche de prédiction non supervisée, y compris un modèle de langage masqué (MLM : Masked Language Model) qui est responsable de la compréhension du contexte et de la réalisation de prédictions (de mots). Ainsi, le modèle peut produire une représentation vectorielle qui peut capturer l'information générale du texte d'entrée. Ces représentations sémantiques de chaque mot dans le texte d'entrée peuvent être améliorées en utilisant un mécanisme d'attention dans le sens où différents mots dans un contexte montrent des effets différents pour renforcer la représentation sémantique. En tant que composant central de l'architecture des transformateurs, le rôle sous-jacent du mécanisme d'attention est d'attribuer moins ou plus de poids à différentes parties du texte vers la sortie (c'est-à-dire différencier la contribution de différentes parties de l'entrée sur la sortie). L'attention peut être considérée comme une fonction qui mappe les requêtes et suit les paires de clés-valeurs et de vecteurs de sortie.

En raison de l'incapacité des méthodes d'espace vectoriel telles que (CV : Count Vectorizer) et TF-IDF à prendre en compte le contexte, l'utilisation de ces représentations avec des classificateurs ML repose sur l'apparition de tokens pour prendre des décisions finales, indépendamment de leur contexte. Ces modèles d'espace vectoriel sont inefficaces pour capturer des motifs sémantiques et contextuels plus profonds, spécifiquement ceux contenus dans l'UGC (par exemple, les tweets). Un avantage majeur de BERT (et de ses variations) dans le cas de Twitter (où l'UGC contient souvent des fautes d'orthographe, du bruit et des abréviations) est l'utilisation de sous-tokens plutôt que de tokens fixes ; il est donc idéal pour être utilisé avec ces données au lieu d'incorporations de mots standard indépendantes du contexte. Bien que le modèle BERT ait réalisé une grande avancée dans la classification de texte, il est coûteux en termes de calcul, car il contient des millions de paramètres.

DistilBERT : Bien que BERT soit plus complexe à entraîner (en fonction du nombre de paramètres utilisés), une variation de BERT, appelée DistilBERT, fournit un nombre de paramètres plus simple et raisonnable par rapport à BERT (réduction de BERT de 40 % en taille tout en conservant 97 % de ses capacités de compréhension linguistique), permettant ainsi un entraînement plus rapide (60 % plus rapide).

RoBERTa : Robustly optimized BERT approach, introduit par Facebook. Il s'agit simplement d'un ré-entraînement de BERT avec une méthodologie de formation améliorée (i) en supprimant la tâche de prédiction de la phrase suivante du processus de pré-entraînement ; (ii) RoBERTa a été entraîné sur dix fois plus de données, et (iii) en introduisant un masquage dynamique utilisant des tailles de lots plus grandes pour que le token masqué change pendant l'entraînement plutôt que le modèle de masquage statique utilisé dans BERT. Ainsi, RoBERTa introduit une approche de pré-entraînement différente de celle de BERT.

4. Les méthodes de détection des fausses nouvelles basée sur l'analyse des sentiments

Le sentiment exprimé et sa force constituent un élément substantiel pour déterminer de manière fiable le degré de véracité d'une information. Parmi les développements de la dernière décennie, deux types d'approches ont émergé pour détecter les fausses nouvelles en ce qui concerne l'utilisation de l'analyse de sentiment (SA). D'une part, il existe un ensemble d'approches qui considèrent l'analyse de sentiment comme la base fondamentale de leur stratégie de détection des fausses nouvelles, généralement complétée par l'utilisation d'autres informations extraites à la fois du contenu des nouvelles et du contexte de leur propagation sur les réseaux sociaux. Ces approches sont celles auxquelles nous consacrons la Section 4.1. D'autre part, nous avons un ensemble plus important de modèles dans lesquels le sentiment exprimé dans un article de presse est considéré comme une caractéristique, parmi d'autres caractéristiques obtenues à partir du texte et du contexte de l'article. Les articles qui ont choisi cette approche sont l'objet de la Section 4.2.

4.1. Systèmes de détection de fausses nouvelles basés sur l'analyse de sentiment (SA)

Cette section explore les systèmes de détection de fausses nouvelles qui utilisent l'analyse de sentiment comme élément central de leur approche, en se concentrant sur la manière dont la polarité des sentiments exprimés peut indiquer la véracité des informations.

(Dey, et al., 2018) ont appliqué plusieurs méthodes de traitement automatique du langage naturel (étiquetage des parties du discours, reconnaissance des entités nommées, analyse de sentiment) à un ensemble de 200 tweets portant sur l'élection présidentielle américaine de 2016. Ils ont constaté que les tweets crédibles avaient principalement une polarité positive ou neutre, tandis que les tweets contenant du contenu faux avaient une forte tendance à la négativité. Cependant, leur ensemble de données était trop petit pour obtenir des résultats concluants. (Bhutani, et al., 2019) ont basé leur détecteur de fausses nouvelles sur l'analyse de sentiment en partant de l'hypothèse que le sentiment exprimé dans la rédaction d'un article de presse servirait de facteur décisif crucial dans le processus de caractérisation de l'information en vraie ou fausse. Ils ont appliqué un classificateur Naïve-Bayes pour déterminer le sentiment des textes, puis l'ont utilisé comme principale caractéristique des classificateurs Naïve-Bayes multinomiaux et des forêts aléatoires pour détecter les fausses nouvelles, ce dernier obtenant les meilleurs résultats.

(Ajao, et al., 2019) ont également exploité l'hypothèse selon laquelle il existe une relation entre les fausses nouvelles et le sentiment du texte publié en ligne. Cette hypothèse a été testée statistiquement sur un corpus de rumeurs (Zubiaga, et al., 2016) et a été confirmée ultérieurement par des résultats expérimentaux comparant plusieurs classificateurs classiques et d'apprentissage en profondeur utilisant le sentiment par rapport à une base de référence d'apprentissage en profondeur considérant uniquement les caractéristiques textuelles. Les meilleurs résultats pour la détection de fausses nouvelles ont été obtenus par des classificateurs sensibles au sentiment, un SVM dans le cas des modèles classiques et un LSTM avec attention hiérarchique dans le cas des modèles d'apprentissage en profondeur.

(Cui, et al., 2019) ont trouvé des preuves statistiques sur l'ensemble de données FakeNewsNet indiquant que la polarité du sentiment des commentaires sous de fausses nouvelles était plus grande que sous de vraies nouvelles. En conséquence, ils ont décidé d'incorporer les sentiments latents des utilisateurs dans un cadre d'incorporation en profondeur de bout en bout pour la détection de fausses nouvelles. Ils ont utilisé trois réseaux neuronaux pour traiter les images d'actualités, le texte d'actualités et les profils d'utilisateurs, tandis qu'un mécanisme adversarial a été introduit pour préserver la similarité sémantique et garantir la cohérence de représentation entre le texte et l'image. Enfin, ils ont modélisé le sentiment des utilisateurs pour l'incorporer dans le cadre proposé. Une nouveauté de ce travail était l'utilisation de l'apprentissage adversarial pour trouver des corrélations sémantiques entre différentes modalités de contenus d'actualités. Le système de détection de fausses nouvelles résultant a surpassé d'autres classificateurs basés sur l'apprentissage classique et en profondeur. Des expériences d'ablation ont montré que le composant contribuant le plus à la performance du système était l'analyse de sentiment.

(Vicario, et al., 2019) ont introduit un cadre d'alerte précoce sur les possibles cibles de désinformation sur les réseaux sociaux. Ils ont compilé un ensemble de données avec de vraies informations provenant de publications Facebook de journaux officiels italiens et de fausses nouvelles provenant de publications Facebook de sites italiens connus pour la diffusion de désinformation. Pour chaque publication, ils ont extrait les entités associées au contenu textuel et la polarité du sentiment exprimé dans le texte. Pour chaque entité, ils ont calculé la "distance de présentation" (la différence absolue entre les valeurs maximales et minimales parmi les scores de sentiment de toutes les publications contenant l'entité) et la "distance de réponse moyenne" (la différence absolue entre le score de sentiment moyen des publications contenant l'entité et le score de sentiment moyen de leurs commentaires). À partir de ces deux valeurs, ils ont établi la controverse et la perception de l'entité selon des seuils dérivés empiriquement. Ils ont remarqué que la distance de présentation était un bon indicateur de l'attention reçue par une entité en termes de likes et de commentaires et que les entités controversées et captivantes étaient beaucoup plus présentes dans les fausses nouvelles, mettant ainsi en évidence le potentiel de telles propriétés pour identifier les sujets susceptibles de faire l'objet de désinformation. Enfin, ils ont utilisé ces mesures pour dériver des caractéristiques basées sur le sentiment qui, avec d'autres basées sur les propriétés textuelles (par exemple, nombre de caractères, mots, etc.) et le comportement des utilisateurs (par exemple, nombre de commentaires, likes, etc.), ont alimenté plusieurs

classificateurs d'apprentissage automatique classiques pour détecter les fausses nouvelles. La meilleure performance a été obtenue par un classificateur de régression logistique.

Les fausses nouvelles en matière de santé présentent certaines caractéristiques qui les rendent plus difficiles à détecter que les fausses nouvelles dans d'autres domaines. Par exemple, elles peuvent induire en erreur le lecteur en déclarant une association comme une relation de cause à effet ou en mélangeant le risque absolu et le risque relatif, ce qui ne nécessite que de légères modifications basées sur des informations véridiques. (Dai, et al., 2020) ont mené une analyse exploratoire pour comprendre les caractéristiques des ensembles de données pour la détection de fausses nouvelles en matière de santé, analyser les motifs utiles et valider la qualité des ensembles de données FakeHealth. En ce qui concerne l'engagement social sur les nouvelles de santé, ils ont constaté que les réponses aux vraies nouvelles étaient plus positives. Dans le même domaine, (Kadan, et al., 2020) ont ciblé la détection de fausses nouvelles en matière de santé dans les sources médiatiques en ligne qui ressemblaient à des journaux traditionnels, étant donné que les informations étaient sous forme d'articles avec des informations fiables et un récit orienté vers les émotions abondant. Pour cette raison, ils ont basé leur approche de détection de fausses nouvelles sur les différents types de caractéristiques affectives affichées dans les articles de fausses et vraies nouvelles de santé. Les caractéristiques émotionnelles ont été extraites d'un lexique pour alimenter les classificateurs classiques et d'apprentissage en profondeur, et les résultats ont montré que l'information émotionnelle augmentait les performances pour tous les classificateurs. Ils ont également mené des expériences préliminaires sur la détection de fausses nouvelles sur le COVID-19, où ils ont constaté une présence significative de contenu émotionnel dans les récits, indiquant l'applicabilité de la détection orientée vers les émotions pour identifier les fausses nouvelles sur cette pandémie.

(Zhang, et al., 2021) ont considéré que la plupart des travaux existants sur les fausses nouvelles étaient basés sur les signaux émotionnels des contenus transmis par les éditeurs mais ne se concentraient que rarement sur les émotions des commentaires suscités dans la foule, même lorsque la propagation virale était alimentée par l'évocation d'émotions à forte intensité. Ils ont précisément exploré si les émotions dans les commentaires d'actualités et leur relation avec celles du contenu étaient utiles pour la détection de fausses nouvelles. Ils ont testé l'approche en utilisant divers classificateurs basés sur l'apprentissage en profondeur sur des ensembles de données en anglais et en chinois. Les résultats sur un ensemble de données chinois de fausses nouvelles étaient bons, tandis que les résultats sur l'ensemble de données en anglais étaient assez faibles, probablement parce que l'ensemble de données était initialement conçu pour la détection de rumeurs, pas de fausses nouvelles.

4.2.SA comme caractéristique pour les systèmes de détection de fausses nouvelles

Dans cette section, nous explorons comment SA est intégrée comme caractéristique clé au sein de systèmes complexes visant à détecter les fausses nouvelles.

(Popat, et al., 2016) ont abordé le problème d'évaluation de la crédibilité des affirmations textuelles arbitraires exprimées librement dans un contexte ouvert en trouvant

automatiquement des sources dans les actualités et les médias sociaux. Ces sources ont ensuite été utilisées pour alimenter un classificateur de régression logistique afin de déterminer si l'affirmation était vraie ou fausse. Un élément clé de leur approche était l'analyse du style dans lequel une affirmation était rapportée dans un article, en supposant qu'une affirmation vraie serait rapportée dans un langage objectif et impartial. Ils ont capturé le style linguistique au moyen d'un ensemble de fonctionnalités basées sur un lexique telles qu'une liste de mots à opinions positives et négatives ; des listes de verbes assertifs, factifs et descriptifs ; des mots de précaution ; des mots impliquants ; et des marqueurs de discours. Cette approche supposait que des preuves ou contre-preuves substantielles pouvaient être facilement obtenues à partir d'un instantané statique du web, mais cela n'était pas vrai pour les affirmations nouvellement émergentes avec une présence limitée sur le web. Pour surmonter cette limitation, (Popat, et al., 2017) ont proposé d'améliorer l'approche en déterminant l'orientation, la fiabilité et la tendance des sources de preuves ou de contre-preuves récupérées, qui ont été utilisées avec les fonctionnalités déjà définies dans (Popat, et al., 2016) pour alimenter un classificateur CRF.

(Varol, et al., 2017) ont étudié le développement de méthodes informatiques pour la détection précoce de campagnes d'information pouvant être utilisées pour propager des fausses nouvelles, de la propagande ou une manipulation des marchés financiers. En particulier, ils ont essayé de déterminer si un hashtag Twitter était promu sur la base d'informations disponibles même dans les cas où la nature d'une tendance est inconnue. Il s'agit d'une tâche difficile car une minorité de conversations promues est mélangée à une majorité de contenus organiques. De plus, les hashtags promus peuvent exister avant le moment où ils obtiennent le statut promu et peuvent avoir été créés sous une forme entièrement régulière, affichant ainsi des caractéristiques largement indiscernables de celles des autres hashtags réguliers sur le même sujet jusqu'au moment de la promotion. Ils ont proposé d'utiliser 487 fonctionnalités extraites de la structure du réseau et des schémas de diffusion, de l'information linguistique, du contenu et de l'information sentimentale, des signaux de synchronisation et des métadonnées utilisateur. Les fonctionnalités basées sur le sentiment incluaient la polarité et la force du sentiment, les émoticônes positives et négatives, le score de bonheur, l'excitation, les scores de valence et de dominance, et le score émotionnel. Parmi celles-ci, seules les mesures liées aux émoticônes figuraient parmi les 10 meilleures fonctionnalités pour les expériences réalisées avec des classificateurs d'apprentissage automatique.

Les différences de contenu entre les fausses et les vraies nouvelles politiques ont été étudiées par (Horne & Adali, 2017). Ils ont suivi la polarité du sentiment pour chaque phrase, les caractéristiques stylistiques (le nombre de fois que chaque partie du discours apparaît dans un article, le nombre de mots vides, la ponctuation, les citations, les négations, les mots informels/injurieux, les interrogations et les mots en majuscules), la complexité des mots (nombre de syllabes dans les mots, ratio de mots uniques, et nombre de mots communs et spécialisés), et la complexité des phrases (nombre de mots par phrase, profondeur de l'arborescence syntaxique de la phrase, et profondeur des arbres syntaxiques pour les groupes nominaux et verbaux). Ils ont constaté que les sentiments positifs et négatifs étaient des caractéristiques statistiquement significatives pour différencier le texte

principal des vraies et fausses nouvelles dans l'ensemble de données politiques de Buzzfeed Silverman. Cela ne s'est pas produit lorsque seuls les titres de nouvelles ont été considérés, probablement en raison de leur courte longueur.

(Rashkin, et al., 2017) ont étudié le langage des médias d'informations dans le contexte de la vérification des faits politiques et de la détection de fausses nouvelles. À cette fin, ils ont estimé l'utilisation de mots fortement et faiblement subjectifs avec un lexique de sentiments en supposant que les mots subjectifs étaient utilisés pour dramatiser ou sensationnaliser les reportages d'actualité. Les autres types de signaux lexicaux qu'ils ont considérés étaient les mots de précaution, les intensificateurs, les comparatifs, les superlatifs, les adverbes d'action, les adverbes de manière et les adverbes modaux. Ils ont constaté que les mots utilisés pour exagérer (mots subjectifs, superlatifs et adverbes modaux) étaient tous plus fréquemment utilisés dans les fausses nouvelles. Cependant, les fonctionnalités supplémentaires n'étaient utiles qu'avec des classificateurs classiques, car les améliorations pour les classificateurs basés sur des réseaux neuronaux profonds étaient négligeables par rapport à l'utilisation du texte comme seule caractéristique d'entrée.

(Vosoughi, et al., 2018) ont analysé les histoires vraies et fausses diffusées sur Twitter de 2006 à 2017. Ils ont constaté que les histoires fausses suscitaient les émotions de peur, de dégoût et de surprise dans les réponses, tandis que les histoires vraies suscitaient de l'anticipation, de la tristesse, de la joie et de la confiance. Ils en concluent que les émotions exprimées en réponse aux fausses nouvelles peuvent mettre en lumière des facteurs supplémentaires qui incitent les gens à partager de fausses nouvelles. Bien que l'analyse des émotions ne soit pas la même que l'analyse de sentiment, elles sont étroitement liées car toutes deux analysent le contenu subjectif exprimé dans un texte, et les modèles de classification utilisés sont très similaires dans les deux cas, la plus grande différence résidant dans l'ensemble de classes avec lesquelles les textes à traiter sont annotés. Pour cette raison, nous avons inclus cet article dans l'analyse.

(Yang, et al., 2018) ont analysé l'ensemble de données du détecteur de BS et ont conclu que la polarité du sentiment dans les vraies et fausses nouvelles était différente, les fausses nouvelles ayant tendance vers un sentiment négatif. L'écart type des fausses nouvelles sur le sentiment négatif était également plus grand que celui des vraies nouvelles, ce qui signifie que certaines des fausses nouvelles pouvaient avoir un sentiment négatif très fort. En ce qui concerne l'utilisation du langage dans les vraies et fausses nouvelles, ils ont observé que les fausses nouvelles avaient moins de mots et de phrases, et que la variance de ces valeurs dans les vraies nouvelles était beaucoup plus petite que celle dans les fausses nouvelles ; les vraies nouvelles avaient moins de points d'interrogation ; les fausses nouvelles avaient un ratio beaucoup plus élevé de lettres majuscules ; la médiane des négations dans les fausses nouvelles était beaucoup plus petite ; les fausses nouvelles avaient moins de pronoms de première et deuxième personne et plus de pronoms de troisième personne ; les vraies nouvelles avaient plus de diversité lexicale ; et il y avait des différences dans l'utilisation des mots dans les titres des fausses et vraies nouvelles. En plus de ces caractéristiques explicites, l'approche de (Yang, et al., 2018) était basée sur l'idée qu'il existe des motifs cachés dans les mots et les images utilisés dans les fausses nouvelles qui peuvent être

capturés avec un ensemble de caractéristiques latentes extraites via les multiples couches convolutionnels dans un réseau neuronal profond. L'originalité de cette approche résidait dans le fait qu'elle proposait un modèle unifié pour analyser à la fois le texte et les images des fausses nouvelles. En particulier, ils ont utilisé deux CNN parallèles pour extraire des caractéristiques latentes à partir des informations textuelles et visuelles. Ensuite, les caractéristiques explicites et latentes étaient projetées dans le même espace de caractéristiques.

(Reis, et al., 2019) ont extrait un grand ensemble de fonctionnalités à partir du contenu des actualités en utilisant des techniques de traitement du langage ainsi que des sources d'actualités (biais, crédibilité et fiabilité, et emplacement du domaine) et de l'environnement (nombre de likes, de partages et de commentaires, et le rythme auquel les commentaires sont postés). Une des fonctionnalités de contexte était la subjectivité et les scores de sentiment du texte des actualités. Toutes ces fonctionnalités ont été utilisées pour alimenter plusieurs classificateurs classiques, avec Random Forest et XGBoost obtenant les meilleurs résultats sur l'ensemble de données BuzzFace. À partir de leurs résultats, les auteurs ont observé qu'il était possible de choisir un seuil pour classer correctement presque toutes les fausses nouvelles (TPR proche de 1) tout en classant incorrectement 40 % des vraies nouvelles (FPR de 0,4), et ils ont considéré que cela pourrait être utile pour aider les vérificateurs de faits à identifier les histoires qui méritent d'être investiguées.

(Shu, et al., 2020b) ont analysé l'ensemble de données de FakeNewsNet, constatant que les gens exprimaient leurs émotions ou opinions à l'égard des fausses nouvelles à travers des publications sur les réseaux sociaux telles que des opinions sceptiques et des réactions sensationnelles, les vraies nouvelles ayant une plus grande proportion de réponses neutres par rapport aux réponses positives et négatives, tandis que les articles faux avaient un ratio plus élevé de sentiment négatif. Dans leurs expériences préliminaires pour classifier les fausses nouvelles, ils ont utilisé des fonctionnalités de base provenant uniquement du texte, donc ils n'ont pas fourni d'informations sur l'impact que l'analyse de sentiment a sur la détection des fausses nouvelles dans cet ensemble de données.

4.3. Étude comparative des systèmes de détection de fausses nouvelles utilisant l'analyse de sentiment

Les tableaux 2-1 et 2-2 présentent les systèmes pour lesquels des mesures de performance ont été fournies. Nous pouvons constater comment les premiers modèles utilisaient des ensembles de données spécialement créés pour mener les expériences décrites dans chacun des articles. Plus tard, au fil du temps, de plus en plus d'ensembles de données ont été utilisés et pourraient être considérés comme des normes dans le sens où ils sont disponibles pour être utilisés par d'autres chercheurs afin de corroborer les résultats ou de tester leurs propres approches pour la tâche. Nous pouvons également observer que jusqu'en 2019, la plupart des travaux testaient un seul système d'apprentissage automatique pour la phase finale du système de détection, tandis que, à partir de cette année-là, il est courant qu'un article donné rapporte des résultats pour différents classificateurs. Pour mettre ces résultats en contexte avec ceux obtenus avec d'autres méthodes de détection de fausses nouvelles ne faisant pas

intervenir l'analyse de sentiment, Oshikawa et al. (Oshikawa, et al., 2020) ont rapporté que le meilleur système sur l'ensemble de données LIAR a atteint une précision de 0,457. Dans le cas de FEVER, le meilleur système a atteint une précision de 0,647. En revanche, la précision des meilleurs systèmes augmente à 0,944 pour Buzzfeed Political News Data et à 0,938 pour la partie PolitiFact de FakeNewsNet. En ce qui concerne FakeNewsNet, Zhou et Zafarini (Zhou & Zafarani, 2020) ont fourni une analyse de l'évolution des performances à mesure que des fonctionnalités d'analyse lexicale, syntaxique, sémantique et discursive ont été considérées.

Il est difficile de comparer les résultats entre les systèmes en raison de l'utilisation d'une grande variété d'ensembles de données différents et de différentes mesures de performance. Alors que de nombreux systèmes rapportent la valeur du score F1 ainsi que celle de la précision P et du rappel R, c'est parce que F1 est en fait une mesure agrégée de P et R. D'autre part, les systèmes qui fournissent la valeur de l'exactitude Acc ne rapportent pas la valeur de F1 et vice versa. Seuls quelques systèmes rapportent la valeur d'Acc et AUC, l'un fournit AUC et F1 et un autre montre Acc et F1. En pratique, cette diversité de mesures n'est pas si pertinente car seuls les systèmes créés par les mêmes auteurs peuvent être comparés sur le même ensemble de données d'évaluation. D'un point de vue qualitatif, il n'est pas rare que, bien que les résultats de nombreux articles puissent coïncider dans bon nombre de leurs conclusions, ils présentent des divergences dans certains aspects spécifiques. Nous considérons que cela est dû à la jeunesse relative du domaine de la détection de fausses nouvelles. Avec l'organisation de tâches partagées dédiées en 2020 et 2021 et la création de grands ensembles de données ces dernières années, nous prévoyons qu'à l'avenir, il sera de plus en plus courant de voir des systèmes évalués sur ces nouveaux corpus standard.

Pour les méthodes SA utilisées dans ces systèmes, une mesure de la performance des systèmes SA qui ont été appliqués n'est généralement pas donnée. Cela nous empêche de savoir si une amélioration de la performance de l'analyse de sentiment induit une amélioration significative de la performance finale de détection de fausses nouvelles. De plus, la plupart des articles utilisent des systèmes SA relativement simples basés sur des lexiques. Les modèles qui constituent l'état de l'art en SA sont basés sur l'apprentissage automatique. Ces systèmes sont plus coûteux à construire et nécessitent plus de ressources computationnelles que les systèmes basés sur des lexiques.

Référence	Langage	SA Méthode	Détection Méthode	Jeu de données	Performance
(Popat, et al., 2016)	Anglais	Approche basée sur le lexique	Logistic Regression	Snopes	Acc = 71.96; AUC = 0.80
(Popat, et al., 2017)	Anglais	Approche basée sur le lexique	CRF	Snopes	Acc = 84.02; AUC = 0.86
(Horne & Adali, 2017)	Anglais	Approche basée sur le lexique	SVM	Silverman's Buzzfeed Political News	Acc = 0.77
				Ad-hoc (news articles)	Acc = 0.71
(Rashkin, et al., 2017)	Anglais	Approche basée sur le lexique	MaxEnt	Fact Checking	F1 = 0.55
			LSTM		F1 = 0.56
(Varol, et al., 2017)	Anglais	Approche basée sur le lexique	KNN	Ad-hoc from Twitter	Acc = 0.97; F1 = 0.81
(Dey, et al., 2018)	Anglais	Approche basée sur le lexique	KNN	Ad-hoc from Twitter	Acc = 0.66
(Yang, et al., 2018)	Anglais	N/A	TI-CNN	BS Detector	P = 0.9220; R = 0.9277; F1 = 0.9210
(Bhutani, et al., 2019)	Anglais	Naïve-Bayes	Random Forest	LIAR	AUC = 0.63
				Fake vs. real news project	AUC = 0.88
(Ajao, et al., 2019)	Anglais	Approche basée sur le lexique	SVM	Rumors	Acc = 0.86; P = 0.86; R = 0.86; F1 = 0.86
			LSTM-HAN		Acc = 0.86; P = 0.86, R = 0.82; F1 = 0.84
(Cui, et al., 2019)	Anglais	Approche basée sur les règles	Ad-hoc deep neural network	FakeNewsNet PolitiFact	F1 = 0.7724
				FakeNewsNet GossipCop	F1 = 0.8042
(Vicario, et al., 2019)	Italien	Commercial tool (Dandelion API)	Linear Regression	Ad-hoc from Facebook	P = 0.90; R = 0.90; FPR = 0.11; F1 = 0.90
			Logistic Regression		P = 0.91; R = 0.91; FPR = 0.08; F1 = 0.91
			KNN		P = 0.87; R = 0.87; FPR = 0.13; F1 = 0.87
			Decision tree		P = 0.89; R = 0.89; FPR = 0.12; F1 = 0.89

Tableau 2-1 : Principales caractéristiques des systèmes de détection de fausses nouvelles utilisant SA : systèmes fournissant des résultats de performance quantitatifs sur la tâche.

Référence	Langage	SA Méthode	Détection Méthode	Jeu de données	Performance
(Reis, et al., 2019)	Anglais	Approche basée sur les règles	KNN	BuzzFace	AUC = 0.80; F1 = 0.75
			Naïve Bayes		AUC = 0.72; F1 = 0.75
			Random Forest		AUC = 0.85 ; F1 = 0.81
			SVM		AUC = 0.79 ; F1 = 0.76
			XGBoost		AUC = 0.86 ; F1 = 0.81
(Kadan, et al., 2020)	Anglais	Approche basée sur le lexique	Naïve Bayes	HWB	Acc = 0.790
			KNN		Acc = 0.925
			SVM		Acc = 0.900
			Random Forest		Acc = 0.840
			Decision Tree		Acc = 0.940
			AdaBoost		Acc = 0.0.965
			CNN		Acc = 0.910
			LSTM		Acc = 0.920
(Zhang, et al., 2021)	Anglais	Approche basée sur le lexique + Commercial (Baidu AI)	BiGRU	RumourEval-19	F1 = 0.340
			BERT		F1 = 0.346
			NileTMRG		F1 = 0.342
	Chinois	Approche basée sur le lexique + Commercial (NVIDIA)	BiGRU	Weibo-20	F1 = 0.855
			BERT		F1 = 0.867
			HSA-BLSTM		F1 = 0.908

Tableau 2-2 : Principales caractéristiques des systèmes de détection de fausses nouvelles utilisant l'analyse de sentiment : systèmes fournissant des résultats de performance quantitatifs sur la tâche.

5. Conclusion

Le chapitre portait spécifiquement sur la technologie de la détection des fausses nouvelles par l'analyse des sentiments, qui semble être une percée prometteuse dans le NLP. Étant donné que l'analyse des sentiments permet un aperçu du texte manifesté par les émotions, son

implémentation dans les systèmes de détection peut leur donner une compréhension plus fine et, potentiellement, améliorer la performance.

Chapitre 3 : Analyse des sentiments pour la détection des fausses nouvelles : Système proposé

1. Introduction

Dans ce chapitre, nous exposons notre principale contribution à la résolution du problème à l'étude. Nous décrivons, tout d'abord, les objectifs spécifiques de cette contribution et justifions l'approche choisie. Ensuite, nous présentons l'architecture de système développé en justifiant le rôle de chacune de ses composantes. Nous concluons enfin ce chapitre par la méthodologie d'évaluation choisie pour mesurer la performance du système.

2. Objectifs du système proposé pour la détection des fausses nouvelles

L'analyse des sentiments est une méthode fondamentale en NLP qui relève de l'identification et de l'extraction des opinions d'un texte. Elle s'avère particulièrement pratique pour bien comprendre les émotions, les opinions transmises dans le document textuel : articles de journaux, critiques en ligne, messages sur les réseaux sociaux. Dans le domaine de la détection des fausses nouvelles, elle apparaît comme un outil formellement précieux, en tant que moyen supplémentaire pour la détermination de la véracité d'une information, en ce sens qu'en règle générale, les articles de fausses nouvelles sont souvent rédigés dans un ton qui est celui de l'extrême, d'un côté comme de l'autre (au niveau positif : enthousiasme, admiration, etc., au niveau négatif : ou indignation, colère, etc.), ce qui marque une différence par rapport à l'information réelle. Par conséquent, intégrer l'analyse des sentiments dans les modèles de détection des fausses nouvelles peut améliorer la performance de la détection en ajoutant des caractéristiques supplémentaires considérant l'intention du texte.

Le but principal de cette recherche est de proposer un modèle pour la détection des fausses nouvelles qui s'appuie sur une analyse de sentiment prise comme un des points principaux permettant d'améliorer le modèle de détection proposé. Ce modèle basé sur une approche d'apprentissage profond permet de traiter un vaste corpus de fausses nouvelles pour en extraire, grâce à l'analyse des sentiments, des caractéristiques susceptibles d'enrichir le modèle proposé par une information complémentaire basée sur l'analyse des sentiments des articles d'actualité. Ces caractéristiques viennent compléter les caractéristiques du contenu des nouvelles au sein du modèle de détection proposé pour permettre une plus grande performance du détecteur.

L'approche adoptée repose sur l'intégration de l'analyse des sentiments dans le processus de détection des fausses nouvelles, et ce pour plusieurs raisons. Tout d'abord, les sentiments présents dans les titres des nouvelles peuvent souvent signaler des indices discrets mais signifiants sur l'intention de l'auteur, ce qui est d'un intérêt particulier dans le cadre des fausses nouvelles. Ensuite, en utilisant une méthode d'apprentissage profond, il est possible de saisir des liens complexes et non linéaires entre les diverses caractéristiques, y compris les émotions, ce qui pourrait entraîner une meilleure généralisation et une détection plus précise des fausses nouvelles. Finalement, cette méthode permet de traiter de manière efficace de grandes quantités de données, ce qui est essentiel dans le domaine de la détection des fausses nouvelles où les ensembles de données sont souvent étendues et diversifiées.

Pour résumer, l'approche suggérée cherche à intégrer les avantages de l'analyse des sentiments et de l'apprentissage profond afin de créer un modèle solide capable de repérer de manière efficace les fausses informations, ce qui contribue à la lutte contre la désinformation.

3. Architecture générale du Système Proposé

Pour concevoir un système fiable et efficace de détection de fausses nouvelles, il est fondamental de déterminer une architecture robuste qui allie différentes étapes du traitement et de l'analyse des sentiments. La section qui suit décrit en détail l'architecture du système proposé, en explicite le rôle de chaque élément dans le processus général de détection.

L'architecture est fondée sur deux unités principales, représentées sur la figure 3-1, l'une dédiée à l'analyse des sentiments, l'autre à la classification de texte. Dans un premier temps le titre des nouvelles extraites de l'ensemble des données subit une phase de prétraitement de texte puis le modèle applique l'unité d'analyse de sentiments, qui va permettre d'évaluer le ton émotionnel du titre, puis le modèle fait passer le texte dans l'unité de classification, qui va analyser les caractéristiques linguistiques du texte. Les résultats des deux unités sont ensuite combinés et traités par une fonction d'activation sigmoïde dédiée à la classification binaire.

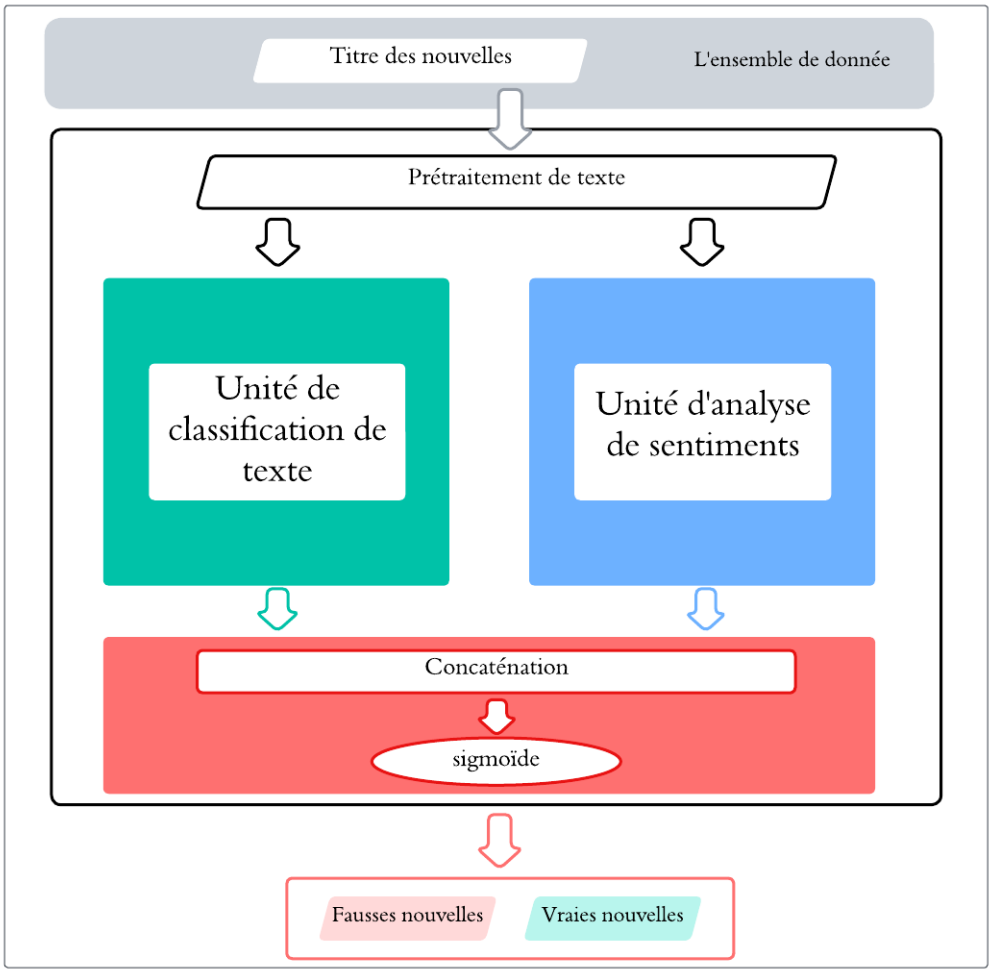

Figure 3-1: Détection des fausses nouvelles basée sur l'analyse des sentiments

3.1.L'ensemble de données

L'ensemble de données joue un rôle fondamental au sein du processus d'entraînement et d'évaluation du modèle puisqu'ils constituent les exemples nécessaires pour apprendre à faire la distinction entre les vraies nouvelles et les fausses nouvelles. Les données sont en général composées de titres et contenus de nouvelles avec des étiquettes indiquant si une nouvelle est vraie ou fausse.

3.2.Étapes de prétraitement de texte

Le prétraitement des données est une étape indispensable dans la construction de tout système d'analyse de texte, il permet de préparer les données brutes afin de rendre possibles leurs traitements par les algorithmes de traitement du langage naturel (NLP) et de classification. Dans le cadre de notre système proposé, le prétraitement de texte consiste à nettoyer puis à normaliser les titres d'actualités, avant leur passage dans les unités d'analyse de sentiment et de classification.

Le texte a été traité à l'aide de techniques de prétraitement qui sont :

(1) SUPPRESSION DES LIGNES VIDES ET DES TABULATIONS, Cette étape consiste à éliminer les lignes qui ne contiennent pas de texte utile et à retirer les tabulations, qui peuvent introduire des espaces inutiles dans les données.

(2) SUPPRESSION DES BALISES HTML, on élimine les balises HTML du texte, pour empêcher éléments de mise en forme ou balises de perturber l'analyse. Cette étape permet de ne garder que l'information textuelle.

(3) SUPPRESSION DES LIENS, URL et autres liens web sont enlevés car ils ne contiennent pas d'information susceptible d'intéresser l'analyse textuelle, mais plutôt du bruit.

(4) EXCLUSION DES CARACTERES ACCENTUES, les caractères accentués sont convertis en leur forme non accentuée pour standardiser le texte et neutraliser les complications dues à la reconnaissance des caractères spéciaux.

(5) EXPANSION DES CONTRACTIONS, les contractions usuelles (comme « it's » pour « it is ») sont développées dans leur forme classique ; cela permet d'améliorer la cohérence du texte et de le rendre plus facilement manipulable par les outils d'analyse.

(6) SUPPRESSION DES CARACTERES SPECIAUX SAUF (!, ?), les caractères spéciaux autres que les points d'exclamation et d'interrogation sont supprimés afin de réduire le bruit dans les données textuelles et de garder uniquement le nécessaire.

(7) SUPPRESSION DES MOTS VIDES, Les mots vides (ou « StopWords ») tels « the », « and », « or », sont exclus car en général n'apportent aucune valeur à l'analyse sémantique.

3.3.Unité de classification de texte

L'unité de classification de texte, un composant clé du système de détection de fausses nouvelles. Cette unité est responsable de l'analyse des caractéristiques linguistiques et

syntaxiques du texte afin d'identifier les indices susceptibles d'indiquer si une nouvelle est vraie ou fausse.

3.3.1. Modèle basé sur l'apprentissage profond : Bi-LSTM et CNN

Cette unité extrait des caractéristiques textuelles de tous les exemples étiquetés de fausses nouvelles et de vraies nouvelles grâce à un classificateur d'apprentissage profond comme illustré dans la figure 3-2.

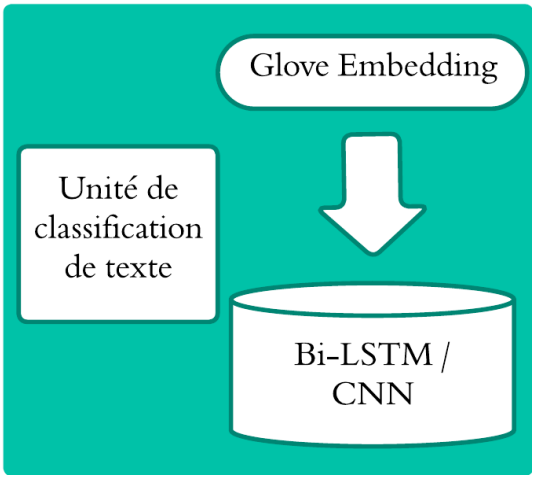

Figure 3-2 : Unité de classification de texte des modèles d'apprentissage profond

Cette unité commence par recevoir les titres des actualités après avoir été traités à l'aide des techniques de prétraitement. Ensuite, après les avoir traités, ils sont divisés en un ensemble pour entraîner le classificateur et un autre pour le tester. Cela est fait en utilisant 80% du total des données à des fins d'entraînement et 20% pour tester ensemble.

En second lieu, Glove, un modèle de word embedding pré-entraîné, a été utilisé pour les représentations de titre et pour apprendre les contextes des mots dans l'ensemble d'entraînement. L'embedding est une approche de représentation de document qui représente les mots et les distribue en un vecteur selon leur sémantique et leur syntaxique. Il convertit chaque mot de la chaîne de texte en vecteur de n dimensions, où n est la dimension d'embedding qui est de 300 dimensions dans ce modèle proposé. La distance entre les deux vecteurs d'embedding représente la proximité des mots en termes de relation sémantique. Par exemple, les mots « anxiété » et « dépression » sont sémantiquement liés car ils appartiennent à la même catégorie liée à la santé mentale, et de même les mots « mauvais » et « bon » sont également proches en embedding. Nous avons utilisé le modèle Glove pré-entraîné (840B tokens, 2.2M vocab, 300d vectors) (https://nlp.stanford.edu/projects/glove/, consulté le 18 octobre 2024) dans notre modèle pour prendre en compte à la fois les majuscules et les minuscules. Cela est bénéfique pour détecter les fausses nouvelles car les mots en majuscules sont fréquemment utilisés, et la taille du vocabulaire est un autre facteur crucial.

Enfin, après que chaque titre a été représenté sous forme de vecteur de mots à l'aide de Glove pré-entraîné, les séquences de vecteurs sont ensuite utilisées comme entrées pour le classificateur d'apprentissage profond une par une.

- Bi-LSTM

Le classificateur Bi-LSTM comprend deux couches de mémoire à long terme (LSTM), et comme le montre la figure 3-3, la couche LSTM avant et la couche LSTM arrière combinent de longues périodes d'informations contextuelles des deux directions, avant et arrière, d'une certaine période. La structure du Bi-LSTM capture le plus grand nombre de caractéristiques saillantes des deux directions. La couche avant apprend la séquence des entrées.

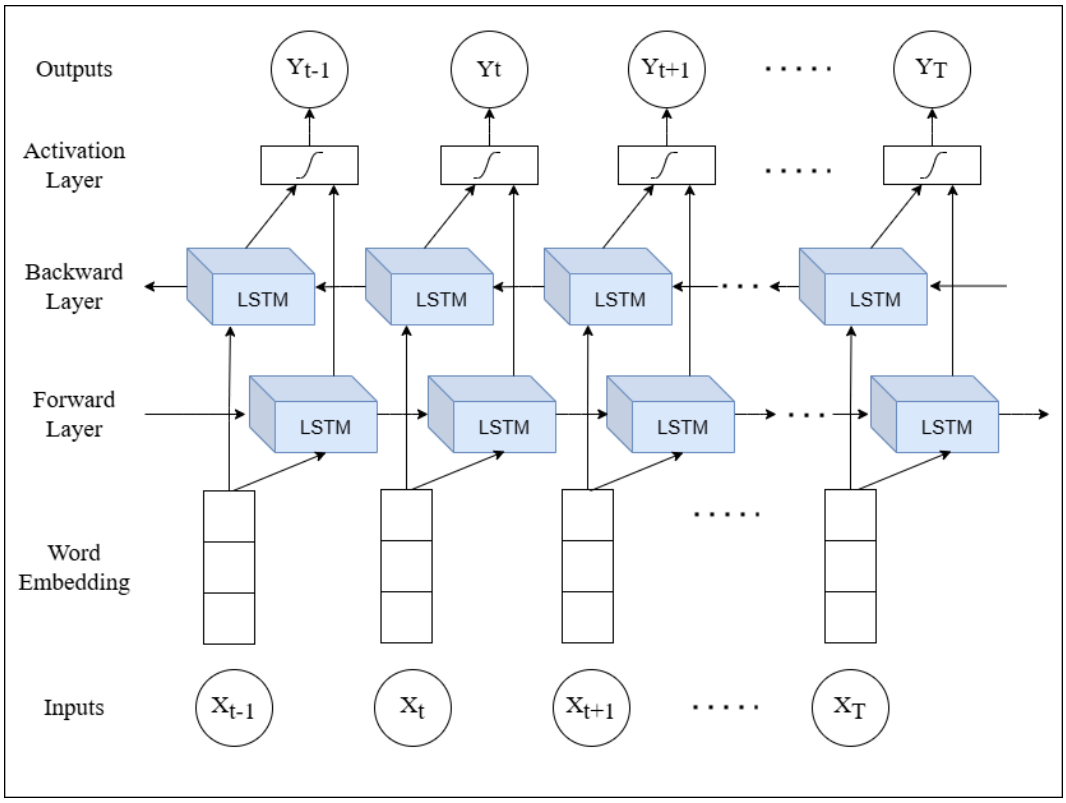

Figure 3-3 : L'architecture de base de Bi-LSTM utilise Word embedding.

Ces informations sont traitées par le LSTM avant de gauche à droite. Son état caché peut être illustré par la formule suivante :

$$\overrightarrow{h}t = \text{LSTM}\left(x_t , \overrightarrow{h}t - 1\right) \quad (1)$$

La couche arrière apprend l'inverse de la séquence de ces entrées ; c'est-à-dire que l'information sera traitée par le LSTM arrière de droite à gauche, et son état caché peut être représenté par la formule suivante :

$$\overleftarrow{h}t = \text{LSTM}\left(x_t, \overleftarrow{h}t + 1\right) \quad (2)$$

Ces deux couches de LSTMs sont connectées à la couche de sortie unique, comme indiqué dans la Formule (3).

Elles traversent la séquence d'entrée dans deux directions différentes simultanément.

$$h_t = \left[\overrightarrow{h}t, \overleftarrow{h}t\right] \quad (3)$$

La figure 3-4 montre l'architecture de base d'un réseau de neurones de type Bi-LSTM, c'est une architecture en couches, avec une entrée de séquence de 300 mots dans la première couche qui est une couche d'entrée textuelle, ensuite s'ajoute une couche Embedding qui transforme les mots du texte en vecteurs de taille 300.

Deux couches Bi-LSTM sont utilisées pour traiter des séquences de mots dans deux directions possibles. Ces réseaux sont particulièrement efficaces pour capturer des dépendances à long terme et des relations contextuelles au sein des données textuelles et extraire des motifs plus complexes dans les séquences.

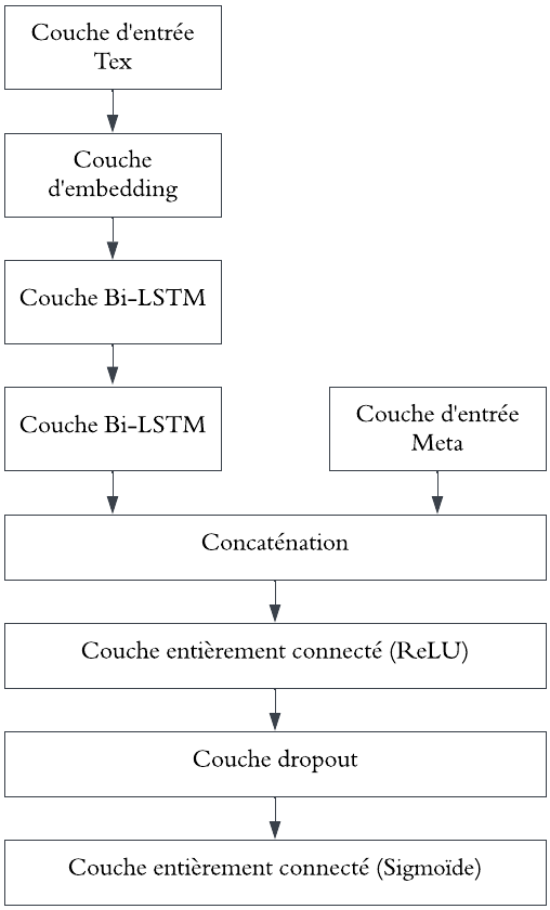

Figure 3-4 : Diagramme en couche de Bi-LSTM

Une fois les sorties des couches Bi-LSTM traitées, ces dernières sont concaténées avec des métadonnées, fournies par la deuxième couche d'entrée (Meta_input) qui permet d'enrichir la représentation linguistique textuelle résultant des précédents traitements.

La couche Concatenate fusionne les informations des différentes sources, qui sont ensuite traitées par une couche entièrement connecté avec une fonction d'activation ReLU. Cette activation permet de modéliser des relations non linéaires et d'augmenter la capacité d'apprentissage du modèle. Pour prévenir le surapprentissage et améliorer la généralisation, une couche de Dropout est intégrée dans l'architecture.

Enfin, le réseau se termine par une autre couche entièrement connecté finale, équipée d'une fonction d'activation sigmoïde, qui est conçue pour la classification binaire.

- CNN

Un réseau de neurones convolutionnel (CNN) est un type de réseau de neurones artificiel à propagation avant. Les CNN ont obtenu des résultats étonnants dans les domaines de la vision par ordinateur et de nombreuses tâches de traitement du langage naturel (NLP), telles que l'analyse des sentiments, la détection de fausses nouvelles, la détection de spam, la modélisation de phrases, et bien d'autres tâches connexes. Un CNN possède une ou plusieurs couches cachées, qui peuvent extraire des caractéristiques à partir de l'entrée (image, audio, vidéo, texte) et une couche entièrement connectée pour produire la sortie souhaitée.

Le CNN se compose principalement de trois types de couches comme illustré dans la Figure 3-5 :

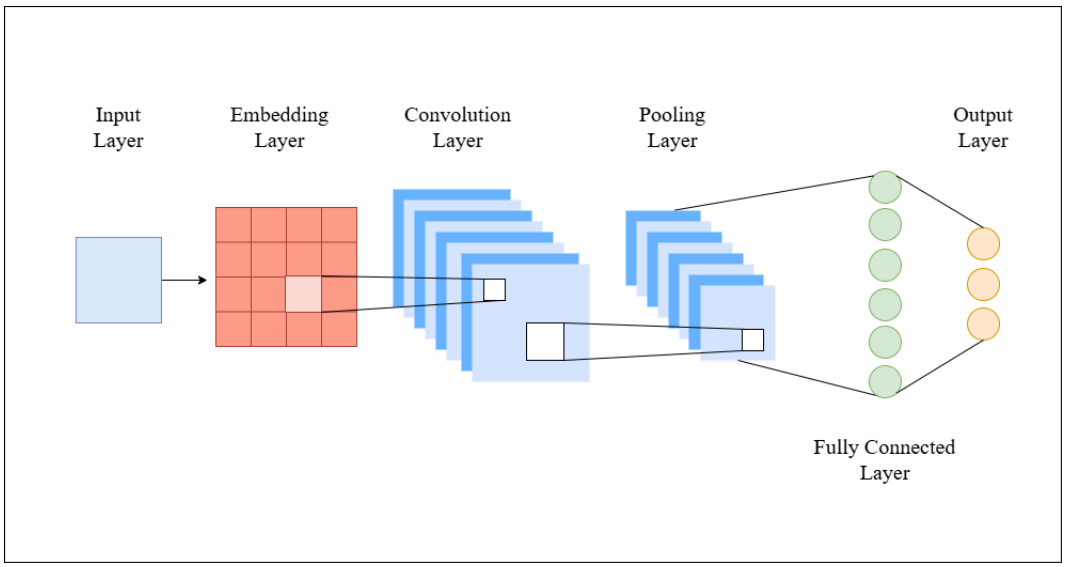

Figure 3-5 : L'architecture de base de CNN utilise Word embedding.

La couche de convolution (CONV) applique un ensemble de filtres qui peuvent identifier et reconnaître des caractéristiques ou des traits dans l'entrée (image, audio, vidéo, texte). La sortie de cette couche est une carte de caractéristiques, qui détermine les positions et les profondeurs des caractéristiques dans l'entrée en multipliant le filtre (ensemble de poids) avec la matrice d'entrée. Les paramètres les plus importants utilisés par la couche de convolution sont (1) le nombre de noyaux et (2) la taille des noyaux.6

La couche de convolution h est formée en appliquant la fonction d'activation $f(\cdot)$ à la matrice d'entrée X qui est convoluée avec la matrice de poids W_k et ajoutée au terme de biais b_k pour chaque couche. Les éléments de la i-ème ligne et de la j-ème colonne de W_k et X sont respectivement appelés $w_{i,j}^k$ et $x_{i,j}$. La carte de caractéristiques k de la h-ème couche résultante a une dimension $C \times H \times W$, où C, H et W représentent respectivement le canal, la hauteur et la largeur.
On peut créer une couche de convolution, notée h, en utilisant k petits filtres (également appelés noyaux) de taille $N_i \times N_j$ comme le montre l'équation 4. Ces filtres effectuent une opération de corrélation croisée, convoluant le pixel d'entrée $x_{u,v}$ pour obtenir $h_{u,v}^k$. (4)

$$h_{u,v}^k(X_{u,v}) = f\left(\sum_{i=1}^{N_i}\sum_{j=1}^{N_j} w_{i,j}^k x_{u+i,v+j} + b^k\right)$$

La couche de Pooling (couche de sous-échantillonnage) réduit la dimension de la carte de caractéristiques en appliquant des fonctions de max-pooling ou d'average-pooling.

La couche entièrement connectée (couche de classification) est la dernière couche ajoutée à la fin du modèle. Elle est utilisée pour afficher et prédire la sortie finale en utilisant la fonction d'activation sigmoid (pour la classification binaire) ou la fonction d'activation softmax (pour la classification multi-classe) pour calculer les scores des classes. Enfin, la couche de sortie présente l'étiquette de la classe.

Dans l'architecture représentée par la figure 3-6, la première couche de ce réseau de neurones est la couche d'entrée textuelle qui reçoit une séquence de 300 mots en entrée. La deuxième couche est Embedding, laquelle représente les mots du texte en utilisant des vecteurs de dimension 300, ce qui permet de capturer des relations sémantiques entre les mots.

Au sein de cette architecture, se trouvent plusieurs couches de réseau convolutionnels (Conv1D) pour tenter d'identifier différents motifs dans le texte par des filtres de taille adéquate. Des couches MaxPooling1D vont réduire les dimensions à la suite du traitement opéré par les précédentes couches de convolution et à réduire le nombre de paramètres du réseau. Pour améliorer les performances et prévenir le surapprentissage, on ajoute une couche Dropout dans le modèle, juste après la couche Flatten qui aide à régulariser l'apprentissage du modèle.

Les données sont ensuite concaténées. Après cette concaténation, une couche dense avec fonction d'activation ReLU traite l'information combinée, suivi d'une nouvelle couche Dropout à même de réduire encore le surapprentissage. Enfin, la sortie de la dernière couche est utilisée pour la classification binaire à l'aide d'une couche entièrement connecté avec une fonction d'activation sigmoïde.

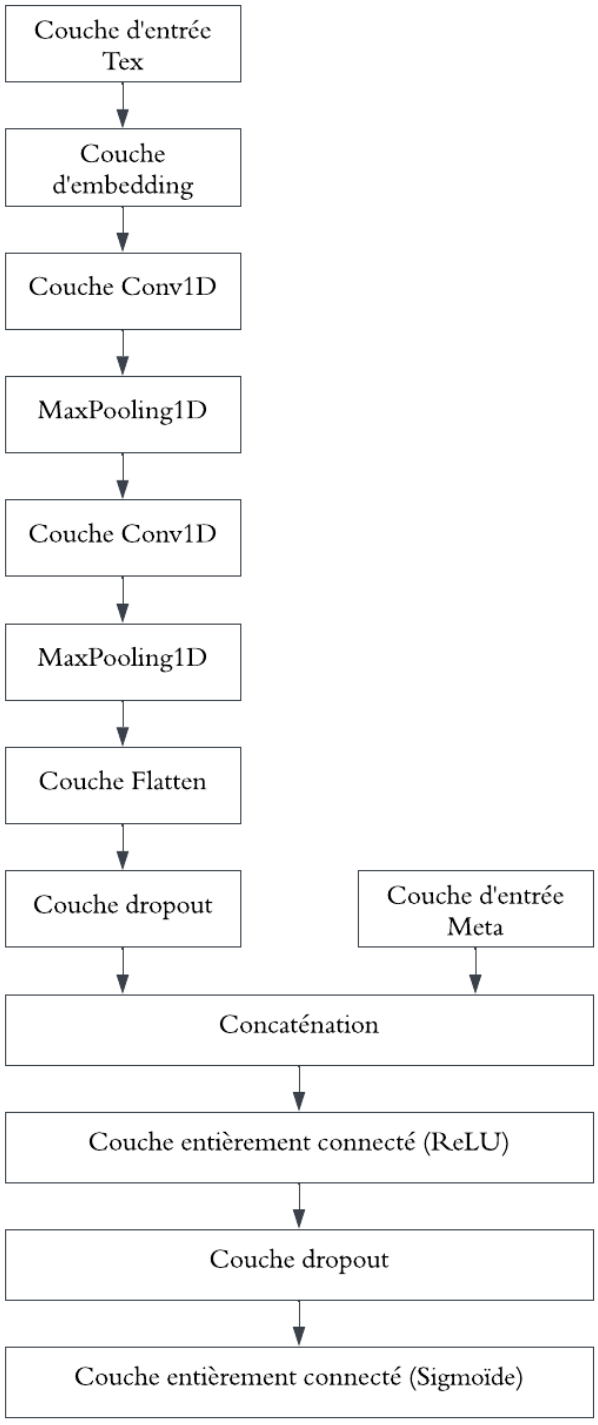

Figure 3-6 : Diagramme en couche de CNN

3.3.2. Modèle basé sur les transformateurs : BERT

BERT est une technique de machine virtuelle basée sur transformateur préformée à partir de données non étiquetées qui sont extraites de Wikipédia (langue : anglais) et BookCorpus. Transformer est équipé d'un encodeur et d'un décodeur qui lit les données d'entrée et prédit le résultat. BERT ne nécessite que la composante d'encodeur du Transformer pour accomplir sa tâche. Afin de clarifier le sens des termes vagues dans les données, BERT tente de fournir du contexte aux données environnantes. Le cadre BERT a été pré-entraîné à l'aide d'un grand volume de textes de Wikipédia.

L'unité de classification du texte repose sur le modèle pré-entraîné BERT, permet de capturer les relations sémantiques profondes et contextuelles des titres de nouvelles.

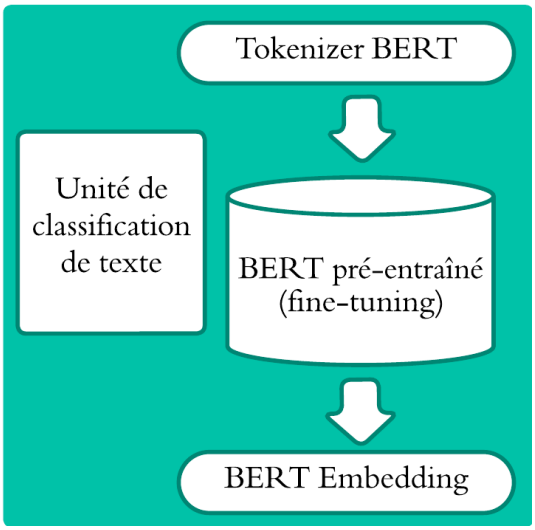

Figure 3-7 : Unité de classification de texte du modèle basé sur les transformateurs : BERT

Comme l'illustre la figure 3-7, pour la représentation d'entrée, le BERT Tokenizer segmente le texte en sous-mots et prépare l'entrée en ajoutant des jetons spéciaux comme `[CLS]`, qui marque le début de la première phrase pour les tâches de classification, et `[SEP]`, qui sépare les phrases suivantes.

BERT considère le contexte de chaque mot du texte. Le vecteur d'un mot dans les premières techniques d'embedding de mots était toujours le même, quel que soit l'endroit où le mot apparaissait dans le texte. Contrairement à ces techniques, BERT fournit des vecteurs différents pour les mêmes mots en fonction de leur position.

Les séquences de tokens générées par le tokenizer sont ensuite passées dans un modèle BERT pré-entraîné. Un modèle BERT pré-entraîné à la pointe de la technologie peut être adapté ou affiné pour diverses tâches. Ces tâches peuvent inclure des questions-réponses, la reconnaissance de langue, détection des fausses nouvelles, etc. Le modèle BERT est considéré comme à la pointe de la technologie car il atteint systématiquement les meilleures précisions dans plusieurs tâches de traitement du langage naturel (Devlin, et al., 2019).

Le processus d'adaptation de BERT comprend à la fois le pré-entraînement « Pre-training » et l'affinage « Fine-tuning » comme illustré dans la figure 3-8.

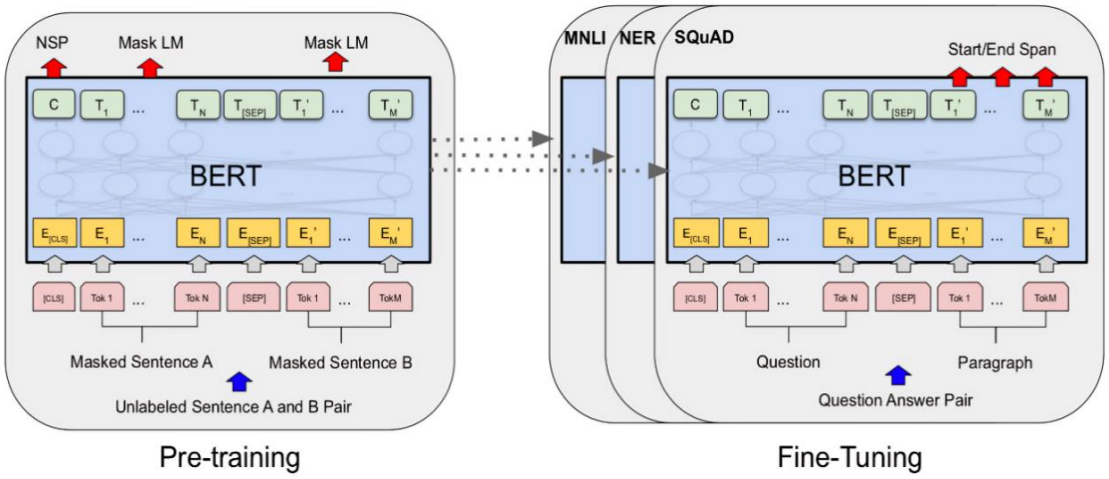

Figure 3-8 : Pre-training et fine-tuning du modèle BERT par (Devlin, et al., 2019)

Pre-training : Au cours de la phase de pré-entraînement du modèle BERT, celui-ci a été entraîné pour diverses tâches pré-entraînées en utilisant des données non étiquetées (Devlin, et al., 2019). Deux stratégies majeures que BERT utilise pour l'entraînement sont le masquage de mots (MLM) et la prédiction de la phrase suivante (NSP). Dans la technique MLM, 15 % des mots dans une phrase sont sélectionnés aléatoirement et masqués. En se basant sur le contexte des autres mots (qui ne sont pas masqués), le modèle tente de prédire le mot masqué. Dans la technique NSP, le modèle reçoit des paires de phrases en entrée. Le modèle apprend à prédire si la deuxième phrase d'une paire sélectionnée est la phrase suivante dans le document original. Le pré-entraînement de BERT est extrêmement coûteux. Il nécessite quatre jours sur quatre à seize Cloud TPUs.

Fine-tuning : La méthode d'affinage engendre des coûts plus faibles par rapport au processus de pré-entraînement. Le processus d'affinage commence par l'utilisation des paramètres pré-entraînés. Ces paramètres sont ensuite mis à jour avec des données étiquetées préparées en fonction du type d'étude. À l'exception des couches de sortie, les architectures utilisées pour l'affinage et le pré-entraînement étaient les mêmes. Cette méthode a été appliquée dans notre étude, permettant d'adapter le modèle pré-entraîné à nos données spécifiques pour obtenir des résultats optimaux.

BERT Embedding génère une représentation vectorielle dense pour chaque token du texte, capturant ainsi le sens et le contexte dans la séquence. Les embeddings générés par BERT sont envoyés à des couches entièrement connectées pour traiter davantage la sortie de BERT.

Dans notre système proposé, l'architecture en couches de BERT est représentée dans la figure 3-9, Le modèle commence avec 2 types d'entrée : le premier est les données textuelles, sur lequel un modèle BERT pré-entraîné va être appliqué pour produire des embeddings, le second étant constitué par les métadonnées directement sur lesquelles

nous allons travailler. Ces deux sources d'information sont ensuite concaténées pour créer une représentation unifiée.

Pour éviter le sur-apprentissage, nous allons utiliser une couche dropout qui désactive aléatoirement certains neurones.

Cette représentation va être transmise en succès à plusieurs couches entièrement connectées, la première étant activée par ReLU afin de trouver des relations plus complexe en profondeur. Une normalisation va stabiliser le processus d'apprentissage, suivie de nouveau d'une couche dropout pour renforcer l'idée de régularisation.

Enfin une couche avec activation sigmoïde produira la sortie finale qui devrait être présente sous forme de probabilité, adaptée à la classification binaire.

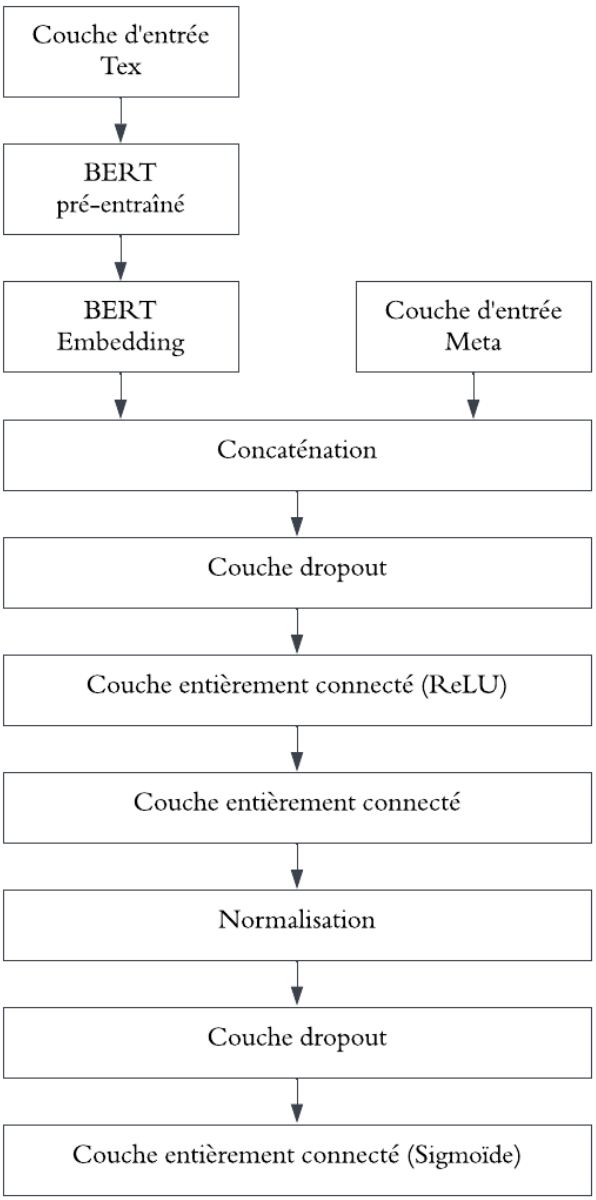

Figure 3-9 : Diagramme en couche de BERT

3.4.Unité d'analyse de sentiment

L'unité d'analyse de sentiment (SA) analyse également le sentiment des titres des nouvelles. L'analyse est faite sur le sentiment après que le texte a été traité à l'aide des techniques d'analyse de texte. L'analyseur de sentiment basé sur le lexique utilisé est TextBlob. TextBlob (https://pypi.org/project/textblob/0.9.0/, consulté le 27 octobre 2024) est une bibliothèque NLP basée sur le langage Python qui utilise le Natural Language Toolkit (NLTK) .

La polarité est la sortie du TextBlob, qui retourne une polarité variant de (-1 à +1) et où (+1) correspondra à la valeur la plus positive (ex : « great » et « best ») alors que (-1) à la valeur la plus négative (ex : « disgusting », « terrible » et « pathetic »). TextBlob reçoit en entrée textuelle une phrase qui est souvent décomposée dans une collection de mots. Après le scoring individuel de chaque mot, le sentiment ultime est déterminé à l'aide d'une procédure de regroupement en faisant la moyenne de tous les sentiments..

Comme le montre la figure 3-10, pour les modèles d'apprentissage profond, les résultats de l'analyse de sentiments des titres sont utilisés comme une caractéristique, qui sera ensuite concaténée avec les autres caractéristiques de l'unité de classification de textes au niveau de la couche de concaténation, afin d'optimiser la détection des fausses nouvelles par le modèle.

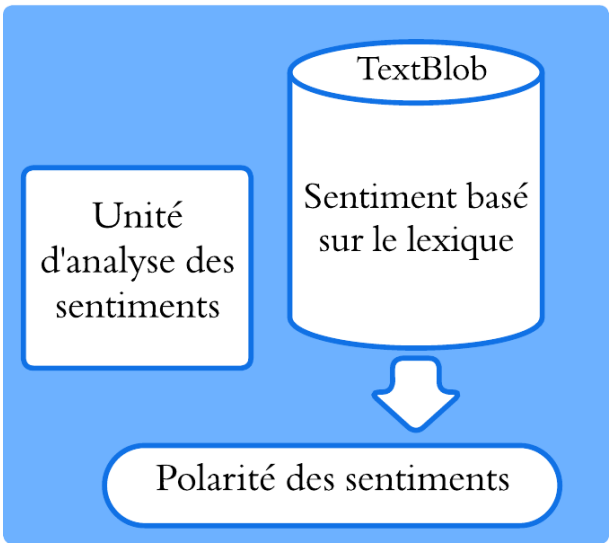

Figure 3-10 : Unité d'analyse de sentiment pour les modèles d'apprentissage profond

Dans des applications comme la détection des fausses nouvelles BERT peut être combiné avec d'autres unités comme une unité d'analyse de sentiment. Grâce à sa structure robuste et à sa capacité d'apprentissage profond, BERT est devenu un modèle de référence en NLP qui offre des performances intéressantes pour de nombreuses tâches linguistiques.

BERT est un modèle pré-entraîné sur des représentations de texte contextuelles. Il est conçu pour capturer des nuances linguistiques complexes, mais il peut ne pas bénéficier de l'ajout direct d'une valeur numérique continue comme le score de polarité. En revanche, une représentation binaire simplifiée du sentiment se combine efficacement avec les

représentations de BERT en tant qu'indicateur complémentaire, aidant à capturer la tonalité générale sans perturber ses représentations contextuelles riches.

Dans le modèle BERT, l'unité d'analyse des sentiments se différencie de celle utilisée dans les modèles d'apprentissage profond comme le montre la figure 3-11. En effet, la polarité des sentiments est extraite par TextBlob et classée en catégories de sentiments positifs, négatifs et neutres avant d'être transformée en valeurs numériques 1 pour les sentiments positifs et les sentiments neutres alors que 0 représente les sentiments négatifs. Ces valeurs sont ensuite intégrées aux représentations du modèle BERT. Cette fusion vient enrichir la décision finale facilitant ainsi la capacité à classer une nouvelle comme vraie ou fausse par le modèle.

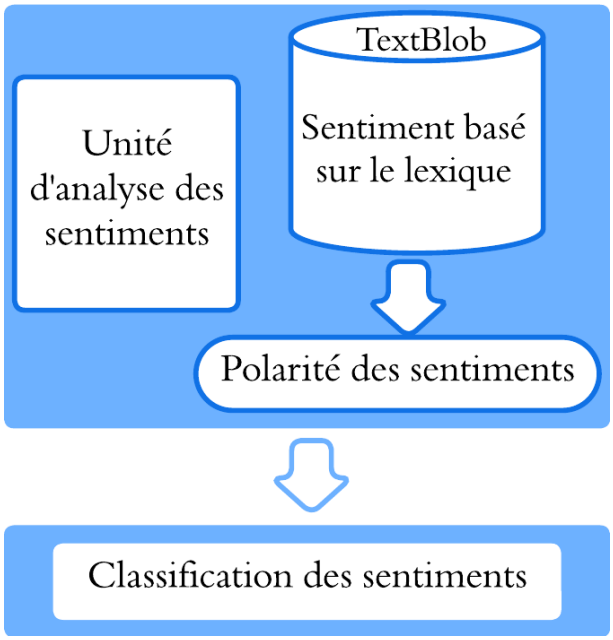

Figure 3-11 : Unité d'analyse de sentiment pour les modèles basés sur les transformateurs : BERT

4. Méthodologie d'évaluation

Pour évaluer les performances des systèmes de détection de fausses nouvelles basés sur l'apprentissage profond implémentés, leurs performances doivent d'abord être évaluées à l'aide de métriques d'évaluation appropriées qui ont été proposées dans des travaux antérieurs. Ici, les métriques utilisées pour évaluer nos modèles de détection de fausses nouvelles sont discutées.

En général, le problème de la détection des fausses nouvelles est un problème de classification binaire. Selon eux, pour déterminer si un article de presse est faux ou non, les comptes de vrais positifs, de vrais négatifs, de faux positifs et de faux négatifs issus des résultats sont pris en compte lors du calcul de ces métriques.

Les quatre comptes suivants sont importants dans ces métriques :

- ✓ Vrai Positif (TP) : La proportion des échantillons de fausses nouvelles prédites qui sont effectivement annotées/marquées comme fausses dans le jeu de données.

- ✓ Vrai Négatif (TN) : La proportion des échantillons de vraies nouvelles prédites qui sont effectivement annotées/marquées comme fausses dans le jeu de données.
- ✓ Faux Négatif (FN) : La proportion des échantillons de vraies nouvelles prédites qui sont effectivement annotées/marquées comme fausses dans le jeu de données.
- ✓ Faux Positif (FP) : La proportion des échantillons de fausses nouvelles prédites qui sont effectivement annotées/marquées comme fausses dans le jeu de données.

En utilisant les comptes/métriques ci-dessus, les métriques suivantes d'un problème de classification typique sont formulées :

- Exactitude (Accuracy)

L'exactitude mesure la similitude entre les informations fausses prédites et les véritables informations fausses. Une meilleure exactitude indique une performance globale supérieure du classificateur.

$$Accuracy = \frac{|TP| + |TN|}{|TP| + |TN| + |FP| + |FN|}$$

Figure 3-12 : Formule de l'exactitude (Accuracy)

- Précision (Precision)

La précision mesure la fraction de toutes les fausses nouvelles identifiées qui sont reconnues comme fausses. Cette mesure est utilisée pour identifier quelles nouvelles sont fausses. Une plus grande précision peut être obtenue en créant moins de prédictions positives lorsque l'ensemble de données est déséquilibré.

$$\Pr ecision = \frac{|TP|}{|TP| + |FP|}$$

Figure 3-13 : Formule de la précision (Precision)

- Rappel (Recall)

Le rappel est utilisé pour mesurer la perception. Si la valeur de la précision et du rappel est élevée, cela signifie que le classificateur produit des résultats de haute précision et de rappel élevés.

$$\mathrm{Re}\,call = \frac{|TP|}{|TP| + |FN|}$$

Figure 3-14 : Formule du rappel (Recall)

- Score F1 (F1-Score)

Ainsi, le score F1 est utilisé pour trouver un équilibre entre la précision et le rappel, ce qui peut donner une performance globale de prédiction de la détection des fausses nouvelles.

$$F1 - score = 2\ x\ \frac{precision \times recall}{precision + recall}$$

Figure 3-15 : Formule du score F1

Ces métriques sont généralement utilisées dans les approches d'apprentissage automatique et permettent ainsi d'évaluer les performances d'un classificateur sous différents angles. Si un classificateur donne des valeurs élevées de précision, rappel, F1 et exactitude, cela signifie une meilleure performance du classificateur.

5. Conclusion

En somme, ce chapitre a donné une vue d'ensemble de l'architecture du système proposé en expliquant ses principaux composants et de leur manière de fonctionner. Par ailleurs, la méthodologie d'évaluation a été expliquée pour mesurer la performance et la pertinence du système développé.

Chapitre 4 : Résultats expérimentaux

1. Introduction

Au sein du présent chapitre, nous visons à présenter en détail le processus d'implémentation des différentes méthodes de classification des fausses nouvelles, à savoir : Bi-LSTM, CNN, BERT. Après avoir parlé des jeux de données utilisés pour la classification, nous parlerons du pré-traitement de texte, de l'analyse des sentiments et de la classification de texte. Ensuite, nous présenterons les résultats obtenus grâce à plusieurs expérimentations tout en discutons les performances des différents modèles. Enfin, nous conclurons par une étude comparative.

2. Configuration des modèles du système proposé

Nous avons implémenté plusieurs modèles de classification, notamment Bi-LSTM, CNN, et BERT pour la détection des fausses nouvelles dans ISOT et GossipCop. Pour chaque modèle, nous avons ajusté les hyperparamètres afin d'optimiser les performances.

La sélection adéquate des hyperparamètres est une étape cruciale dans toute solution de Deep Learning.

2.1.Environnement d'exécution

Dans cette présente section, nous décrivons l'environnement d'exécution qui a été utilisé pour les expériences concernant nos modèles de classification. Afin d'assurer une puissance de calcul assez importante et de tenir les performances lors de l'apprentissage de modèles DL complexes, nous avons opté pour Google Colab.

Figure 4-1 : Google Colab

Ce service met à la disposition des utilisateurs un environnement de développement Python cloud, de plus, associé à des bibliothèques populaires pour le Machine Learning, le traitement du langage naturel, et qui permet, de manière gratuite, l'accès aux unités de traitement graphique (GPU) importante notamment pour l'accélération de l'entraînement des modèles en relation avec des réseaux de neurones profonds tels que CNN, Bi-LSTM ou encore BERT utilisés dans ce travail. Pour les expériences que nous avons réalisées, l'environnement choisi fait appel à un GPU Tesla T4 , connu pour sa puissance de calcul et son efficacité énergétique, ce qui est un atout pour le traitement de grands quantités de données. Ce GPU est doté de 15 Go de VRAM, qui permettent d'accueillir des lots de données conséquents et de conserver temporairement les poids de modèles volumineux, un facteur décisif pour éviter les erreurs de mémoire en cours de traitement. En outre, l'utilisation de Python dans cet environnement est le plus appropriée pour la mise en œuvre que nous souhaitons réaliser, grâce à son écosystème riche en bibliothèques efficaces

(PyTorch, TensorFlow…) qui facilitent la création, l'entraînement et la mise en œuvre des modèles de l'apprentissage profond.

Figure 4-2 : Bibliothèques python

2.2.Paramétrage du modèle Bi-LSTM

Cette section décrit la configuration des hyperparamètres du modèle Bi-LSTM, qui est conçu pour capturer les relations séquentielles dans le texte.

Les hyperparamètres du modèle de type Bi-LSTM spécifiquement déployé dans les tableaux 4-1 et 4-2 ont été choisis pour garantir les meilleures performances possibles du modèle sur les ensembles de test ISOT et GossipCop. En ayant une dimension d'embedding de 300, il est possible d'obtenir des représentations beaucoup plus riches du texte, les taux de dropout choisis veillant à prévenir le surapprentissage dans le cadre d'un processus d'optimisation des performances. Les régularisateurs L1 et L2 intervenant également comme moyens de régularisation qui vont, à la fois réduire la complexité du modèle et en renforcer les capacités de généralisation. Étant donné la nature binaire de la classification, une fonction d'activation sigmoïde est utilisée, associée à la fonction de perte Binary_Crossentropy, adaptée aux tâches de classification binaire. L'optimiseur Adam assure une convergence rapide. La taille de lots et un nombre d'époques limité à 5 suggèrent une approche qui vise l'efficacité en termes de temps de calcul, ce qui peut être adapté à un grand volume de données comme dans ISOT et GossipCop.

Nom de l'hyperparamètre	Type ou valeur
Dimension d'embedding	300
Dropout (dropout rate)	0.2
Dropout récurrent	0.1
Régularisateur L1	0.0001
Régularisateur L2	0.001
Fonction d'activation	Sigmoïde
Fonction de perte (Loss Function)	Binary_Crossentropy
Optimiseur (Optimizer)	Adam
Taille de lots (batch size)	512
No. d'époques (No. of epochs)	5

Tableau 4-1 : Hyperparamètres du modèle Bi-LSTM dans ISOT

Nom de l'hyperparamètre	Type ou valeur
Dimension d'embedding	300
Dropout (dropout rate)	0.2
Dropout récurrent	0.1
Régularisateur L1	0.0001
Régularisateur L2	0.01
Fonction d'activation	Sigmoïde
Fonction de perte (Loss Function)	Binary_Crossentropy
Optimiseur (Optimizer)	Adam
Taille de lots (batch size)	512
No. d'époques (No. of epochs)	5

Tableau 4-2 : Hyperparamètres du modèle Bi-LSTM dans GossipCop

2.3.Paramétrage du modèle CNN

Les hyperparamètres du modèle CNN ont été ajustés pour maximiser ses performances sur les ensembles de données ISOT et GossipCop, comme montré dans les tableaux 4-3 et 4-4. Ils sont configurés pour extraire des caractéristiques textuelles détaillées, permettant au modèle de détecter des motifs complexes dans les données.

La dimension d'embedding de 300 convertit les mots en vecteurs de taille fixe, capturant ainsi les relations sémantiques entre eux. La taille des kernels régit l'étendue des filtres employés pour extraire des caractéristiques locales, permettant de détecter des motifs textuels de différentes longueurs.

Les techniques de régularisation intégrées minimisent le risque de surapprentissage, améliorant ainsi la généralisation du modèle. En outre, des fonctions de perte et d'activation adaptées à la classification binaire permettent une optimisation spécifique à cette tâche. L'optimiseur Adam assure une convergence rapide en ajustant dynamiquement le taux d'apprentissage pour chaque paramètre.

La taille des lots détermine le nombre d'exemples utilisés pour chaque mise à jour des poids, équilibrant ainsi précision et efficacité, tandis que le nombre d'époques correspond aux passages complets sur les données, permettant au modèle de mieux affiner ses paramètres pour une généralisation optimale.

Nom de l'hyperparamètre	Type ou valeur
Dimension d'embedding	300
Taille de kernel	5
Régularisateur L1	0.001
Régularisateur L2	0.01
Dropout	[0.4,0.5]
Fonction d'activation	Sigmoïde
Fonction de perte (Loss Function)	Binary_Crossentropy
Optimiseur (Optimizer)	Adam
Taille de lots (batch size)	512
No. d'époques (No. of epochs)	5

Tableau 4-3 : Hyperparamètres du modèle CNN dans ISOT

Nom de l'hyperparamètre	Type ou valeur
Dimension d'embedding	300
Taille de kernel	3
Régularisateur L1	0.001
Régularisateur L2	0.01
Dropout	[0.3,0.4]
Fonction d'activation	Sigmoïde
Fonction de perte (Loss Function)	Binary_Crossentropy
Optimiseur (Optimizer)	Adam
Taille de lots (batch size)	512
No. d'époques (No. of epochs)	5

Tableau 4-4 : Hyperparamètres du modèle CNN dans GossipCop

2.4.Paramétrage du modèle BERT

Les hyperparamètres du modèle BERT ont été soigneusement ajustés pour optimiser ses performances sur les ensembles de données ISOT et GossipCop, comme indiqué dans les tableaux 4-5 et 4-6.

La dimension d'embedding détermine la taille de l'espace vectoriel dans lequel le texte est représenté, influençant la richesse des informations capturées par le modèle. La fonction

d'activation choisie, Sigmoïde, aide le modèle à générer des probabilités pour les différentes classes, facilitant la classification binaire entre vraies et fausses nouvelles. La fonction de perte utilisée, CrossEntropyLoss, permet d'évaluer la performance du modèle en mesurant l'écart entre les prédictions et les étiquettes réelles, ce qui guide l'optimisation. L'optimiseur AdamW est sélectionné pour ajuster les poids du modèle en réduisant la fonction de perte tout en gérant le poids de désintégration, ce qui est essentiel pour la régularisation. Le taux d'apprentissage contrôle la vitesse à laquelle le modèle met à jour ses poids pour minimiser la perte. La taille de lots définit le nombre d'exemples utilisés par itération, impactant la stabilité et la rapidité de l'entraînement. Enfin, le nombre d'époques indique le nombre total de passages sur les données d'entraînement, influençant la capacité du modèle à généraliser aux nouvelles données sans surapprentissage.

Nom de l'hyperparamètre	Type ou valeur
Dimension d'embedding	768
Fonction d'activation	Sigmoïde
Fonction de perte (Loss Function)	CrossEntropyLoss
Optimiseur (Optimizer)	AdamW
Taux d'apprentissage (Learning rate)	1e-5
Taille de lots (batch size)	64
No. d'époques (No. of epochs)	10

Tableau 4-5 : Hyperparamètres du modèle BERT dans ISOT

Nom de l'hyperparamètre	Type ou valeur
Dimension d'embedding	768
Fonction d'activation	Sigmoïde
Fonction de perte (Loss Function)	CrossEntropyLoss
Optimiseur (Optimizer)	AdamW
Taux d'apprentissage (Learning rate)	1e-5
Taille de lots (batch size)	64
No. d'époques (No. of epochs)	10

Tableau 4-6 : Hyperparamètres du modèle BERT dans GossipCop

3. Résultats et discussions

Dans cette section, les résultats issus de l'application des modèles Bi-LSTM, CNN et BERT aux jeux de données ISOT et GossipCop sont présentés. Les performances de chaque modèle sont développées en termes de l'exactitude, la précision, le rappel et le score F1.

3.1.Visualisation des données

La visualisation des données est une étape cruciale pour comprendre la distribution et la structure des données textuelles dans nos ensembles. Avant de procéder à la formation du modèle, il est essentiel d'explorer visuellement les données afin de comprendre leurs caractéristiques clés.

3.1.1. Ensemble de données ISOT

L'ensemble de données ISOT[1] se compose de 45.000 articles d'actualité, presque également répartis entre les catégories vrai et faux.
Les vrais articles ont été extraits du site de Reuters; Les faux articles provenant de diverses sources ont été signalés comme étant des fausses sources par Wikipédia et Politifact. Les ensembles de données ont inclus le corps complet de chaque article, le titre, la date ainsi que le sujet. Les principaux sujets de cet article sont la politique, les nouvelles du monde et les dates sont compris entre 2015 et 2018. La répartition par nombre et par type est affichée dans le tableau 4-7.

News	Nombre totale des articles	Type	Nombre des articles
Réel	21417	World news	10,145
		Politics news	11,272
Faux	23481	Government news	1570
		Middle east	778
		US news	783
		Left news	4459
		Politics	6841
		News	9050

Tableau 4-7 : La répartition de l'ensemble de données ISOT.

Le graphique à barres illustré dans la figure 4-3 permet d'apprécier la répartition des articles en fonction de leurs sujets et de leur statut de véracité (vrai ou faux).

- PoliticsNews et WorldNews sont majoritairement des vraies nouvelles, au-delà de 10 000 articles chacun.
- News affiche une répartition assez équilibrée entre vraies et fausses nouvelles.
- Politics, Government News, Left-news et US News contiennent moins d'articles et affichent un nombre significatif de faux articles pour certains d'entre eux.

[1] ISOT Fake News Dataset., 2017 Available at: https://www.uvic.ca/engineering/

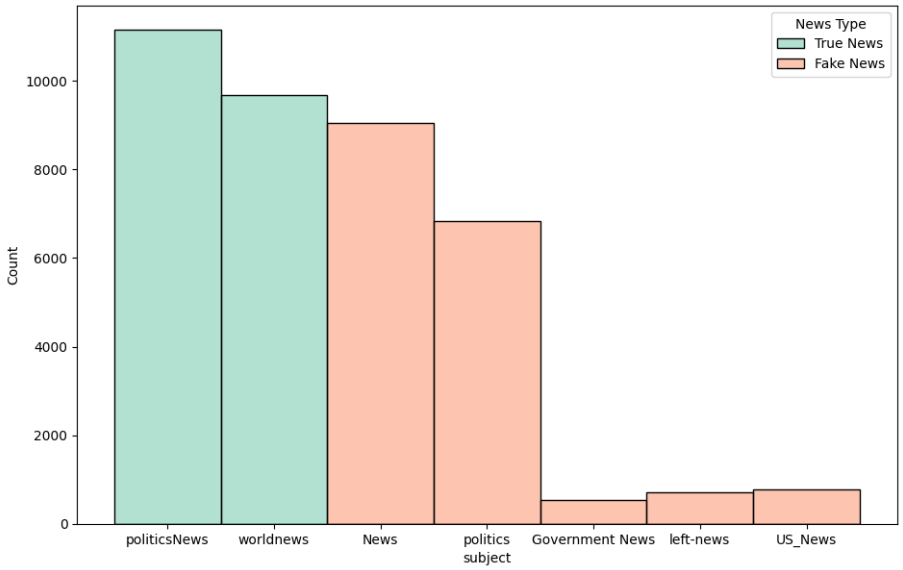

Figure 4-3 : Répartition des articles des nouvelles selon la véracité

Le graphique illustré dans la figure 4-4 représente l'évolution dans le temps du nombre de vraies et fausses nouvelles entre 2015 et 2018. On remarque que les articles (vrais et faux) sont en constante augmentation depuis 2015. Cette augmentation est particulièrement importante après 2017, avec un pic prononcé entre septembre 2017 et janvier 2018. Ceci mettrait en avant une tendance à une production plus importante de nouvelles avec en grande partie des articles de vraies nouvelles publiées durant cette période-là.

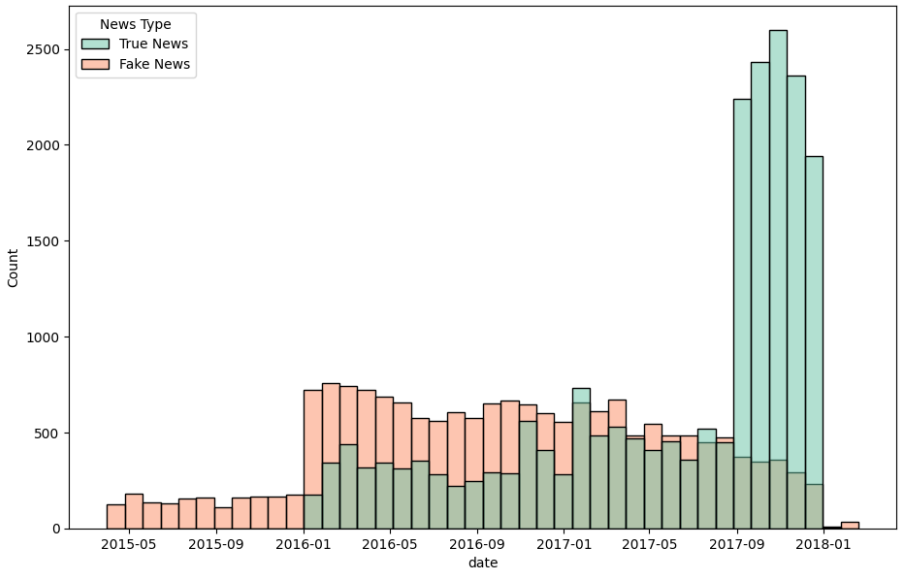

Figure 4-4 : Distribution temporelle des articles selon la véracité

Le graphique affiché dans la figure 4-5 est un diagramme à barres, dont le but est de faire apparaître la longueur moyenne des titres mesurée en caractères pour deux types d'articles. On constate que les titres des fausses nouvelles affichent des titres plus longs en moyenne avec une longueur d'environ 90 caractères que pour les vraies nouvelles dont les titres sont plus courts avec une moyenne d'environ 65 caractères. Une telle

différence interpelle puisque cette tendance serait à mettre en lien avec une façon exhaustive d'élaborer les titres des fausses nouvelles, permettant alors d'attirer l'attention ou encore plus, de manipuler l'audience.

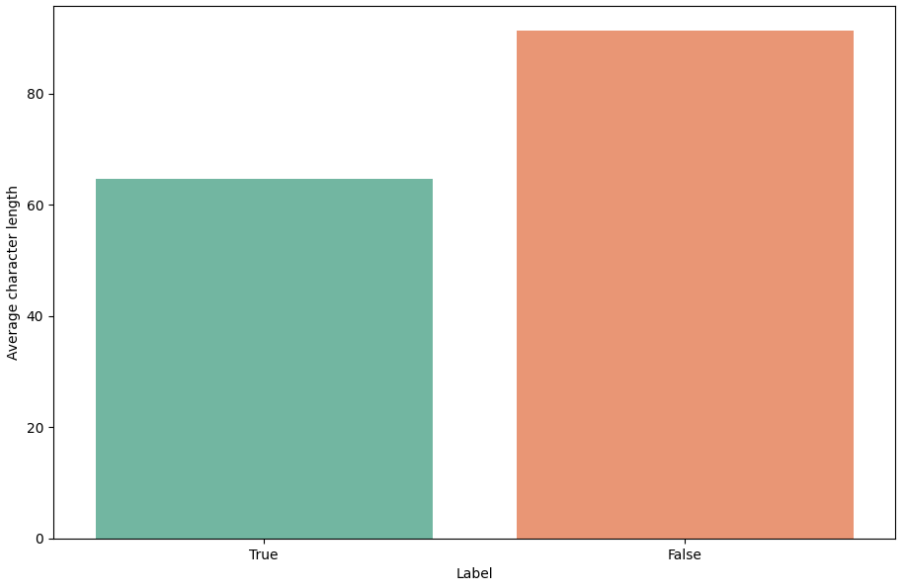

Figure 4-5 : Comparaison de la longueur moyenne des titres entre les vraies et fausses nouvelles dans ISOT

Comme montre la figure 4-6, c'est un nuage de mots (word cloud) dans lequel fréquence des mots rencontrés dans un corpus de texte est représentée. Les mots les plus importants sont en plus gros caractères et en gras. Dans ce nuage, nous remarquons plusieurs termes qui peuvent être identifiés comme particulièrement comptables : « Trump », « U.S. », « say », « president », « video », « republican », « Obama », « Russia », « House » évoquant des articles probablement consacrés à Donald Trump, à sa présidence, aux élections aux États-Unis, à la Russie, à des vidéos et à des prises de paroles politiques.

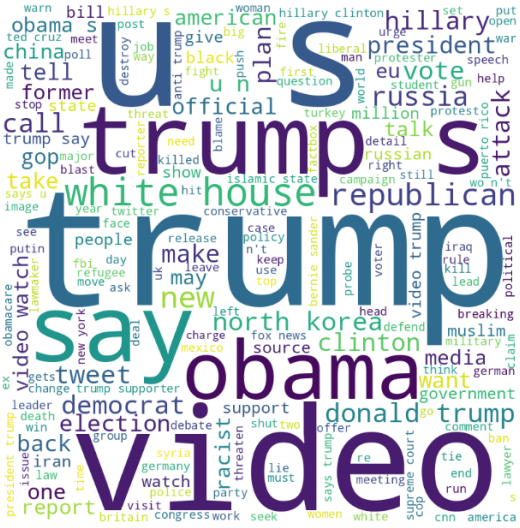

Figure 4-6 : Nuage de mots de l'ensemble de données ISOT

3.1.2. Ensemble de données FakeNewsNet

Cet ensemble de données complet (https://github.com/KaiDMML/FakeNewsNet, consulté le 18 octobre 2024) a été construit en rassemblant des informations provenant de deux sites web de vérification des faits pour obtenir du contenu d'actualités sur les fausses nouvelles et les vraies nouvelles telles que PolitiFact et GossipCop.

Dans PolitiFact, les journalistes et les experts du domaine passent en revue les nouvelles politiques et fournissent des résultats d'évaluation de vérification des faits pour prétendre que les articles sont faux ou réels. Dans GossipCop, les histoires de divertissement, provenant de divers médias, sont évaluées par une note sur une échelle de 0 à 10 comme le degré de faux au réel.

L'ensemble de données contient environ 900 nouvelles politiques et 20 000 nouvelles de ragots comme montrées dans le tableau 4-8.

Ensemble de données	Faux	Réel	global
PolitiFact	161	205	366
GossipCop	4,927	16,693	21,620

Tableau 4-8: Les statistiques de l'ensemble de données FakeNewsNet

Comme dit ci-dessus, FakeNewsNet a deux sous-ensembles, GossipCop et Politifact.com. Nous avons choisi de travailler sur les données de GossipCop, qui sont moins violentes, parce que bien que les ragots soient moins sérieux que la politique, ils sont également un acteur de la désinformation sur les réseaux sociaux. En effet, les nouvelles entourant les célébrités sont facilement diffusées et, par conséquent, peuvent rapidement modifier l'opinion publique et son expérience, changeant à terme la perception des événements ou des individus entre les gens. Il est donc primordial de comprendre et d'identifier la désinformation dans ce secteur afin d'assurer une information plus exacte, même en situation non politique.

De plus, notre premier ensemble de données, ISOT, étant axé sur des contenus politiques, l'étude de GossipCop nous permet d'introduire une diversité dans notre recherche, en analysant des types de désinformation dans des contextes différents.

La table illustré dans la figure 4-7 fournit un instantané de plusieurs enregistrements du jeu de données GossipCop. Elle comprend des colonnes comme ID, News_URL, Title et Tweet_IDs.

- ID semble être un identificateur unique pour chaque article de nouvelles, tandis que News_URL donne le lien source.
- Title contient le titre ou la description de l'article, et Tweet_IDs énumère les ID des tweets liés à l'article.

	id	news_url	title	tweet_ids
0	gossipcop-882573	https://www.brides.com/story/teen-mom-jenelle-...	Teen Mom Star Jenelle Evans' Wedding Dress Is ...	912371411146149888\t912371528343408641\t912372...
1	gossipcop-875924	https://www.dailymail.co.uk/tvshowbiz/article-...	Kylie Jenner refusing to discuss Tyga on Life ...	901989917546426369\t901989992074969089\t901990...
2	gossipcop-894416	https://en.wikipedia.org/wiki/Quinn_Perkins	Quinn Perkins	931263637246881792\t931265332022579201\t931265...
3	gossipcop-857248	https://www.refinery29.com/en-us/2018/03/19192...	I Tried Kim Kardashian's Butt Workout & Am For...	868114761723936769\t868122567910936576\t868128...
4	gossipcop-884684	https://www.cnn.com/2017/10/04/entertainment/c...	Celine Dion donates concert proceeds to Vegas ...	915528047004209152\t915529285171122176\t915530...
...	...	...	...	...
20857	gossipcop-6702260693	www.huffingtonpost.com/2012/09/11/september-11...	September 11: Celebrities Remember 9/11 (TWEETS)	245643768638894080
20858	gossipcop-6051845337	www.dailymail.co.uk/news/article-4915674/NASCA...	NASCAR owners threaten to fire drivers who pro...	912048333413330944\t912048571482087424\t912049...
20859	gossipcop-2435526162	www.telegraph.co.uk/men/the-filter/7-signs-dav...	The 7 signs that David Beckham is definitely h...	897794716447539200\t897804460830928896\t897842...
20860	gossipcop-4576152851	www.vanityfair.com/style/2016/09/ryan-gosling-...	Ryan Gosling and Eva Mendes Did Not Get Marrie...	778678901572710400\t778681718714740736\t778683...
20861	gossipcop-919334865	www.lifeandstylemag.com/posts/jamie-foxx-katie...	Jamie Foxx Spends the Day With Katie Holmes an...	913137595424608258\t913139996059717632\t913146...

Figure 4-7 : Ensemble de données GossipCop

Dans l'histogramme de la figure 4-8, la longueur moyenne des titres, exprimée en nombre de caractères, semble être très proche pour les articles « Vrai » et « Faux » dont les différences sont très faibles. Cela suggère que la longueur d'un article n'est pas un indicateur fort de sa véracité, car les vraies et fausses nouvelles peuvent avoir des articles de même longueur. Cette information peut aider à se concentrer sur d'autres caractéristiques pour distinguer les fausses nouvelles, comme l'utilisation des sentiments exprimées dans les titres des nouvelles.

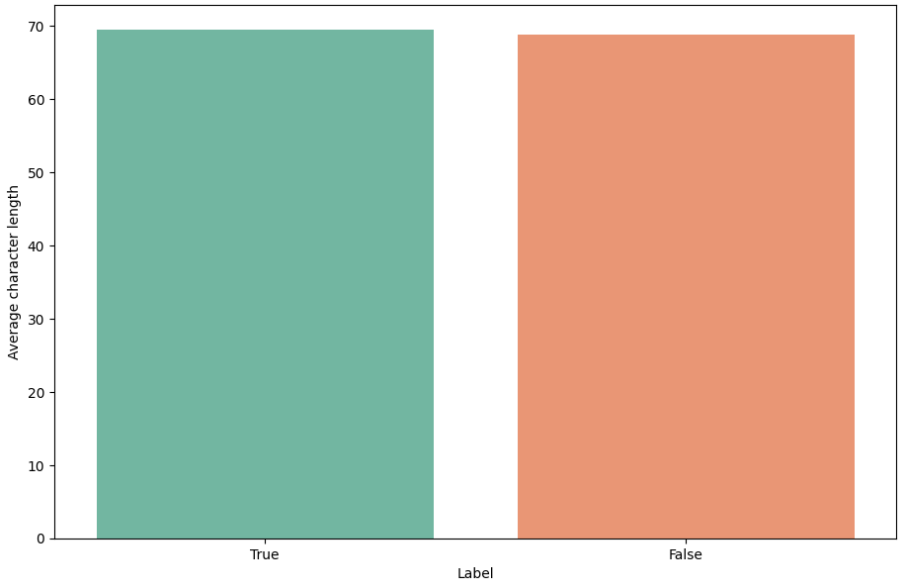

Figure 4-8 : Longueur moyenne des titres pour les vraies et fausses nouvelles dans GossipCop

Le nuage de mots illustré par la Figure 4-9 montre les mots qui apparaissent fréquemment dans les titres des nouvelles dans l'ensemble de données. Les termes les plus importants sont en plus gros et en gras , comme "star", "now", "say", "reveal", et des noms de célébrités tel que "Kim Kardashian", "Meghan Markle", etc.

Le nuage montre l'important poids des termes concernant des stars et les verbes d'énonciation relatives à des révélations, des annonces, ou des évènements actuels (« watch », « reveal », « celebrate », « say »), ce qui témoigne sans doute, l'accent des fausses nouvelles sur des histoires de célébrités, souvent liées à des scandales ou à des événements sensationnels.

Figure 4-9 : Nuage de mots de l'ensemble de données GossipCop

3.2.Prétraitement de texte

Le prétraitement est une étape essentielle qui permet de transformer les données textuelles brutes en un format compatible avec les modèles d'apprentissage. Les opérations qui font intervenir ce processus sont citées dans le chapitre précédent.

Celles-ci ont pour effet d'améliorer la qualité et la cohérence des données d'entrée aux modèles. Ceci permet donc de réduire le bruit et de favoriser la focalisation sur les informations pertinentes.

L'étape de prétraitement de texte a été appliquée de manière identique aux deux ensembles de données. La figure 4-10 représente les données de GossipCop avant le prétraitement de texte et la figure 4-11 les données après le prétraitement.

Dans la figure 4-10, les mots communs comme "s", "and", "to", "is", et "the" sont très fréquents, mais ce sont souvent des mots vides (StopWords) qui n'apportent pas beaucoup de valeur pour l'analyse sémantique.

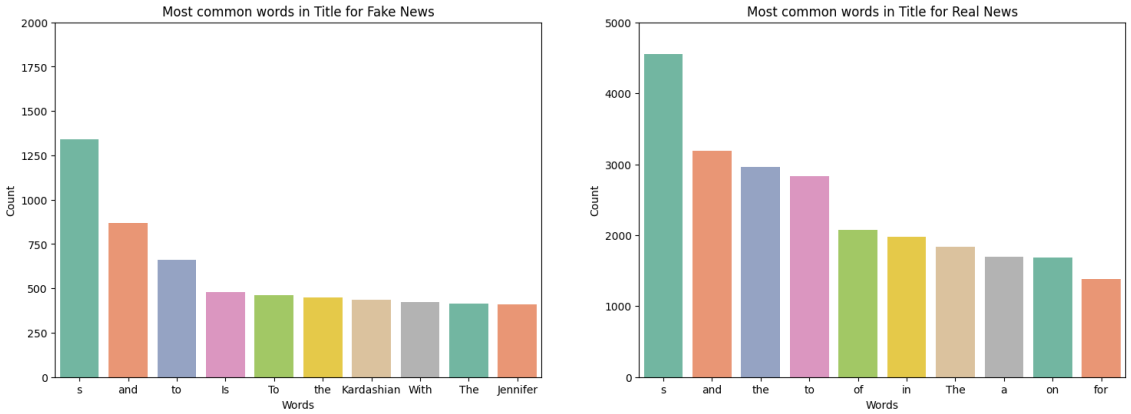

Figure 4-10 : Les mots fréquents dans les titres des vraies et fausses nouvelles avant le prétraitement de texte dans GossipCop

Après le prétraitement, comme illustré dans la figure 4-11, ces mots vides ont été supprimés, et des mots plus spécifiques apparaissent, comme des noms de célébrités (Jennifer, Kardashian, Brad, etc.) ou des termes clés liés à des événements (Awards, Season, Kim). Cela permet de mieux identifier les thématiques principales des titres.

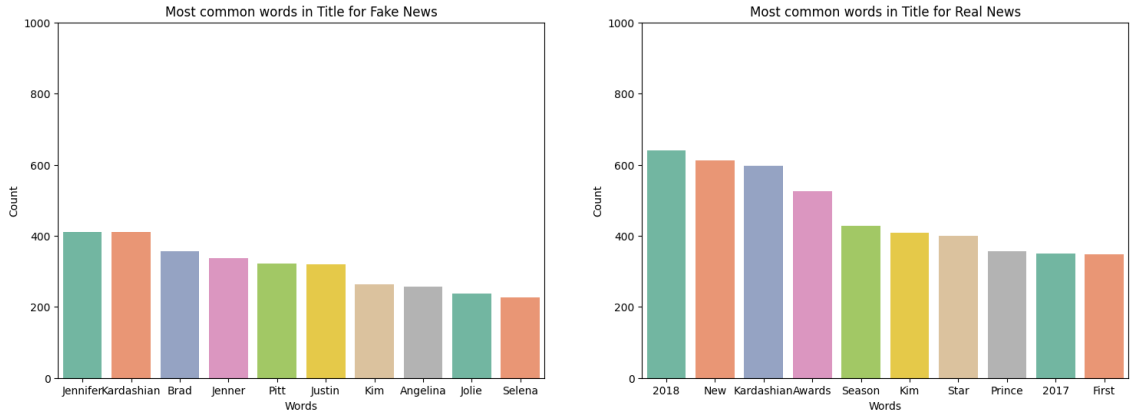

Figure 4-11 : Les mots fréquents dans les titres des vraies et fausses nouvelles après le prétraitement de texte dans GossipCop

3.3.Analyse des sentiments

La bibliothèque TextBlob a été exploitée pour faire ressortir ce score de sentiment pour les titres d'articles de nouvelles. Celui-ci peut varier entre -1 correspondant à un sentiment extrêmement négatif et 1 à un sentiment extrêmement positif. La figure 4-12 illustre la colonne « polarity » qui révèle pour les divers titres de nouvelles de l'ensemble de données ISOT le score de sentiment déjà extrait.

Pour les modèles Bi-LSTM et CNN, le résultat de la polarité est combiné directement avec le classificateur pour détecter les fausses nouvelles.

	text	subject	title	polarity
0	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	'As U.S. budget fight looms , Republicans flip...	0.0
1	WASHINGTON (Reuters) - Transgender people will...	politicsNews	' U.S . military accept transgender recruits M...	-0.1
2	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	'Senior U.S. Republican senator : Let Mr. Muel...	0.0
3	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	'FBI Russia probe helped Australian diplomat t...	0.0
4	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	'Trump wants Postal Service charge much Amazon...	0.2

Figure 4-12 : analyse des sentiments avec TextBlob dans ISOT

Pour le modèle BERT, quelques étapes doivent être suivies avant de combiner les deux unités, à savoir l'analyse des sentiments et la classification.

Le tableau illustré dans la figure 4-13 représente la première étape suivant le calcul de la polarité, où chaque article se voit attribuer un score de polarité par l'outil TextBlob pour déterminer s'il est porteur d'un sentiment positif, neutre ou négatif. Ces scores permettent de catégoriser les articles en fonction de leur tonalité : une polarité positive (> 0) indique un sentiment positif, une polarité négative (< 0) traduit un sentiment négatif, et une polarité proche de zéro suggère un sentiment neutre. Cette classification qualitative fournit une première évaluation de la tonalité des articles.

	text	subject	title	polarity	sentiment
0	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	U budget fight looms Republicans flip fiscal s...	0.00	neutre
1	WASHINGTON (Reuters) - Transgender people will...	politicsNews	U military accept transgender recruits Monday ...	-0.10	negative
2	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	Senior U Republican senator Let Mr Mueller job	0.00	neutre
3	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	FBI Russia probe helped Australian diplomat ti...	0.00	neutre
4	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	Trump wants Postal Service charge much Amazon ...	0.20	positive
...	...	...	...	...	...
38713	21st Century Wire says All the world s a stage...	US_News	White House Theatrics Gun Control	0.00	neutre
38714	Randy Johnson 21st Century WireThe majority ...	US_News	Activists Terrorists Media Controls Dictates N...	0.00	neutre
38715	Tune in to the Alternate Current Radio Network...	US_News	BOILER ROOM Surrender Retreat Heads Roll EP 38	0.00	neutre
38716	21st Century Wire says A new front has just op...	US_News	Federal Showdown Looms Oregon BLM Abuse Local ...	0.00	neutre
38717	21st Century Wire says It s not that far away....	US_News	Troubled King Chicago Rahm Emanuel Desperate S...	-0.55	negative

Figure 4-13 : Classification des sentiments dans ISOT

Le graphe illustré dans la figure 4-14 présente la distribution des sentiments pour les vraies et fausses nouvelles

la majorité des nouvelles réelles présentent un sentiment neutre et son effectif est plus fort que celui des nouvelles fausses. On peut donc soutenir qu'elles prennent la forme d'informations plutôt que d'expressions de sentiments positifs ou négatifs.

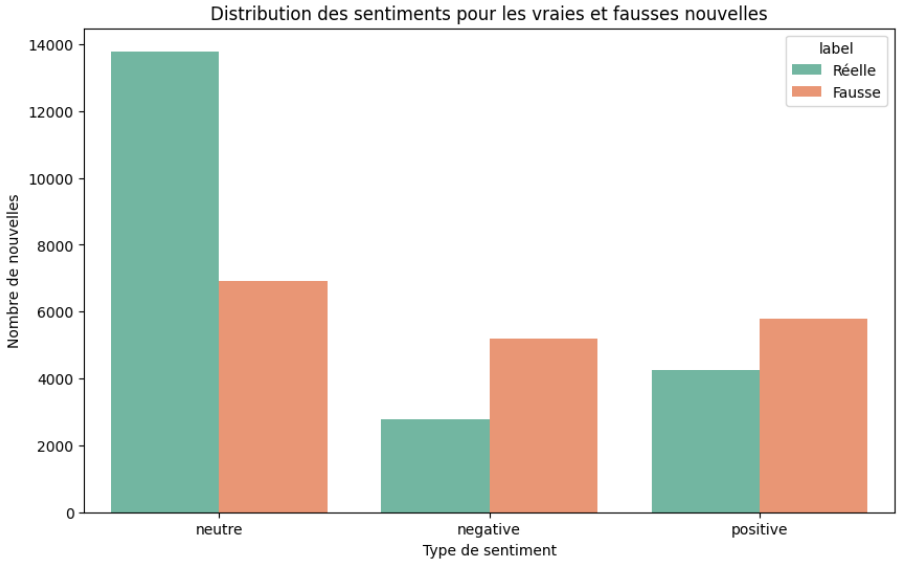

Figure 4-14 : Distribution des sentiments pour les vraies et fausses nouvelles dans ISOT

Les nouvelles fausses affichent une part plus importante de sentiments négatifs que les nouvelles authentiques. Cela pourrait montrer que les informations erronées sont souvent conçues de telle sorte qu'elles provoquent des émotions négatives telles que peur, colère ou méfiance et attirent ainsi d'autant plus d'attention. Les nouvelles fausses et réelles présentent tout de même toutes deux quelques nouvelles à sentiment positif, mais leur nombre est sensiblement inférieur à celui des nouvelles à sentiment neutre.

Cette distribution peut refléter une stratégie dans les fausses nouvelles d'exploiter les émotions pour capter l'attention et influencer le lecteur, contrairement aux vraies nouvelles qui ont tendance à présenter les informations de manière plus équilibrée.

Le tableau se présentant dans la figure 4-15, représente ensuite la seconde étape de transformation des sentiments en représentations numériques, facilitant leur intégration dans le modèle BERT. Les sentiments positifs et neutres sont représentés par 1 et les sentiments négatifs sont représentés par 0. Cette conversion permet de simplifier la représentation des sentiments pour que le modèle BERT puisse les utiliser comme une composante supplémentaire dans ses représentations textuelles.

En fusionnant, d'une part, les informations textuelles et, d'autre part, ces indicateurs de sentiment, le modèle apparait à même de mieux appréhender la tonalité émotionnelle standard de chacun des articles, afin d'optimiser sa capacité à discriminer des nouvelles vraies des nouvelles fausses.

	text	subject	title	polarity	sentiment
0	WASHINGTON (Reuters) - The head of a conservat...	politicsNews	U budget fight looms Republicans flip fiscal s...	0.00	1
1	WASHINGTON (Reuters) - Transgender people will...	politicsNews	U military accept transgender recruits Monday ...	-0.10	0
2	WASHINGTON (Reuters) - The special counsel inv...	politicsNews	Senior U Republican senator Let Mr Mueller job	0.00	1
3	WASHINGTON (Reuters) - Trump campaign adviser ...	politicsNews	FBI Russia probe helped Australian diplomat ti...	0.00	1
4	SEATTLE/WASHINGTON (Reuters) - President Donal...	politicsNews	Trump wants Postal Service charge much Amazon ...	0.20	1
...	...	...	...	...	...
38713	21st Century Wire says All the world s a stage...	US_News	White House Theatrics Gun Control	0.00	1
38714	Randy Johnson 21st Century WireThe majority ...	US_News	Activists Terrorists Media Controls Dictates N...	0.00	1
38715	Tune in to the Alternate Current Radio Network...	US_News	BOILER ROOM Surrender Retreat Heads Roll EP 38	0.00	1
38716	21st Century Wire says A new front has just op...	US_News	Federal Showdown Looms Oregon BLM Abuse Local ...	0.00	1
38717	21st Century Wire says It s not that far away....	US_News	Troubled King Chicago Rahm Emanuel Desperate S...	-0.55	0

Figure 4-15 : Conversion des sentiments en valeurs numériques

3.4.Classification de texte

Cette sous-section présente la performance de chaque modèle dans la tâche de classification des fausses nouvelles. Les performances sont évaluées à l'aide de plusieurs métriques, notamment l'exactitude, la précision, le rappel et le score F1.

3.4.1. Résultats de classification pour l'ensemble de données ISOT

Les graphiques ci-dessous illustrent les performances d'un modèle Bi-LSTM appliqué à la détection de fausses nouvelles.

La figure 4-16 montre l'évolution de l'exactitude et de la perte pour l'entraînement et la validation du modèle Bi-LSTM au fil des époques. On remarque que le modèle apprend bien : l'exactitude en entraînement grimpe presque jusqu'à 99 %, tandis que celle en validation se stabilise autour de 98 %. En parallèle, les pertes pour l'entraînement et la validation diminuent progressivement, ce qui reflète une amélioration continue des performances du modèle sur les données de test.

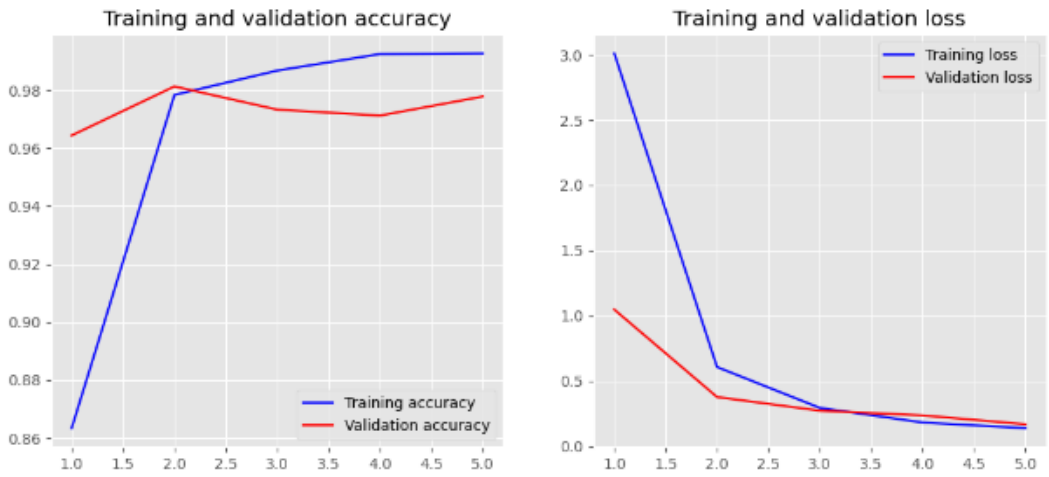

Figure 4-16 : Performance d'entraînement et de validation du modèle Bi-LSTM dans ISOT

La matrice de confusion dans la figure 4-17 nous donne un aperçu des performances du modèle. Celui-ci a correctement identifié 4081 échantillons de fausses nouvelles et 3491 échantillons de vraies nouvelles. Les erreurs de classification sont faibles : seulement 91 fausses nouvelles ont été incorrectement identifiées comme vraies, et 82 vraies nouvelles ont été classées comme fausses.

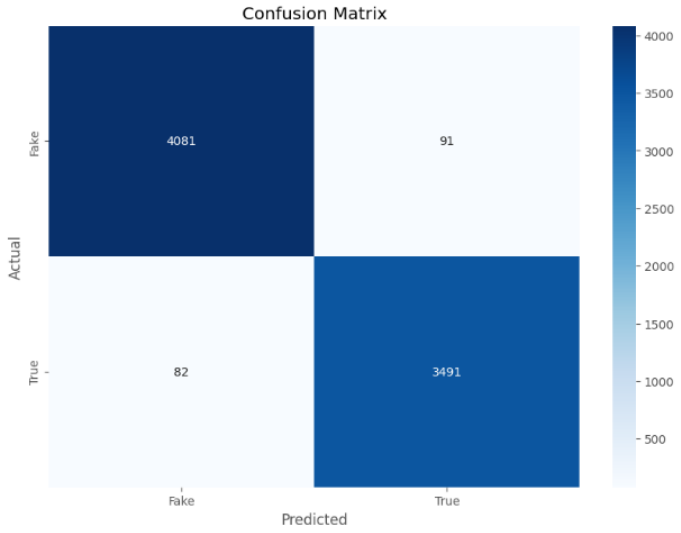

Figure 4-17 : Matrice de confusion du modèle Bi-LSTM dans ISOT

Les graphiques ci-dessous illustrent les performances d'un modèle CNN appliqué à la détection de fausses nouvelles.

Comme on peut le voir dans la figure 4-18, l'exactitude d'entraînement progresse rapidement, atteignant presque 99 %. De son côté, l'exactitude de validation se stabilise autour de 98 %. Quant à la perte en entraînement et en validation, elle diminue fortement dès la première époque pour se rapprocher progressivement de zéro au fil des époques.

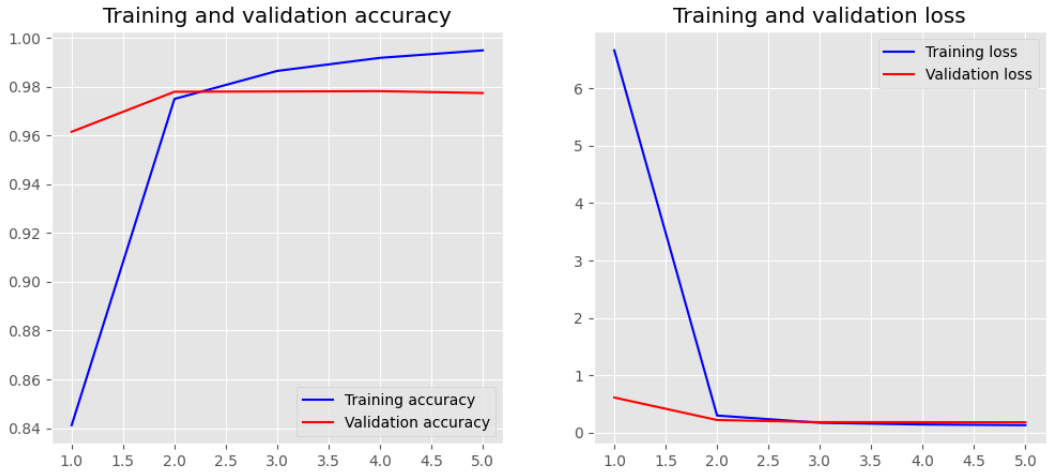

Figure 4-18 : Performance d'entraînement et de validation du modèle CNN dans ISOT

La matrice de confusion présenté dans la figure 4-19 montre que le modèle classifie correctement la plupart des fausses et vraies nouvelles, avec seulement 66 fausses nouvelles mal classées comme vraies et 109 vraies nouvelles mal classées comme fausses, ce qui est faible par rapport au nombre total de prédictions.

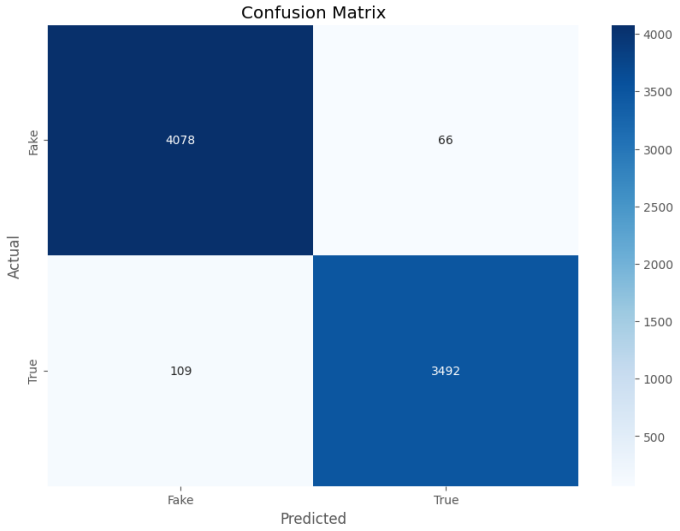

Figure 4-19 : Matrice de confusion du modèle CNN dans ISOT

Les graphiques ci-dessous illustrent les performances d'un modèle BERT appliqué à la détection de fausses nouvelles.

Le modèle BERT entraîné sur l'ensemble de données ISOT montre d'excellentes performances de classification, comme le révèlent les courbes de perte et de précision présentées dans la figure 4-20. Dès les premières époques, l'exactitude tant pour l'entraînement que pour la validation grimpe rapidement à environ 95 %, avant de continuer à progresser lentement jusqu'à presque 100 %. De plus, les courbes de perte diminuent de façon continue pour l'entraînement comme pour la validation, et la différence entre elles reste faible.

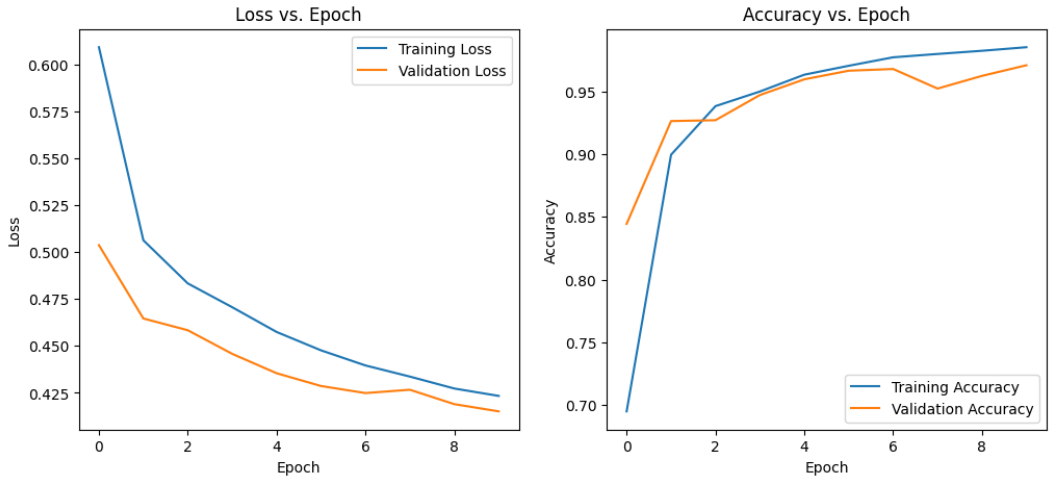

Figure 4-20 : Performance d'entraînement et de validation du modèle BERT dans ISOT

La matrice de confusion dans la figure 4-21 montre également une prédiction précise, avec un seul échantillon mal classé (un faux négatif) et aucun faux positif, ce qui traduit une grande précision et un bon rappel.

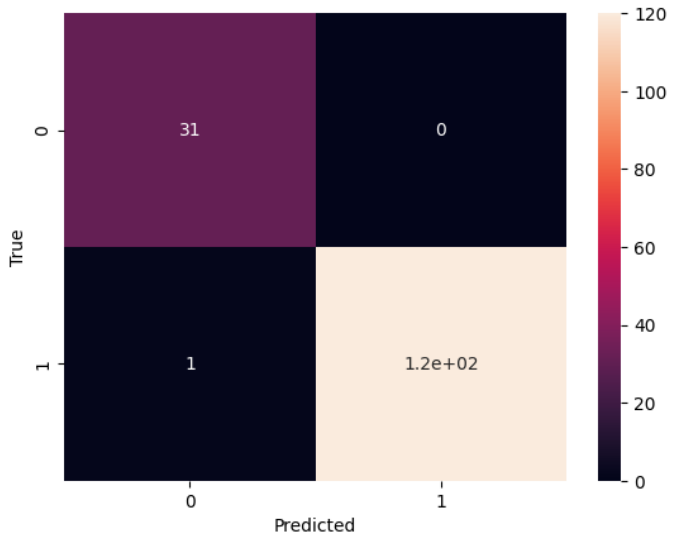

Figure 4-21 : Matrice de confusion du modèle BERT dans ISOT

Les performances des trois modèles CNN, Bi-LSTM et BERT montrent une amélioration notable en termes d'exactitude et de scores d'évaluation, avec le modèle BERT surpassant les deux autres.

Comme le montre le tableau 4-9, le modèle Bi-LSTM atteint une exactitude de 97.76% et un score F1 de 97.58%. Le modèle CNN présente des performances similaires avec une exactitude de 97.74 %, une précision de 98.14 %, un rappel de 96.97 % et un score F1 de 97.55 %, soulignant son efficacité à identifier des motifs locaux dans les données.

Cependant, c'est BERT qui se démarque vraiment avec une exactitude remarquable de 99,34 %, une précision de 100 % et un score F1 de 99,58 %. Ces résultats mettent en

lumière les capacités de BERT en tant que modèle de langage pré-entraîné, particulièrement performant pour la classification de textes grâce à sa capacité à saisir les dépendances contextuelles et les nuances sémantiques.

Modèles	Exactitude (Accuracy)	Précision (Precision)	Rappel (Recall)	Score F1
CNN	97.74%	98.14%	96.97%	97.55%
Bi-LSTM	97.76%	97.45%	97.70%	97.58%
BERT	99.34%	100%	99.17%	99.58%

Tableau 4-9 : Performance des modèles sur l'ensemble de données ISOT

3.4.2. Résultats de classification pour l'ensemble de données GossipCop

Les graphiques montrent les performances d'un modèle Bi-LSTM sur les données d'entraînement et de validation, ainsi que sa matrice de confusion.

Dans le graphique qui est présenté dans la figure 4-22, la courbe de l'exactitude croît régulièrement atteignant environ 90 % pour l'entraînement et 84 % pour la validation au bout de cinq époques. La perte, est pour sa part, rapidement descendue pour l'entraînement et la validation pour se stabiliser ensuite à des valeurs proches de zéro, ce qui témoigne de bonne convergence du modèle.

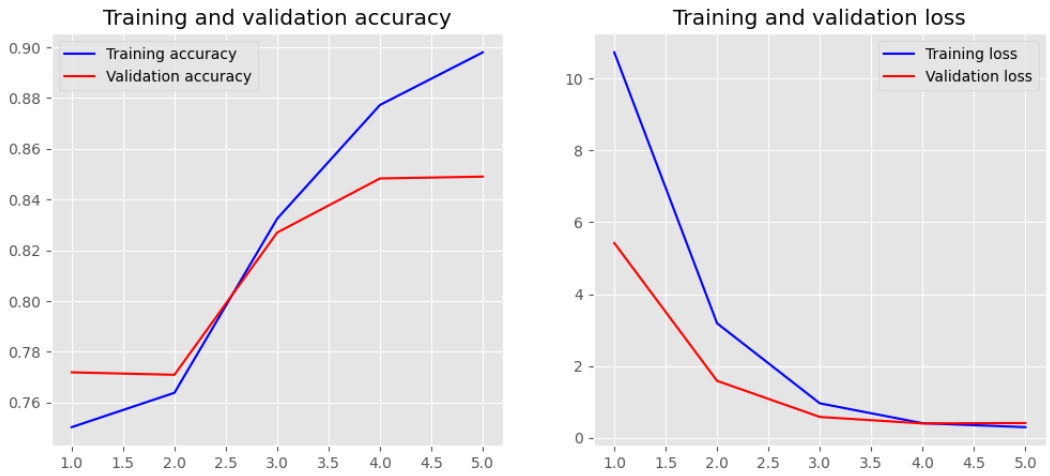

Figure 4-22 : Performance d'entraînement et de validation du modèle Bi-LSTM dans GossipCop

Les résultats du modèle de classification Bi-LSTM sont montrés dans la figure 4-23, sous la forme d'une matrice de confusion. Ainsi, le modèle a correctement identifié 3 027 fausses nouvelles et 514 vraies nouvelles. Il est à noter le nombre considérable d'erreurs de classification puisque 194 fausses nouvelles ont été erronément classées comme vraies ainsi que 436 vraies nouvelles comme fausses.

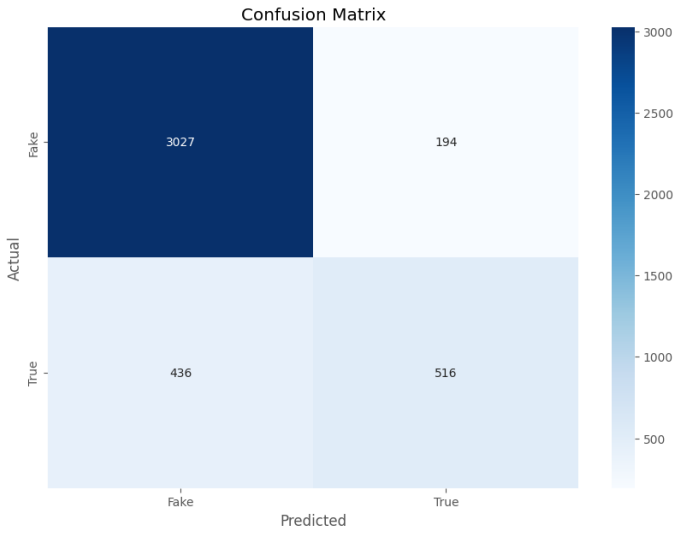

Figure 4-23 : Matrice de confusion du modèle Bi-LSTM dans GossipCop

Dans les deux graphes ci-dessous, nous observons les performances du modèle CNN pour la classification de fausses nouvelles.

La figure 4-24 représente l'évolution de la performance d'exactitude et de perte au cours de l'entraînement et de la validation sur cinq époques. L'exactitude croît de façon significative atteignant environ 84 % pour l'entraînement et pour la validation. De même, La perte décroît régulièrement montrant que le modèle apprend de mieux en mieux sans signe de surapprentissage.

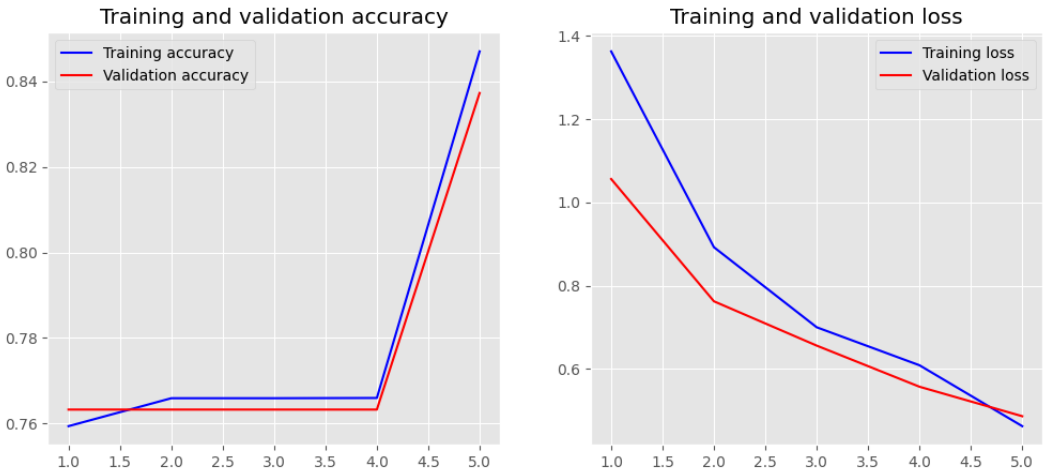

Figure 4-24 : Performance d'entraînement et de validation du modèle CNN dans GossipCop

Le graphique exposé sous la figure 4-25 est une matrice de confusion qui illustre la capacité du modèle à distinguer les vraies informations des fausses. On remarque que le modèle permet une bonne identification des cas de fausses informations, 3039 prédictions exactes, et 455 prédictions exactes pour les vraies nouvelles. Cependant le nombre d'erreurs de classification est en effet considérable puisque 146 fausses nouvelles sont classées à tort comme vraies et 533 nouvelles vraies comme fausses.

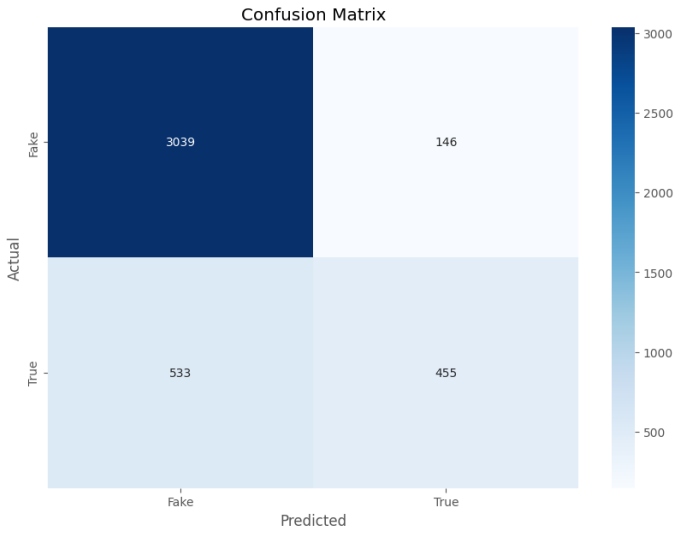

Figure 4-25 : Matrice de confusion du modèle CNN dans GossipCop

Dans la figure 4-26, le graphique montre les courbes de perte et de l'exactitude au cours des époques pour l'entraînement et la validation du modèle BERT. La perte diminue de manière régulière, suggérant une amélioration continue de l'apprentissage du modèle. L'exactitude d'entraînement atteint presque 100%, tandis que l'exactitude de validation stagne autour de 98%, indiquant une bonne généralisation avec un faible surapprentissage.

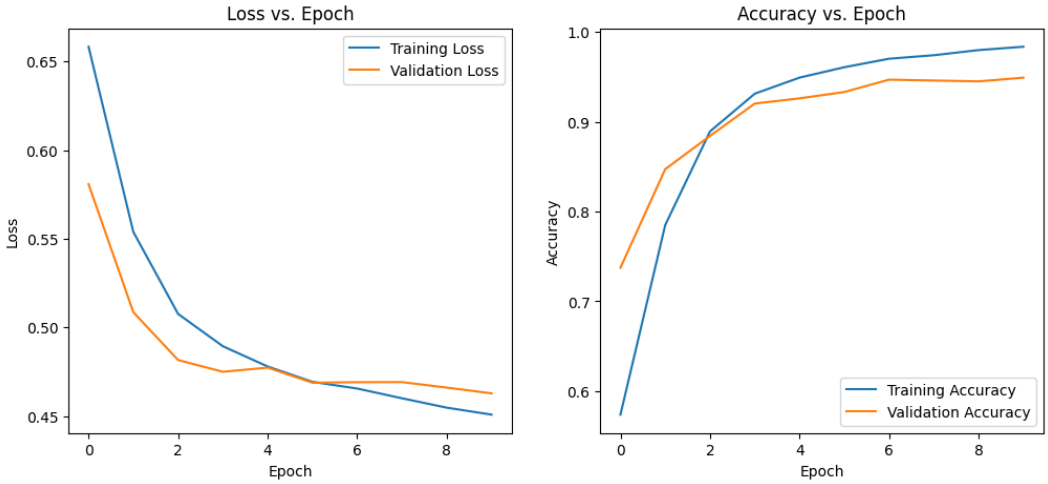

Figure 4-26 : Performance d'entraînement et de validation du modèle BERT dans GossipCop

La figure 4-27 illustre une matrice de confusion qui montre la performance du modèle BERT dans la détection des fausses nouvelles. Elle indique que le modèle a correctement classé 10 exemples des vraies nouvelles et 70 exemples des fausses nouvelles, avec seulement 2 erreurs de classification, 0 fausse nouvelle détectée comme vraie et 2 vraies nouvelles détectées comme fausses.

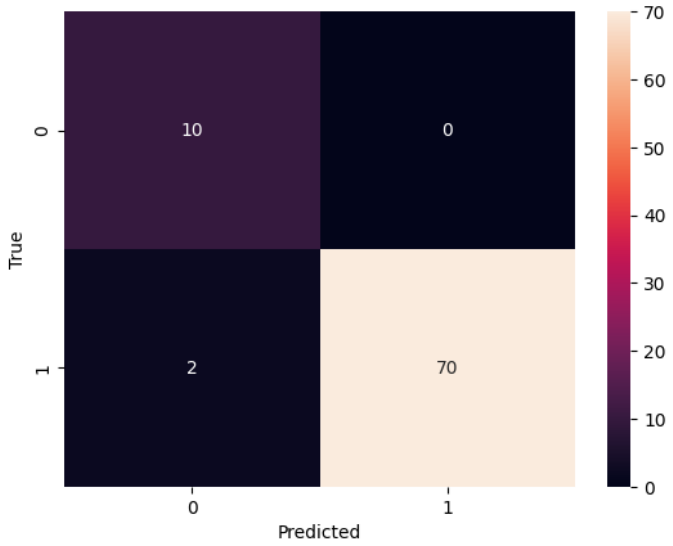

Figure 4-27 : Matrice de confusion du modèle BERT dans GossipCop

Le tableau 4-10 présente les performances des modèles CNN, Bi-LSTM et BERT pour la classification des fausses nouvelles sur l'ensemble de données GossipCop, en utilisant l'exactitude, la précision, le rappel et le score F1 comme métriques d'évaluation.

Le modèle BERT surpasse largement les autres, avec une exactitude de 97.56% et des valeurs de précision, rappel et score F1 extrêmement élevées (100%, 97.22%, et 98.59% respectivement). Ces résultats montrent que BERT est particulièrement efficace pour distinguer les fausses nouvelles, capturant plus fidèlement les bonnes prédictions par rapport aux autres modèles.

En revanche, les modèles Bi-LSTM et CNN affichent des performances similaires mais inférieures, avec des exactitudes respectives de 84.90 % et 83.72 %. Le Bi-LSTM atteint une précision de 72.67%, un rappel de 54.20% et un score F1 de 62.09%, tandis que le CNN obtient une précision de 75.70%, un rappel de 46.05 % et un score F1 de 57.26%.

Modèles	Exactitude (Accuracy)	Précision (Precision)	Rappel (Recall)	Score F1
CNN	83.72%	75.70%	46.05%	57.26%
Bi-LSTM	84.90%	72.67%	54.20%	62.09%
BERT	97.56%	100%	97.22%	98.59%

Tableau 4-10 : Performance des modèles sur l'ensemble de données GossipCop

3.5.Comparaison des performances des modèles et discussions

Le graphique illustré dans la figure 4-28 présente les courbes d'exactitudes d'entraînement et de validation des trois modèles : CNN, Bi-LSTM et BERT, sur une période de 10 époques.

L'exactitude du modèle BERT en entraînement évolue rapidement, dépassant les autres modèles dès les premières époques et atteignant presque une parfaite exactitude d'environ 99 %. D'autre part, le Bi-LSTM présente une progression plus lente alors que le CNN démarre plus tard mais finit par atteindre une exactitude comparable à celle du Bi-LSTM. En phase de validation, BERT est également le modèle le plus performant avec une exactitude très élevée et stable, prouvant ainsi son bon niveau de généralisation, contrairement aux modèles CNN et Bi-LSTM qui commencent à enregistrer à partir de l'époque 4 une légère chute de performances.

En termes de temps d'entraînement, le modèle CNN est bien plus rapide 39,2 secondes en comparaison des modèles Bi-LSTM et BERT, qui nécessitent tous deux un temps d'entraînement respectivement 21 minutes 44 secondes et 21 minutes 31 secondes. Certes, BERT donne les meilleurs résultats d'exactitude et de stabilité, mais son coût en temps d'apprentissages est plus fort, alors que le CNN constitue une alternative plus rapide avec des performances relativement satisfaisantes.

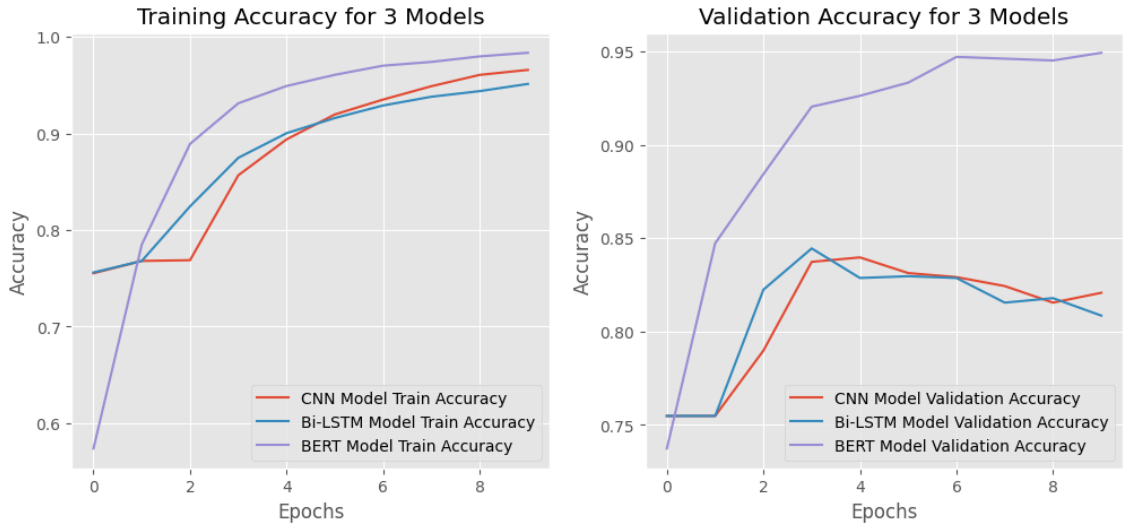

Figure 4-28 : Comparaison des Performances d'Entraînement et de validation des Modèles CNN, Bi-LSTM et BERT

4. Etude comparatives

Dans cette section, nous analysons et comparons les résultats obtenus à partir de deux jeux de données distincts : ISOT et GossipCop. Cette étude met en évidence une comparaison des résultats de notre méthode avec ceux d'autres travaux de détection de fausses informations, en tenant compte des modèles de classification et des techniques NLP utilisés.

Le tableau 4-11 illustre plusieurs travaux de recherche utilisant diverses combinaisons de méthodes NLP et modèles de classification pour la détection des fausses nouvelles. Les approches associant TF-IDF et modèles d'apprentissage statistique classique affichent des performances notables. Par exemple, Ozbay & Alatas (Ozbay & Alatas, 2020) atteignent une précision de 96,80 % avec TF-IDF et l'arbre de décision, tandis que Probierz et al. (Probierz, et al., 2021) ont obtenu 94,19 % avec un modèle SVM basé sur TF-IDF. Toutefois, les méthodes

d'apprentissage profond dépassent souvent les modèles classiques en matière d'exactitude. Par exemple, Nasir et al. (Nasir, et al., 2021) ont proposé un système hybride basé sur Glove avec un RNN et un modèle CNN-RNN, les résultats sont respectivement de 98% et 99%. Aussi, Amer et al. (Amer, et al., 2022) mettent également en avant les modèles profonds avec des résultats de 98,80% avec Stacked LSTM et 99,10% avec Stacked GRU.

Sur le jeu de données GossipCop, les performances sont généralement plus modestes, comme le montre le tableau 4-12, reflétant la complexité de ce corpus. Par exemple, Alghamdi et al. (Alghamdi, et al., 2022) atteignent 86,35 % de précision avec un SVM basé sur TF-IDF, et 86,16 % avec un CNN utilisant BERTbase. L'étude de Guo et al. (Guo, et al., 2023) montre que l'utilisation de modèles avancés comme BERT améliore considérablement les résultats : les précisions obtenues avec CNN, GRU et BERT sont respectivement de 74,10 %, 79,30 %, et 83,60 %.

Référence	**Technique NLP**	**Modèle de classification**	**Exactitude**
(Ozbay & Alatas, 2020)	TF-IDF	Decision tree	96.80%
(Probierz, et al., 2021)	TF-IDF	SVM	94.19%
(Nasir, et al., 2021)	Glove	RNN	98%
		CNN-RNN	99%
(Amer, et al., 2022)	Glove	BERT	96.90%
		Stacked LSTM	98.80%
		Stacked GRU	99.10%
Le système proposé	**Glove**	**CNN**	**97.74%**
		Bi-LSTM	**97.76%**
	BERT Embedding	**BERT**	**99.34%**

Tableau 4-11 : Résultats des modèles de détection de fausses nouvelles sur l'ensemble de données ISOT

Référence	Technique NLP	Modèle de classification	Exactitude
(Alghamdi, et al., 2022)	TF-IDF	SVM	86.35%
	BERT Embedding	CNN	86.16%
(Guo, et al., 2023)	———	CNN	74.10%
		GRU	79.30%
	BERT Embedding	BERT	83.60%
Le système proposé	**Glove**	**CNN**	**83.72%**
		Bi-LSTM	**84.90%**
	BERT Embedding	**BERT**	**97.56%**

Tableau 4-12 : Résultats des modèles de détection de fausses nouvelles sur l'ensemble de données GossipCop

Notre approche se démarque en démontrant une efficacité notable, notamment avec le modèle BERT, qui atteint une exactitude impressionnante de 99,34 % sur ISOT et de 97,54 % sur GossipCop. Ce résultat souligne la capacité de BERT à capturer des nuances complexes dans les textes. De plus, l'intégration de l'analyse de sentiment en tant que caractéristique principale est particulièrement avantageuse en raison de sa capacité à enrichir la compréhension contextuelle et émotionnelle des textes, améliorant ainsi la capacité du modèle à discriminer entre vraies et fausses nouvelles. En effet, ces performances confirment que l'analyse de sentiment apporte une dimension supplémentaire précieuse, permettant d'améliorer la précision et la robustesse de notre approche en tant que détecteur des fausses nouvelles.

5. Conclusion

Avant tout, nous avons expliqué au cours du chapitre présent les résultats expérimentaux de notre système de détection de fausses informations, au moyen de l'exploration et de l'analyse de la performance observée sur deux ensembles de données ISOT et GossipCop dont nous avons testé différents modèles de classification de texte. Les résultats ont montré que le modèle de BERT surpassait les autres, et en obtenant des taux de précision élevés sur les deux jeux de données, il prouve d'ores et déjà sa capacité à détecter des caractéristiques linguistiques plus fines, mais aussi que l'analyse de sentiment a été utile comme caractéristique contextuelle pour mieux discerner le faux du réel. En comparaison avec d'autres travaux de la littérature, notre approche est également plus précise, surtout sur le jeu de données ISOT. En tout état de cause,

la robustesse et l'efficacité de notre approche est confirmée par des résultats aussi intéressants que ceux obtenus dans des contextes de très haute diversité, comme celui du corpus GossipCop.

Conclusion générale

Dans cette recherche, nous avons étudié plusieurs méthodes de détection des fausses nouvelles notamment en considérant l'analyse de sentiments comme un élément central dans le système de classification. Nous avons analysé et comparé plusieurs méthodes de traitement du langage naturel et d'apprentissage profond à partir de jeux de données distincts afin de tester la généralité et l'efficacité des modèles.

Nous avons passé en revue dans un premier temps, des méthodologies de détection des fausses nouvelles à travers des approches supervisées, faiblement supervisées et non supervisées. L'intégration de l'analyse des sentiments a permis d'enrichir le modèle avec une compréhension à la fois contextuelle et émotionnelle des textes, conduisant ainsi à une meilleure différenciation des nouvelles réelles et fausses. Les résultats de nos expérimentations montrent que les techniques basées sur l'apprentissage profond, en particulier avec les modèles tels que BERT, surpassent les modèles classiques de par leur capacité à capturer des nuances sémantiques et contextuelles complexes dans les textes. Notre modèle basé sur BERT a atteint des performances impressionnantes avec une précision de 99,34 % sur le jeu de données ISOT et de 97,54 % sur le corpus GossipCop, soulignant la puissance des modèles de transformateurs pour ce type de tâche.

Cependant, malgré les performances prometteuses de notre approche, plusieurs défis subsistent, notamment la représentativité des ensembles de données comme FakeNewsNet qui peuvent ne pas rendre pleinement compte de la totalité des dynamiques de diffusion des informations fausses ou authentiques, la difficulté de généralisation de ces modèles à d'autres corpus ou d'autres langues, nécessitant des ajustements par rapport aux modèles précédents, comme des modèles multilingues (ensemble des langues ou des corpus linguistiques) ou la standardisation des caractéristiques des engagements des utilisateurs selon les plateformes.

Pour l'avenir, il serait intéressant d'étudier davantage la détection des fausses nouvelles en contexte multilingue et multi-sources, avec des labels plus fins. Tester le modèle sur des bases de données issues de sources variées pourrait renforcer sa robustesse et son adaptabilité. Par ailleurs, un axe de recherche pertinent serait la détection précoce des fausses nouvelles, en étudiant les premiers engagements utilisateurs pour évaluer plus vite la vérité des informations et, ce faisant, ralentir leur propagation en désinformation ; mais l'on pourrait même aussi comprendre comment les réactions initiales influent sur la fiabilité perçue d'une nouvelle.

Pour conclure, cette étude montre que la combinaison de l'analyse des sentiments et de modèles plus avancés d'apprentissage profond constitue une voie prometteuse pour détecter les fausses nouvelles. Les très bonnes performances obtenues montrent bien qu'il est important de prendre en compte les sentiments et les contextes sociaux à travers les réseaux sociaux. Ces résultats renforcent l'idée d'une dimension complémentaire à l'analyse des sentiments dans des modèles de classification pour relever le défi des fausses nouvelles, en vue d'un profit accru en précision et robustesse des systèmes de détection.

Bibliographie

(A. Alkhodair, et al., 2020) A. Alkhodair, S., H.H. Ding, S., C.M. Fung, B. & Liu, J., 2020. Detecting breaking news rumors of emerging topics in social media. *Information Processing & Management,* 57(2).

(Aggarwal, 2018) Aggarwal, C. C., 2018. Neural Networks and Deep Learning. *Springer.*

(Ahmad, et al., 2019a) Ahmad, S., Asghar, M. Z., Alotaibi, F. M. & Awan, I., 2019. Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human-centric Computing and Information Sciences*, 1 12, 9(1), pp. 1 - 23.

(Ahmed, et al., 2019b) Ahmed, S., Hinkelmann, K. & Corradini, F., 2019. *Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks-a survey.* s.l., s.n., p. 8.

(Ainapure, et al., 2023) Ainapure, B. S., Pise, R. N., Reddy, P. & Appasani, B., 2023. Sentiment Analysis of COVID-19 Tweets Using Deep Learning and Lexicon-Based Approaches. *SUSTAINABILITY.*

(Ajao, et al., 2019) Ajao, O., Bhowmik, D. & Zargari, S., 2019. *Sentiment Aware Fake News Detection on Online Social Networks.* s.l., Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, p. 2507–2511.

(Al-Bakri, et al., 2022) Al-Bakri, N. F., Yonan, J., Sadiq, A. & Abid, A., 2022. Tourism Companies Assessment via Social Media Using Sentiment Analysis. *Baghdad Science,* 19(2), p. 422–429.

(Alghamdi, et al., 2022) Alghamdi, J., Lin, Y. & Luo, S., 2022. A Comparative Study of Machine Learning and Deep Learning Techniques for Fake News Detection. *Information,* 13(12), p. 576.

(Alghamdi, et al., 2023) Alghamdi, J., Luo, S. & Lin, Y., 2023. A comprehensive survey on machine learning approaches for fake news detection. *Multimedia Tools and Applications* .

(Al-Makhadmeh & Tolba , 2020) Al-Makhadmeh, Z. & Tolba , A., 2020. Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing,* Volume 102, pp. 501-522.

(Al-Mashhadany, et al., 2022) Al-Mashhadany, A. K., Sadiq, A. T., Ali, S. M. & Ahmed , A. A., 2022. Healthcare assessment for beauty centers using hybrid

sentiment analysis. *Indonesian Journal of Electrical Engineering and Computer Science,* 28(2), p. 890–897.

(Al-Rakhami & Al-Amri, 2020) Al-Rakhami, M. S. & Al-Amri, A. M., 2020. Facts Save: Detecting COVID-19 Misinformation in Twitter. *IEEE Access,* Volume 8, p. 155961–155970.

(Amer, et al., 2022) Amer, E., Kwak, K.-S. & El-Sappagh, S., 2022. Context-Based Fake News Detection Model Relying on Deep Learning Models. *Electronics,* 11(8), p. 1255.

(Ammara , et al., 2019) Ammara , H. et al., 2019. False information detection in online content and its role in decision making: a systematic literature review. *Social Network Analysis and Mining,* Volume 9, p. 50.

(Anis, et al., 2020) Anis, S., Saad, S. & Aref, M., 2020. *Sentiment Analysis of Hotel Reviews Using Machine Learning Techniques.* s.l., Proceedings of the International Conference on Advanced Intelligent System and Informatics, p. 227–234.

(Anto, et al., 2016) Anto, M. P. et al., n.d. *Product rating using sentiment analysis.* s.l., Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), p. 3458–3462.

(Anto, et al., 2016) Anto, M. P. et al., 2016. *Product rating using sentiment analysis.* s.l., Proceedings of the International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), p. 3458–3462.

(Bahad, et al., 2019) Bahad, P., Saxena, P. & Kamal, R., 2019. Fake News Detection using Bi-directional LSTM-Recurrent Neural Network. *Procedia Computer Science,* Volume 165, pp. 74-82.

(Basiri & Kabiri, 2018) Basiri, M. E. & Kabiri, A., 2018. Words Are Important: Improving Sentiment Analysis in the Persian Language by Lexicon Refining. *ACM Transactions on Asian and Low-Resource Language Information Processing,* 17(4), pp. 1 - 18.

(Bhutani, et al., 2019) Bhutani, B., Rastogi, N., Sehgal, P. & Purwar, A., 2019. *Fake News Detection Using Sentiment Analysis.* s.l., Proceedings of the 2019 Twelfth International Conference on Contemporary Computing, pp. 1-5.

(Britz, 2015) Britz, D., 2015. *Recurrent Neural Networks Tutorial, Part 1– Introduction to Rnns.* [Online] Available at: http://www.wildml.com/2015/09/recurrent-neural-networkstutorial-part-1-introduction-to-rnns/ (

(Cao, et al., 2023) Cao, X. et al., 2023. Health Status Recognition Method for Rotating Machinery Based on Multi-Scale Hybrid Features and Improved Convolutional Neural Networks. *Sensors.*

(Chen, et al., 2019) Chen, X. et al., 2019. One-shot generative adversarial learning for MRI segmentation of craniomaxillofacial bony structures. *IEEE Transactions on Medical Imaging,* 39(3), pp. 787-796.

(Cui, et al., 2019) Cui, L., Wang, S. & Lee, D., 2019. *Sentiment-aware multi-modal embedding for detecting fake news.* s.l., Proceedings of the ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, pp. 41-48.

(Cui, et al., 2019) Cui, L., Wang, S. & Lee, D., 2019. *Sentiment-aware multi-modal embedding for detecting fake news.* s.l., Proceedings of the ASONAM '19: International Conference on Advances in Social Networks Analysis and Mining, pp. 41-48.

(Dai, et al., 2020) Dai, E., Sun, Y. & Wang, S., 2020. *Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository.* s.l., Proceedings of the Fourteenth International AAAI Conference onWeb and Social Media, pp. 853-862.

(Dang, et al., 2020) Dang, C., Moreno García, M. N. & De La Prieta, F., 2020. Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics,* 9(3), p. 483.

(de Oliveira, et al., 2020) de Oliveira, N. R., Medeiros, D. S. V. & Mattos, D. M. F., 2020. A Sensitive Stylistic Approach to Identify Fake News on Social Networking. *IEEE Signal Processing Letters,* Volume 27, p. 250–1254.

(Devlin, et al., 2019) Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K., 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding.* s.l., s.n., p. 4171–4186.

(Dey, et al., 2018) Dey, A. et al., 2018. *Fake News Pattern Recognition using Linguistic Analysis.* s.l., Proceedings of the 2018 Joint 7th International Conference on Informatics, Electronics Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision Pattern Recognition (icIVPR), pp. 305-309.

(Dong, et al., 2019) Dong, X., Victor, U., Chowdhury, S. & Qian, L., 2019. Deep Two-path Semi-supervised Learning for Fake News Detection. *arXiv preprint.*

(Dosovitskiy, et al., 2020) Dosovitskiy, A. et al., 2020. *An image is worth 16x16 words: Transformers for image recognition at scale.* s.l., s.n.

(Dou, et al., 2021) Dou, Y. et al., 2021. *User Preference-aware Fake News Detection.* s.l., s.n., pp. 2051-2055.

(Du, et al., 2017) Du, J., Xu, J., Song, H.-y. & Tao, C., 2017. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Medical Informatics and Decision Making,* Volume 17, p. 63–70.

(Du, et al., 2021) Du, J. et al., 2021. Cross-lingual COVID-19 Fake News Detection. *arXiv preprint*

(Dun, et al., 2021) Dun, Y. et al., 2021. *KAN: Knowledge-aware Attention Network for Fake News Detection.* s.l., s.n., pp. 81-89.

(Gangireddy, et al., 2020) Gangireddy, S. C. R., P, D., Long, C. & Chakraborty, T., 2020. *Unsupervised Fake News Detection: A Graph-based Approach.* s.l., s.n., pp. 75-83.

(Ghorbanpour, et al., 2023) Ghorbanpour, F., Ramezani, M., Fazli, M. A. & Rabiee, H. R., 2023. FNR: a similarity and transformer-based approach to detect multi-modal fake news in social media. *arXiv preprint.*

(Ghosh & Senapati, 2021) Ghosh, K. & Senapati, A., 2021. *Technical Domain Classification of Bangla Text using BERT.* s.l., Proceedings of Intelligent Computing and Technologies Conference (ICTCon2021).

(Giachanou, et al., 2019) Giachanou, A., Rosso, P. & Crestani, F., 2019. *Leveraging Emotional Signals for Credibility Detection.* s.l., s.n., p. 877–880.

(Glorot & Bengio, 2010) Glorot, X. & Bengio, Y., 2010. *Understanding the difficulty of training deep feedforward neural networks.* s.l., Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics.

(Guesbaya, et al., 2023) Guesbaya, M., García-Mañas, F., Rodríguez, F. & Megherbi, H., 2023. A Soft Sensor to Estimate the Opening of Greenhouse Vents Based on an LSTM-RNN Neural Network. *SENSORS.*

(Guo, et al., 2021) Guo, Z. et al., 2021. Fuzzy Detection System for Rumors Through Explainable Adaptive Learning. *IEEE Transactions on Fuzzy Systems,* 29(12), pp. 3650-3664.

(Guo, et al., 2023) Guo, Q., Kang, Z., Tian, L. & Chen, Z., 2023. TieFake: Title-Text Similarity and Emotion-Aware Fake News Detection. *International Joint Conference on Neural Networks (IJCNN),* pp. 1-7.

(Gupta & Joshi, 2020) Gupta, I. & Joshi, N., 2020. Enhanced twitter sentiment analysis using hybrid approach and by accounting local contextual semantic. *Journal of intelligent systems,* 29(1), p. 1611–1625.

(Habbat, et al., 2023) Habbat, N., Anoun, H. & Hassouni, L., 2023. Combination of GRU and CNN Deep Learning Models for Sentiment Analysis on French Customer Reviews Using XLNet Model. *IEEE ENGINEERING MANAGEMENT REVIEW.*

(Hamed, et al., 2023) Hamed, S. K., Ab Aziz, M. J. & Yaakub, M. R., 2023. Fake News Detection Model on Social Media by Leveraging Sentiment Analysis of News Content and Emotion Analysis of Users' Comments. *Sensors,* Volume 23, p. 1748.

(Hasan, et al., 2018) Hasan, A., Moin, S., Karim, A. & Shamshirband, S., 2018. Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications,* 9(1).

(Hasan, et al., 2018) Hasan, A., Moin, S., Karim, A. & Shamshirband, S., 2018. Machine Learning-Based Sentiment Analysis for Twitter Accounts. *Mathematical and Computational Applications,* 9(1).

(He, et al., 2015) He, K., Zhang, X., Ren, S. & Sun, J., 2015. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *IEEE International Conference on Computer Vision,* p. 1026–1034.

(Horne & Adali, 2017) Horne, B. D. & Adali, S., 2017. *This Just In: Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News.* s.l., Proceedings of the Workshops of the Eleventh International AAAI Conference on Web and Social Media, p. 759–766.

(Hu, et al., 2019) Hu, G. et al., 2019. *Multi-depth Graph Convolutional Networks for Fake News Detection.* s.l., s.n., pp. 698-710.

(Hu, et al., 2022) Hu, L., Wei, S., Zhao, Z. & Wu, B., 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open,* Volume 3, pp. 133-155.

(Huang, et al., 2019) Huang, Q. et al., 2019. *Deep Structure Learning for Rumor Detection on Twitter.* s.l., s.n., pp. 1-8.

(Jiang, et al., 2019) Jiang, S. et al., 2019. *User-Characteristic Enhanced Model for Fake News Detection in Social Media.* s.l., s.n., pp. 634-646.

(Kadan, et al., 2020) Kadan, A., Padmanabhan, D. & Lajish, V. L., 2020. *Emotion Cognizance Improves Health Fake News Identification.* s.l., Proceedings of the 24th Symposium on International Database Engineering & Applications.

(Kapusta, et al., 2020) Kapusta, J., Hájek, P., Munk, M. & Benko, Ľ., 2020. Comparison of fake and real news based on morphological analysis. *Procedia Computer Science,* Volume 171, pp. 2285-2293.

(Konkobo, et al., 2020) Konkobo, P. M. et al., 2020. *A Deep Learning Model for Early Detection of Fake News on Social Media.* s.l., s.n., pp. 1-6.

(Kumar, et al., 2019) Kumar, A., Singh, S. & Kaur, G., 2019. Fake News Detection of Indian and United States Election Data using Machine Learning Algorithm. *International Journal of Innovative Technology and Exploring Engineering,* 8(11), pp. 1559-1563.

(Kumar, et al., 2019) Kumar, A., Singh, S. & Kaur, G., 2019. Fake News Detection of Indian and United States Election Data using Machine Learning Algorithm. *International Journal of Innovative Technology and Exploring Engineering,* 8(11), pp. 1559-1563.

(Kumar, et al., 2020) Kumar, S. et al., 2020. Fake news detection using deep learning models: A novel approach. *Transactions on Emerging Telecommunications Technologies,* 31(2), p. e3767.

(Lalji & Deshmukh, 2016) Lalji, T. K. & Deshmukh, S. N., 2016. Twitter Sentiment Analysis Using Hybrid Approach. *International Research Journal of Engineering and Technology,* 3(6), p. 2887–2890.

(Li, et al., 2021) Li, X. et al., 2021. A novel self-learning semi-supervised deep learning network to detect fake news on social media. *Multimedia Tools and Applications,* 81(14), pp. 19341-19349.

(Li, et al., 2023) Li, H. et al., 2023. A sentiment analysis approach for travel-related Chinese online review content. *PEERJ. COMPUTER SCIENCE*

(Lin & Chen, 2020) Lin, L. & Chen, Z., 2020. Social rumor detection based on multilayer transformer encoding blocks. *Concurrency and Computation: Practice and Experience,* 33(6), p. e6083.

(Lin, et al., 2020) Lin, H., Zhang, X. & Fu, X., 2020. *A Graph Convolutional Encoder and Decoder Model for Rumor Detection.* s.l., s.n., pp. 300-306.

(Liu & Wu, 2020) Liu, Y. & Wu, Y.-f. b., 2020. Fned: A deep network for fake news early detection on social media.. *ACM Transactions on Information Systems,* Volume 38, p. 1–33.

(Liu, et al., 2022) Liu, J., Liu, Y. & Zhang, Q., 2022. A weight initialization method based on neural network with asymmetric activation function. *Neurocomputing,* Volume 483, pp. 171-182.

(Long, et al., 2018) Long, H., Liao, B., Xu, X. & Yang, J., 2018. A hybrid deep learning model for predicting protein hydroxylation sites. *International Journal of Molecular Sciences.*

(Ma & Gao, 2020) Ma, J. & Gao, W., 2020. *Debunking rumors on Twitter with tree transformer.* s.l., s.n., pp. 5455-5466.

(Ma, et al., 2021) Ma, J. et al., 2021. Improving rumor detection by promoting information campaigns with transformer-based generative adversarial learning. *IEEE Transactions on Knowledge and Data Engineering,* 35(3), pp. 1-14.

(Machová, et al., 2022) Machová, K., Mach, M. & Porezaný, M., 2022. Deep Learning in the Detection of Disinformation about COVID-19 in Online Space. *Sensors,* Volume 22, p. 9319.

(Meel & Vishwakarma, 2021) Meel, P. & Vishwakarma, D. K., 2021. Fake News Detection using Semi-Supervised Graph Convolutional Network. *arXiv preprint.*

(Nandi & Agrawal, 2016) Nandi, V. & Agrawal, S., 2016. Political sentiment analysis using hybrid approach. *International Research Journal ofEngineering and Technology,* 3(5), p. 1621–1627.

(Nasir, et al., 2021) Nasir, J. A., Khan, O. S. & Varlamis, I., 2021. Fake news detection: A hybrid CNN-RNN based deep learning approach. *International Journal of Information Management Data,* p. 100007.

(Ngoge, 2016) Ngoge, L. A., 2016. Real–time sentiment analysis for detection of terrorist activities in kenya. *Strathmore University.*

(Nguyen, et al., 2020) Nguyen, V.-H., Sugiyama, K., Nakov, P. & Kan, M.-Y., 2020. *FANG: Leveraging Social Context for Fake News Detection Using Graph Representation.* s.l., s.n., pp. 1165-1174.

(Omar, et al., 2021) Omar, A., Mahmoud, T. M., Abd-El-Hafeez, T. & Mahfouz, A., 2021. Multi-label Arabic text classification in Online Social Networks. Information Systems, Volume 100.

(Oshikawa, et al., 2020) Oshikawa, R., Qian, J. & Wang, W. Y., 2020. *Survey on Natural Language Processing for Fake News Detection.* s.l., Proceedings of the 12th Language Resources and Evaluation Conference, p. 6086–6093.

(Ozbay & Alatas, 2020) Ozbay, F. A. & Alatas, B., 2020. Fake news detection within online social media using supervised artificial intelligence algorithms. *Physica A: Statistical Mechanics and its Applications,* Volume 540, p. 123174.

(Özgöbek & Gulla, 2017) Özgöbek, Ö. & Gulla, J. A., 2017. Towards an understanding of fake news. *CEUR workshop proceedings,* Volume 2041, p. 35–42.

(Popat, et al., 2016) Popat, K., Mukherjee, S., Strötgen, J. & Weikum, G., 2016. *Credibility Assessment of Textual Claims on the Web.* s.l., Proceedings of the 25th ACM International on Conference on Information and Knowledge Management, p. 2173–2178.

(Popat, et al., 2017) Popat, K., Mukherjee, S., Strötgen, J. & Weikum, G., 2017. *Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media.* s.l., Proceedings of the 26th International Conference onWorldWideWeb Companion, p. 1003–1012.

(Probierz, et al., 2021) Probierz, B., Stefanski, P. & Kozak, J., 2021. Rapid detection of fake news based on machine learning methods. *Procedia Computer Science,* Volume 192, pp. 2893-2902.

(Qi, et al., 2019) Qi, P. et al., 2019. *Exploiting Multi-domain Visual Information for Fake News Detection.* s.l., s.n., p. 518–527.

(Qian, et al., 2021a) Qian, S. et al., 2021. *Hierarchical Multi-modal Contextual Attention Network for Fake News Detection.* s.l., s.n., pp. 153-162.

(Qian, et al., 2021b) Xu, C., Fang, Q., Hu, J. & Qian, S., 2021. Knowledge-aware Multi-modal Adaptive Graph Convolutional Networks for Fake News Detection. *ACM Transactions on Multimedia Computing,* 17(3), pp. 1-23.

(Rashkin, et al., 2017) Rashkin, H. et al., 2017. *Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking.* s.l., Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, p. 2931–2937.

(Rehman, et al., 2021) Rehman, Z. U. et al., 2021. Understanding the Language of ISIS: An Empirical Approach to Detect Radical Content on Twitter Using

Machine Learning. Computers, Materials & Continua, 66(2), pp. 1075-1090.

(Reis, et al., 2019) Reis, J. C. S. et al., 2019. Supervised Learning for Fake News Detection. *IEEE Intelligent Systems,* pp. 76 - 81.

(Ruangkanokmas, et al., 2016) Ruangkanokmas, P., Achalakul, T. & Akkarajitsakul, K., 2016. *Deep Belief Networks with Feature Selection for Sentiment Classification.* s.l., Proceedings of the 2016 7th International Conference on Intelligent Systems,Modelling and Simulation (ISMS), p. 9–14.

(Salur & Aydin, 2020) Salur, M. U. & Aydin, I., 2020. A Novel Hybrid Deep Learning Model for Sentiment Classification. *IEEE Access,* Volume 8, p. 58 080–58 093.

(Sharma & Moh, 2016) Sharma, P. & Moh, T.-S., 2016. Prediction of Indian election using sentiment analysis on Hindi Twitter. IEEE International Conference on Big Data (Big Data), p. 1966–1971.

(Sharma & Moh, 2016) Sharma, P. & Moh, T.-S., 2016. Prediction of Indian election using sentiment analysis on Hindi Twitter. *IEEE International Conference on Big Data (Big Data),* p. 1966–1971.

(Shehu, et al., 2021) Shehu, H. A., Sharif, M. H. U., Datta, R. & Sharif, M. H., 2021. Deep Sentiment Analysis: A Case Study on Stemmed Turkish Twitter Data. *IEEE Access,* Volume 9, p. 56 836–56 854.

(Shrivastava, et al., 2020) Shrivastava, G. et al., 2020. Defensive Modeling of Fake News Through Online Social Networks. *IEEE Transactions on Computational Social Systems,* 7(5), p. 1159–1167.

(Shu & Liu, 2019) Shu, K. & Liu, H., 2019. Detecting fake news on social media. Synthesis Lectures Data. In: *Synthesis Lectures on Data Mining and Knowledge Discovery.* s.l.:s.n., pp. 1-129.

(Shu, et al., 2019a) Shu, K. et al., 2019. *Defend: Explainable fake news detection.* s.l., s.n., pp. 395-405.

(Shu, et al., 2019b) Shu, K., Wang, S. & Liu, H., 2019. *Beyond News Contents: The Role of Social Context for Fake News Detection.* s.l., s.n., pp. 312-320.

(Shu, et al., 2020a) Shu, K. et al., 2020a. Early detection of fake news with multi-source weak social supervision. Dans: *ECML/PKDD.* s.l.:s.n., pp. 650-666.

(Shu, et al., 2020b) Shu, K. et al., 2020. FakeNewsNet: A Data Repository with News Content, Social Context and Spatialtemporal Information for Studying Fake News on Social Media. *Big Data,* p. 171–188.

(Singh, et al., 2022) Singh, C., Imam, T., Wibowo, S. & Grandhi, S., 2022. A Deep Learning Approach for Sentiment Analysis of COVID-19 Reviews. *Applied Sciences,* 12(8).

(Singhal, et al., 2019) Singhal, S. et al., 2019. *SpotFake: A Multi-modal Framework for Fake News Detection.* s.l., s.n., pp. 39-47.

(Song, et al., 2021) Song, C., Shu, K. & Wu, B., 2021. Temporally evolving graph neural network for fake news detection. *Information Processing and Management,* 58(6), p. 102712.

(Sun, et al., 2022) Sun, M., Zhang, X., Zheng, J. & Ma, G., 2022. *DDGCN: Dual Dynamic Graph Convolutional Networks for Rumor Detection on Social Media.* s.l., s.n., pp. 4611-4619.

(Tacchini, et al., 2017) Tacchini, E. et al., 2017. Some Like it Hoax: Automated Fake News Detection in Social Networks. *arXiv.*

(Tam, et al., 2021) Tam, S., Said, R. B. & Tanriöver, Ö. Ö., 2021. A ConvBiLSTM Deep Learning Model-Based Approach for Twitter Sentiment Classification. *IEEE Access,* Volume 9, p. 41 283–41 293.

(Tian, et al., 2021) Tian, L., Zhang, X. & Lau, J. H., 2021. *Rumour Detection via Zero-shot Cross-lingual Transfer Learning.* s.l., s.n., pp. 603-618.

(Varol, et al., 2017) Varol, O., Ferrara, E., Menczer, F. & Flammini, A., 2017. Early detection of promoted campaigns on social media. *EPJ Data Science.*

(Vedova, et al., 2018) Vedova, M. L. D. et al., 2018. Automatic Online Fake News Detection Combining Content and Social Signals. *2018 22nd conference of open innovations association (FRUCT),* p. 272–279.

(Verma & Thakur, 2018) Verma, B. & Thakur, R. S., 2018. *Sentiment Analysis Using Lexicon and Machine Learning-Based Approaches: A Survey.* s.l., Proceedings of International Conference on Recent Advancement on Computer and Communication, p. 441–447.

(Vicario, et al., 2019) Vicario, M. D., Quattrociocchi, W., Scala, A. & Zollo, F., 2019. Polarization and Fake News: Early Warning of Potential Misinformation Targets. *ACM Transactions on the Web.*

(Vosoughi, et al., 2018) Vosoughi, S., Roy, D. & Aral, S., 2018. The spread of true and false news online. *Science,* p. 1146–1151.

(Wang, et al., 2021) Wang, Y. et al., 2021. *Multimodal Emergent Fake News Detection via Meta Neural Process Networks.* s.l., s.n., p. 3708–3716.

(Wu, et al., 2021) Wu, Y. et al., 2021. *Multimodal Fusion with Co-Attention Networks for Fake News Detection.* s.l., s.n., p. 2560–2569.

(Yang, et al., 2018) Yang, Y. et al., 2018. TI-CNN: Convolutional Neural Networks for Fake News Detection. *arXiv.*

(Yang, et al., 2019) Yang, S. et al., 2019. *Unsupervised fake news detection on social media: a generative approach.* s.l., s.n., pp. 5644-5651.

(Yang, et al., 2021) Yang, X. et al., 2021. *Rumor detection on social media with graph structured adversarial learning.* s.l., s.n., pp. 1417-1423.

(Yuan, et al., 2019) Yuan, C. et al., 2019. *Jointly embedding the local and global relations of heterogeneous graph for rumor detection.* s.l., s.n., pp. 796-8.5.

(Yuan, et al., 2020) Yuan, C. et al., 2020. *Early Detection of Fake News by Utilizing the Credibility of News, Publishers, and Users based on Weakly Supervised Learning.* s.l., s.n., pp. 5444-5454.

(Zhang, et al., 2018) Zhang, L., Wang, S. & Liu, B., 2018. Deep learning for sentiment analysis: A survey. *WIREs Data Mining and Knowledge Discovery,* 8(4).

(Zhang, et al., 2021) Zhang, X. et al., 2021. *Mining Dual Emotion for Fake News Detection.* s.l., Proceedings of The World Wide Web Conference WWW 2021.

(Zhou & Zafarani, 2018) Zhou, X. & Zafarani, R., 2018. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *arXiv preprint.*

(Zhou & Zafarani, 2019) Zhou, X. & Zafarani, R., 2019. Network-based Fake News Detection: A Pattern-driven Approach. *ACM SIGKDD Explorations Newsletter,* 21(2), pp. 48-60.

(Zhou & Zafarani, 2020) Zhou, X. & Zafarani, R., 2020. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. *ACM Computing Surveys,* p. 37.

(Zhou, et al., 2020a) Zhou, X., Jain, A., Phoha, V. V. & Zafarani, R., 2020. Fake News Early Detection: A Theory-driven Model. *Digital Threats: Research and Practice,* 1(2), p. 11–25.

(Zhou, et al., 2020b) Zhou, X., Wu, J. & Zafarani, R., 2020. *Similarity-Aware Multi-modal Fake News Detection.* s.l., s.n., pp. 354-367.

(Zhou, et al., 2023) Zhou, J., Zeng, X., Zou, Y. & Zhu, H., 2023. Position-Wise Gated Res2Net-Based Convolutional Network with Selective Fusing for Sentiment Analysis. *Entropy (Basel).*

(Zhu, et al., 2022) Zhu, L. et al., 2022. Deep learning for aspect-based sentiment analysis: a review. *PeerJ Computer Science.*

(Zubiaga, et al., 2016) Zubiaga, A., Liakata, M. & Procter, R., 2016. Learning Reporting Dynamics during Breaking News for Rumour Detection in Social Media. *arXiv.*